Module 7: Duplication and Degradation

Objective

The objectives of this module are:

1. To determine if the gene under investigation has any paralogs
2. To determine if the gene under investigation is possibly a pseudogene.

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

Paralogs

During earlier queries into NCBI BLAST, you might have noticed significant matches of your query gene within the same organism. Orthologs are proteins that share similarity with your protein, but which are found in a different organism. Paralogs, on the other hand, are proteins with similarity to your query that are found in the same organism. Paralogs are the result of a gene duplication event within a genome. If the gene duplication event occurred recently in the evolution of the gene, paralogs will often bear a great deal of sequence similarity to each other. Over time, paralogs often develop different functions, and this sequence similarity may disappear. Another possible fate of a paralog is to become a pseudogene (a nonfunctional gene) Paralogs are normally identified by sequence similarity searches (e.g. BLAST) of a query protein against the rest of the same genome. If you find paralogs for the gene you are annotating, you should determine if your gene or the paralog(s) are pseudogenes.

Pseudogenes (taken from: http://en.wikipedia.org/wiki/Pseudogene)

Pseudogenes are characterized by a combination of homology to a known gene and nonfunctionality. That is, although every pseudogene has a DNA sequence that is similar to some functional gene, they are nonetheless unable to produce functional final protein products. Pseudogenes are sometimes difficult to identify and characterize in genomes, because the two requirements of homology and nonfunctionality are usually implied through sequence alignments rather than biologically proven.

Homology is implied by sequence identity between the DNA sequences of the pseudogene and parent gene. After aligning the two sequences, the percentage of identical base pairs is computed. A high sequence identity means that it is highly likely that these two sequences diverged from a common ancestral sequence (are homologous), and highly unlikely that these two sequences have evolved independently.

Nonfunctionality can manifest itself in many ways. Normally, a gene must go through several steps to a fully functional protein in bacteria: transcription, translation, and protein folding are all required parts of this process. If any of these steps fails, then the sequence may be considered nonfunctional. In high-throughput pseudogene identification, the most commonly identified disablements are premature stop codons and frameshifts, which almost universally prevent the translation of a functional protein product.

It is possible for a gene to be a pseudogene even if no paralogs are found. If a gene is mutated and loses function (becomes a pseudogene) the functionality of that gene is lost to the bacterium.

The psuedogene identification components of this module are among the most complicated to interpret in the annotation exercises.

Procedures

Paralogs

1. To find the top paralog, BLAST your gene against the genome of **YOUR ORGANISM** (*K. sedentarius* used in the example).
2. Navigate to protein BLAST start page at: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
3. Paste the FASTA formatted amino acid sequence of your protein into the search window and enter *Kytococcus sedentarius* DSM 20547 (taxid:478801) in the organism box. Once you start typing the name in the box a dropdown list will appear that will allow you to select *Kytococcus sedentarius* DSM 20547 (taxid:478801). There leave the database as nr (Figure 7.1) and click BLAST.

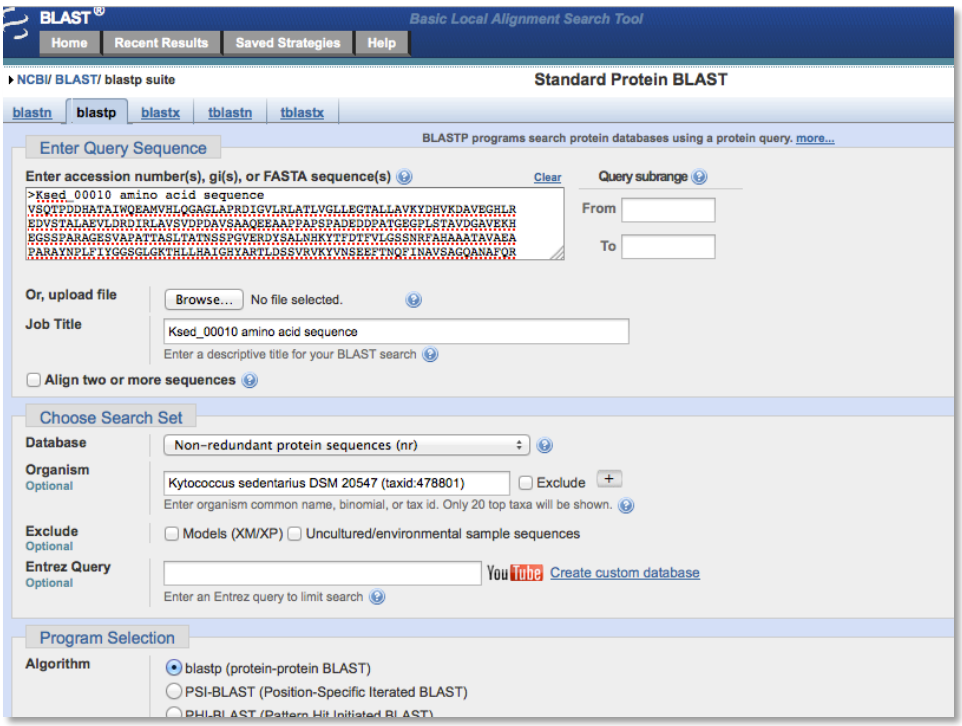


Figure 7.1. The protein BLAST start page for a paralog search.

4. The BLAST results for Ksed_00010 are shown in (Figure 7.2). Note that only one hit has a significant E value and score. Since we expect to find our gene matching itself when searching the *Kytococcus sedentarius* genome, finding only one significant result suggests that there are NO paralogs for Ksed_00010.

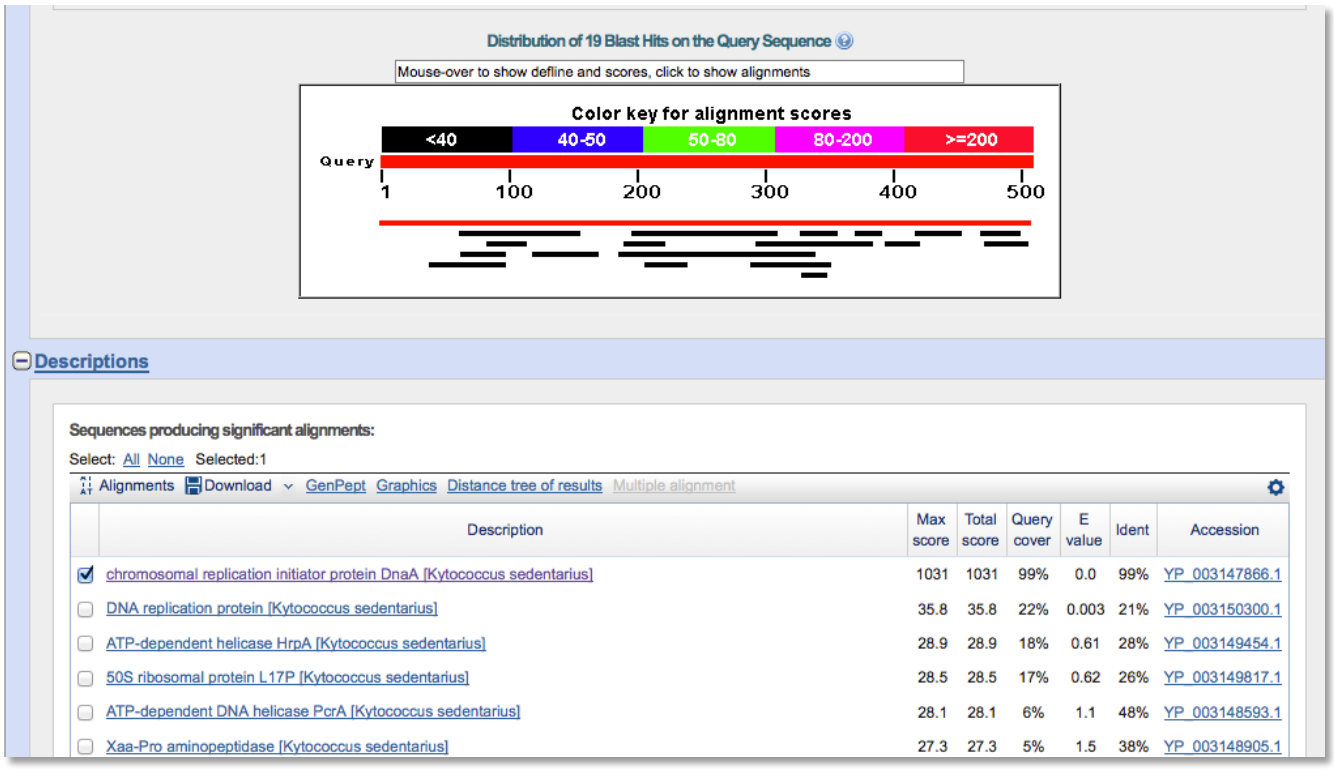


Figure 7.2. The paralog BLAST results for Ksed_00010. Only one significant hit is seen.

5. To confirm that the significant hit is in fact Ksed_00010 matching itself, view the alignment of the top hit and hover your cursor over the Gene hyperlink, if one is present (Figure 7.3) to reveal the locus tag of the hit. (Figure 7.4).

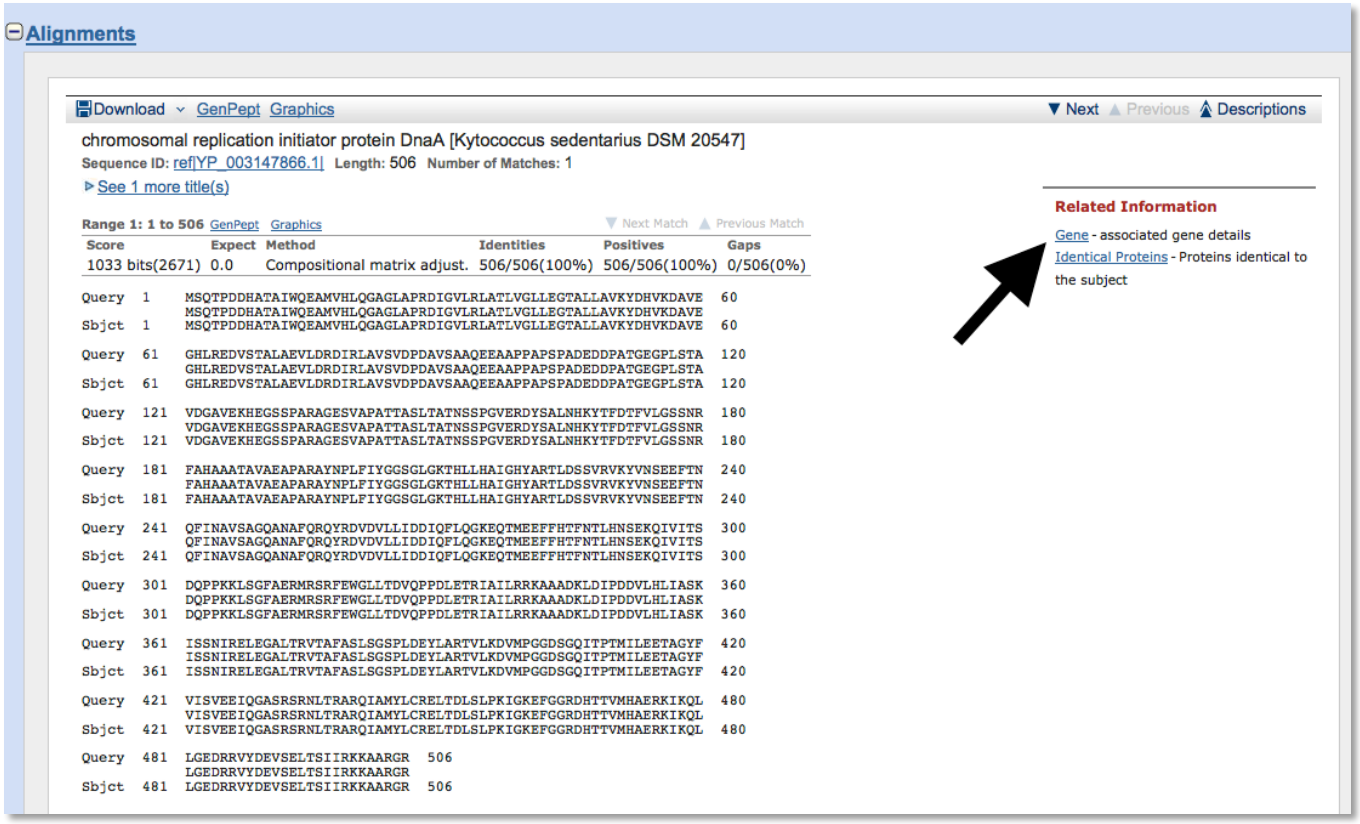


Figure 7.4. The alignment for the first hit in the paralog BLAST search for ksed_00010. The arrow indicates an active hyperlink to a page describing the subject gene. If you hover the cursor over the Gene link without actually selecting it, you will see the locus tag for the subject (Figure 7.5)

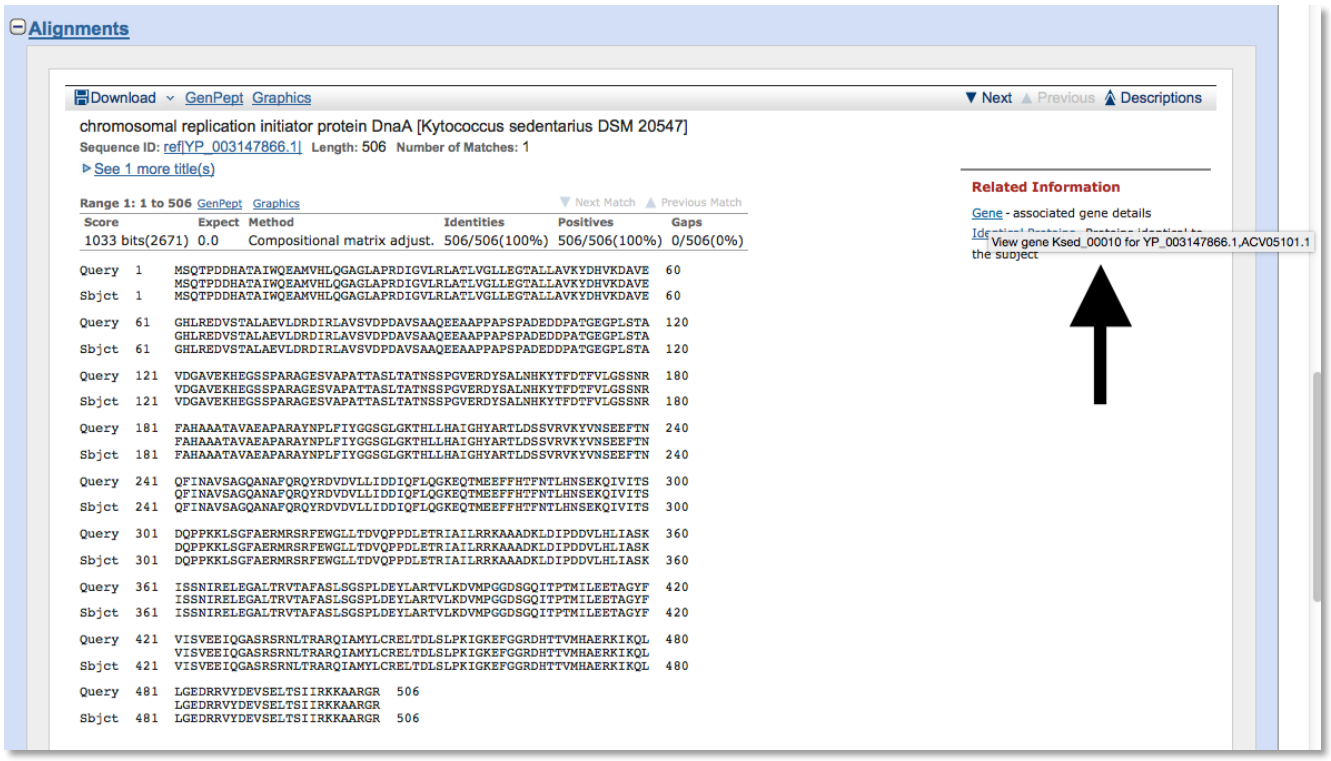


Figure 7.5. The alignment for the first hit in the paralog BLAST search for ksed_00010 with the locus tag for the subject revealed (arrow). The arrow in figure 7.5 point to the link showing that the subject is gene Ksed_00010, confirming the BLAST hit is actually Ksed_00010 and therefore not a paralog.

6. If no paralogs exist, record "No paralogs found" in the "Paralog gene product name" text box of the Lab Notebook page (Figure 7.6).

7. If a single paralog is found, fill in the information for the paralog in each of the paralog notebook text boxes, including the locus tag of the paralog along with the name in the "paralog gene product name" text box (Figure 7.6).

   A. Calculate the alignment length as you did for your BLAST hits in Module 2. Subtract the first number in the alignment of the query gene from the last number in the alignment of the query gene and add 1.

Figure 7.6. The GENI-ACT Paralog notebook. If you find one paralog then simply fill in the information requested. Include the locus tag in the gene product name box as well. See text for what you should do if you find more than one paralog for your gene.

8. For some entries in BLAST the "associated gene details" link may not be visible due to a change in the annotation pipeline at NCBI.

   A. If you find that the single hit in the nr database that has 100% query coverage and 100% identity, then it is likely you have simply found your own protein sequence in the database.

   B. However, there is another way to find the locus tag for potential paralogs. This is illustrated below using the example of blast hits shown in Figure 7.7

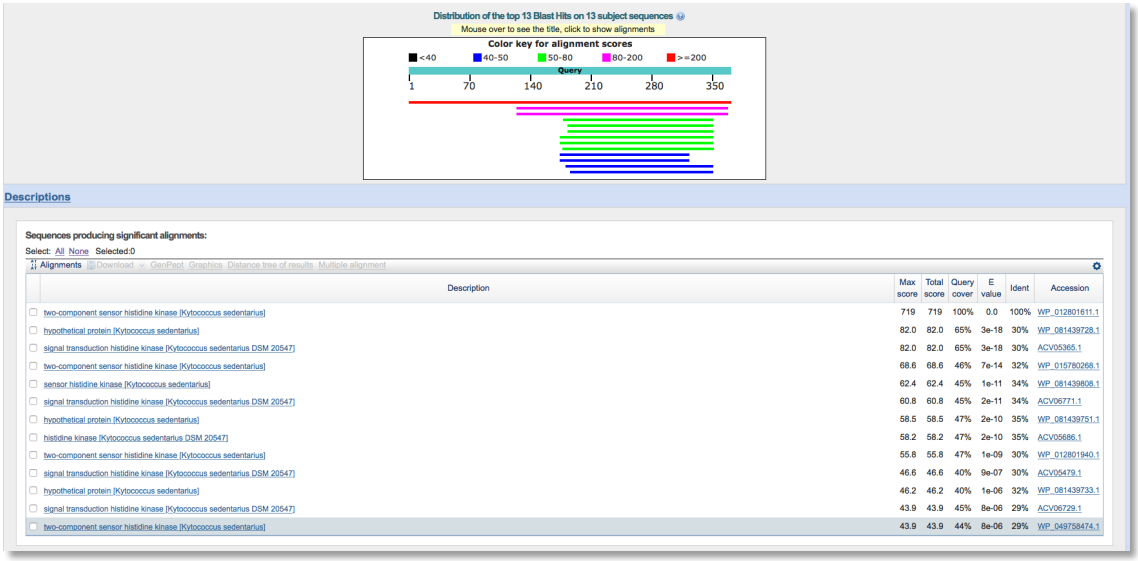
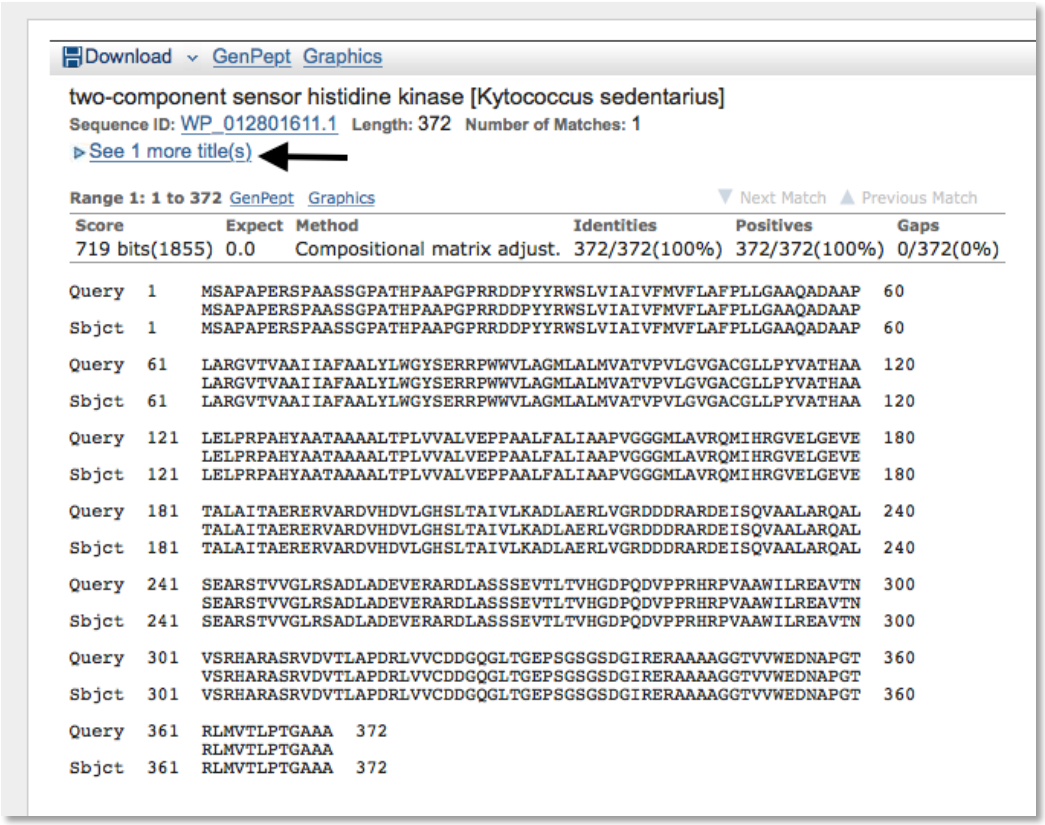


Figure 7.7. The BLAST results for a histidine sensor kinase found in *K. sedentarius* showing a large number of potential paralogs.

C. In figure 7.8 below, you can see the pairwise alignment for the top blast hit from figure 7.7. Ther first thing to note is that the identities and coverage are 100%, suggesting this hit is actually the same as the query sequence that was submitted. There is a WP_ at the front of the sequence ID which means that this is a "reference sequence" designation and hence no gene related information link (as described above) will be found for this sequence. A description of the Prokaryotic RefSeq Genomes project can be found at: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/ . If there is only a WP_ sequence prefix associated with the BLAST hit, you will not find a locus tag linked to the sequence and the sequence would NOT be considered a paralog. However, if you see a "See more titles" link below the main Sequence ID (as shown by the arrow in figure 7.8), it is likely you can follow a link to a locus tag in the genome of the bacterium you are annotating.



Download GenPept Graphics

two-component sensor histidine kinase [Kytococcus sedentarius]

Sequence ID: WP_012801611.1 Length: 372 Number of Matches: 1

See 1 more title(s)

Range 1: 1 to 372 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
719 bits(1855)	0.0	Compositional matrix adjust.	372/372(100%)	372/372(100%)	0/372(0%)

```
Query  1    MSAPAPERSPAASSGPATHPAAPGPRRDDPYYRWSLVIAIVFMVFLAFPLLGAAQADAAP  60
            MSAPAPERSPAASSGPATHPAAPGPRRDDPYYRWSLVIAIVFMVFLAFPLLGAAQADAAP
Sbjct  1    MSAPAPERSPAASSGPATHPAAPGPRRDDPYYRWSLVIAIVFMVFLAFPLLGAAQADAAP  60

Query  61   LARGVTVAAIIAFAALYLWGYSERRPWWVLAGMLALMVATVPVLGVGACGLLPYVATHAA  120
            LARGVTVAAIIAFAALYLWGYSERRPWWVLAGMLALMVATVPVLGVGACGLLPYVATHAA
Sbjct  61   LARGVTVAAIIAFAALYLWGYSERRPWWVLAGMLALMVATVPVLGVGACGLLPYVATHAA  120

Query  121  LELPRPAHYAATAAAALTPLVVALVEPPAALFALIAAPVGGGMLAVRQMIHRGVELGEVE  180
            LELPRPAHYAATAAAALTPLVVALVEPPAALFALIAAPVGGGMLAVRQMIHRGVELGEVE
Sbjct  121  LELPRPAHYAATAAAALTPLVVALVEPPAALFALIAAPVGGGMLAVRQMIHRGVELGEVE  180

Query  181  TALAITAERERVARDVHDVLGHSLTAIVLKADLAERLVGRDDDRARDEISQVAALARQAL  240
            TALAITAERERVARDVHDVLGHSLTAIVLKADLAERLVGRDDDRARDEISQVAALARQAL
Sbjct  181  TALAITAERERVARDVHDVLGHSLTAIVLKADLAERLVGRDDDRARDEISQVAALARQAL  240

Query  241  SEARSTVVGLRSADLADEVERARDLASSSEVTLTVHGDPQDVPPRHRPVAAWILREAVTN  300
            SEARSTVVGLRSADLADEVERARDLASSSEVTLTVHGDPQDVPPRHRPVAAWILREAVTN
Sbjct  241  SEARSTVVGLRSADLADEVERARDLASSSEVTLTVHGDPQDVPPRHRPVAAWILREAVTN  300

Query  301  VSRHARASRVDVTLAPDRLVVCDDGQGLTGEPSGSGSDGIRERAAAAGGTVVWEDNAPGT  360
            VSRHARASRVDVTLAPDRLVVCDDGQGLTGEPSGSGSDGIRERAAAAGGTVVWEDNAPGT
Sbjct  301  VSRHARASRVDVTLAPDRLVVCDDGQGLTGEPSGSGSDGIRERAAAAGGTVVWEDNAPGT  360

Query  361  RLMVTLPTGAAA  372
            RLMVTLPTGAAA
Sbjct  361  RLMVTLPTGAAA  372
```

Figure 7.8. The pairwise alignment from the top BLAST hit shown in Figure 7.7. See the text for a full description.

D. Figure 7.9 shows the display after the "See one more title" hyperlink is connected. A title and accession number more related to the *Kytococcus* genome under investigation in the example is revealed. Clicking on the accession number (arrow in Figure 7.9) will take you to a complete Genbank record for that accession number as shown in Figures 7.10 and 7.11

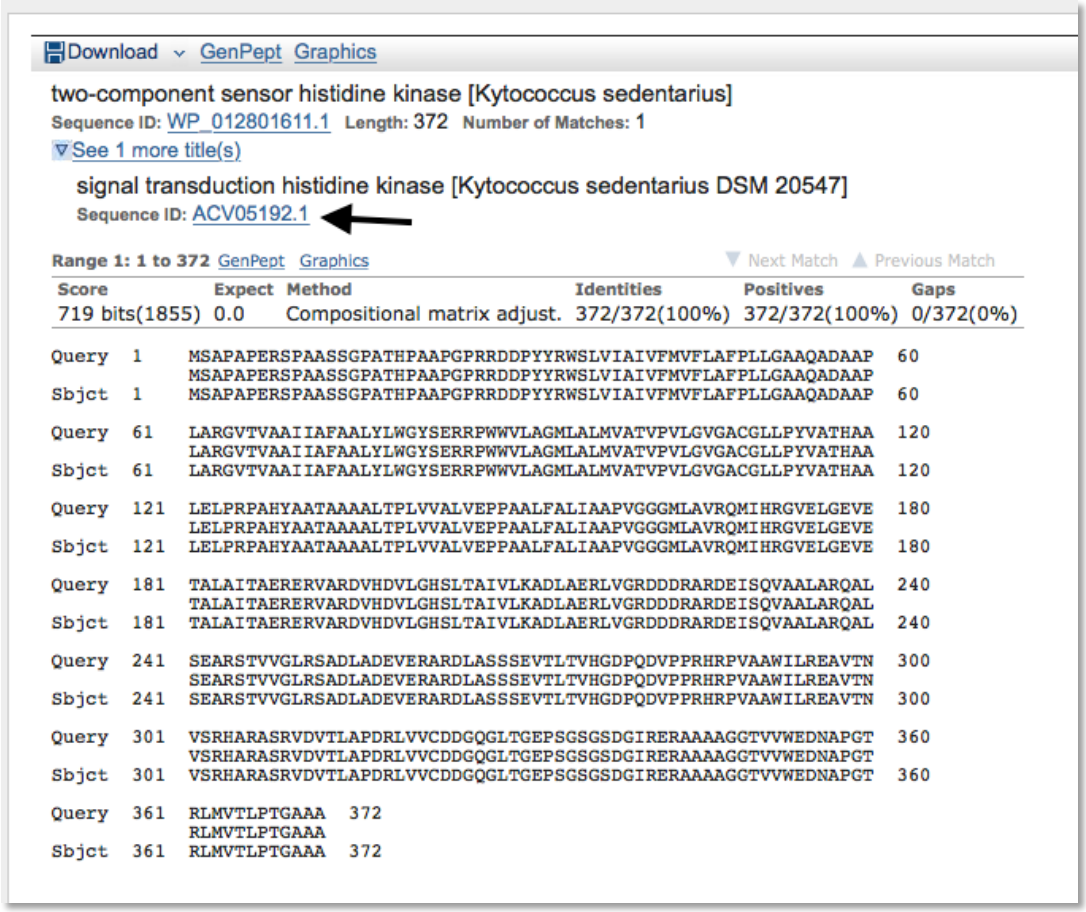


Figure 7.9. The appearance of the BLAST hit from figure 7.8 after clicking on the "See 1 more title" hyperlink. The arrow points to a hyperlink to the full Genbank record that was revealed.

E. Figure 7.10 shows the top most portion of the Genbank record revealed after clicking on the arrow in Figure 7.9. Scrolling down toward the bottom of the page until you come to a section entitled CDS will reveal the locus tag in the genome under investigation as is shown in Figure 7.11

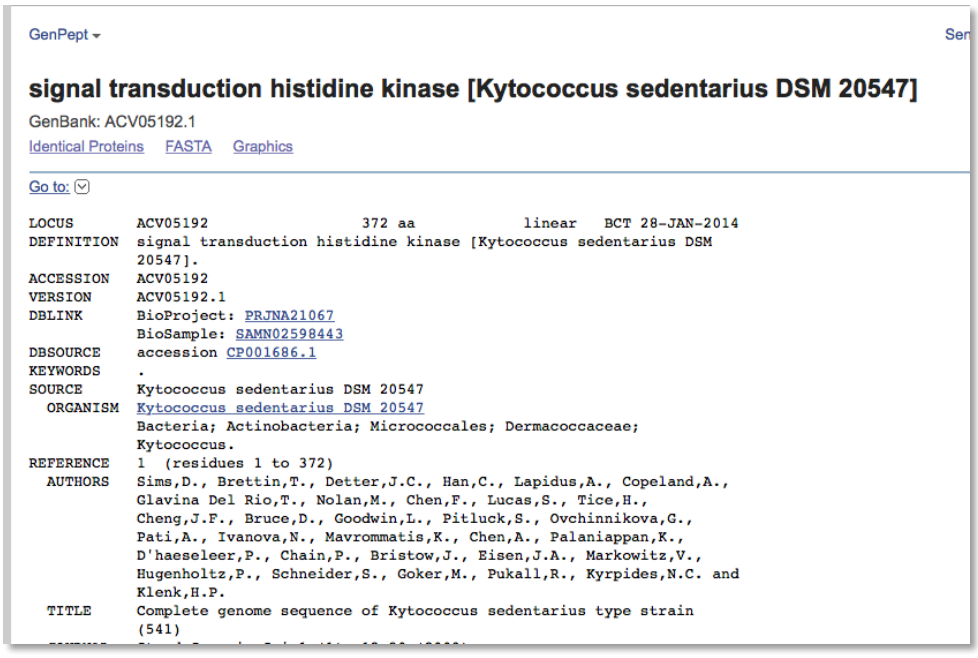


Figure 7.10. The top portion of the full Genbank record for the "See 1 more title" hit shown in Figure 7.9.

```
FEATURES             Location/Qualifiers
     source          1..372
                     /organism="Kytococcus sedentarius DSM 20547"
                     /strain="DSM 20547"
                     /culture_collection="DSM:20547"
                     /db_xref="taxon:478801"
     Protein         1..372
                     /product="signal transduction histidine kinase"
     Region          34..367
                     /region_name="COG4585"
                     /note="Signal transduction histidine kinase [Signal
                     transduction mechanisms]"
                     /db_xref="CDD:226951"
     Region          188..255
                     /region_name="HisKA_3"
                     /note="Histidine kinase; pfam07730"
                     /db_xref="CDD:285031"
     Region          293..367
                     /region_name="HATPase_c"
                     /note="Histidine kinase-, DNA gyrase B-, and HSP90-like
                     ATPase; pfam02518"
                     /db_xref="CDD:280651"
     Site            order(296,300,303,321,323,325,327..328,334..337,353,355,
                     359..360,362)
                     /site_type="other"
                     /note="ATP binding site [chemical binding]"
                     /db_xref="CDD:238030"
     Site            300
                     /site_type="other"
                     /note="Mg2+ binding site [ion binding]"
                     /db_xref="CDD:238030"
     Site            order(325,327,334,336)
                     /site_type="other"
                     /note="G-X-G motif"
                     /db_xref="CDD:238030"
     CDS             1..372
                     /locus_tag="Ksed_01000"
                     /coded_by="complement(CP001686.1:102451..103569)"
                     /note="PFAM: Histidine kinase-, DNA gyrase B-, and
                     HSP90-like ATPase; Histidine kinase"
                     /transl_table=11
ORIGIN
```

Figure 7.11. The lower portion of the full Genbank record. The arrow points to where the locus tag information can be found.

9. It is possible for more than one paralog to be found. In the event that you find more than one paralog for your gene, you should record information for all of the paralogs in your notebook. This can most easily be done inserting the locus tag for each paralog in the paralog gene product name text box from the most significant to the least significant match. You can then scroll down to the alignments portion of the BLAST results and using the Snip tool (PC) or Grab tool (Mac) to capture the multiple alignment and statistics. Upload the images to the "pairwise alignment" box of the notebook. Note that if you upload multiple images they will appear in reverse order in the notebook. That is, the image you load last will be at the top of the notebook section and the one you load first will be at the bottom. You should thus load the images in reverse order to the notebook page (i.e., least significant of the alignments first and most significant of the alignments last).

Pseudogenes

(modified based on a file original obtained courtesy of Seth Axen at the Joint Genome Institute)

As discussed in the background section for this module above, that a pseudogene is a genetic sequence which is nonfunctional. In high-throughput pseudogene identification, the most commonly identified disablements are mutations in the DNA sequence that create either premature stop codons (review background section to this manual) or frameshifts, which almost universally prevent the translation of a functional protein product.

What distinguishes a pseudogene from any other noncoding DNA sequence is that typically, a pseudogene will align well to a known protein as seen on BLAST, CD, and/or Pfam while appearing to be a legitimate coding sequence during structural annotation. Pseudogenes are thought to be formed by two mechanisms in prokaryotes:

- **Duplicated pseudogenes** are sequences which were formed through a duplication of a functional gene, followed by mutagenesis to remove functionality. In this case you would find a paralog or paralogs of the gene under investigation and either the gene under investigation or the paralog ***might*** be a pseudogene.
- **Disabled pseudogenes** are the original sequence of a functional gene which has been disabled through mutagenesis so that the microbe no longer contains a functional copy of the gene. In this case there might not be any paralogs of the gene. If a gene is mutated and loses function (becomes a pseudogene) the functionality of that gene is lost to the bacterium.

The psuedogene identification components of this module are among the most complicated to interpret in the annotation exercises. We will first do a translation of the raw sequence of your gene and compare that to the amino acid sequence of your gene in Genbank. A somewhat quirky aspect of the Genebank record for a pseudogene is that in the Genbank record the mutation causing the gene to be a pseudogene is "corrected" and the fully functional version of the protein is listed. The gene may be annotated as a pseudogene, but the rationale for the annotation is not clearly stated. Comparing the translation of the raw sequence to that of the amino acid sequence will allow us to quickly determine whether a premature stop codon or frameshift (http://en.wikipedia.org/wiki/Frameshift_mutation) has occurred. If you find either you have evidence that your gene is potentially a pseudogene. **Everyone should compare the translation of the raw sequence to that of the amino acid sequence on your gene page. Your instructor may or may not have you perform the more detailed analyses that follow.**

Procedures

Comparison of translation of raw DNA sequence to that of the amino acid sequence under investigation (everyone should do this).

1. Navigate to Gene Search page of IMG/M (https://img.jgi.doe.gov/cgi-bin/m/main.cgi) and perform a search using the locus tag for the gene you are annotating as you did when performing the alternative open reading frame analysis (Figure 7.12).

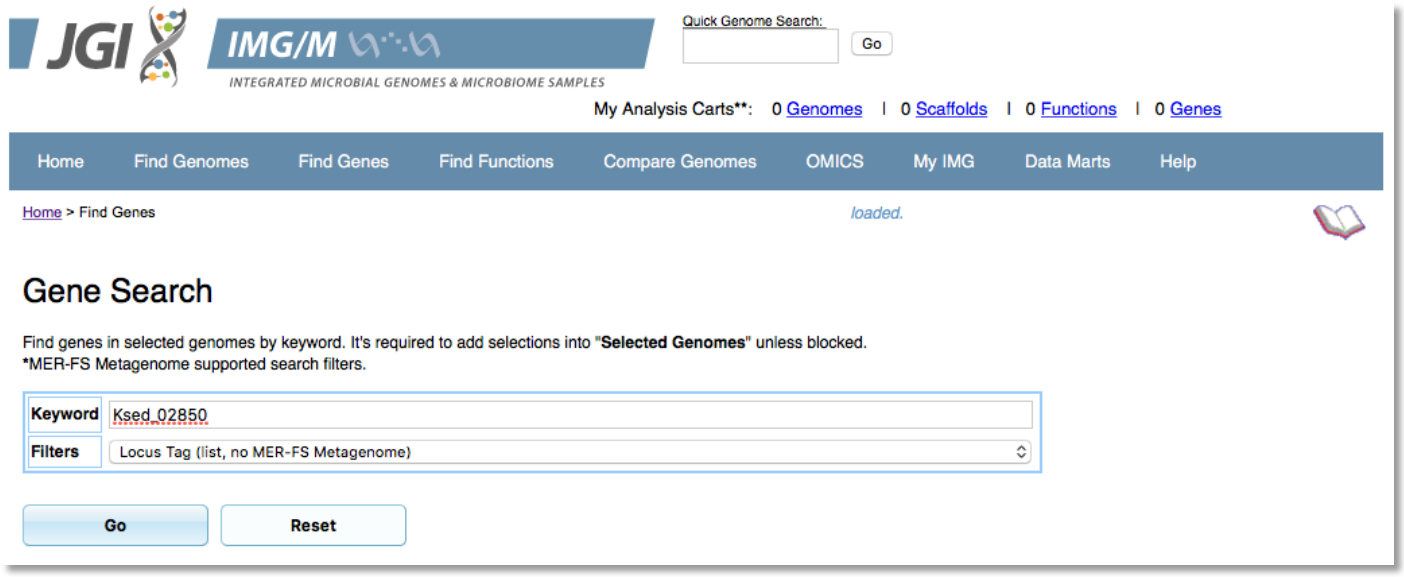


Figure 7.12. The Gene Search window at IMG/M.

1. In the Gene Search window paste the locus tag for your gene in the keyword box, select Locus Tag (list) from the filters pull down menu (Figure 7.12) and click Go. We will use Ksed__02850 for this example.

2. The page that will appear is called the Gene Details page in the IMG Database (Figure 7.13).

Figure 7.13. The IMG Gene Details Page. Only the uppermost portion of the page is shown.

1. Scroll down to the Evidence for Function Prediction section. Click the hyperlink for "Sequence Viewer for Alternate ORF Search" as shown by the arrow in Figure 7.14.

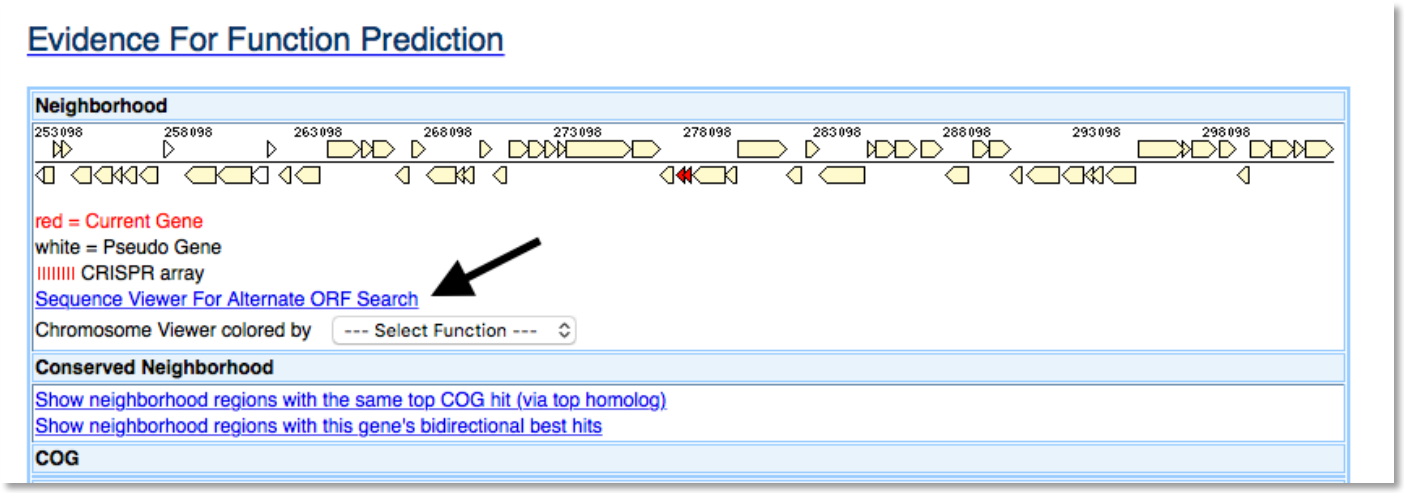


Figure 7.14. Sequence Viewer for Alternate Open Reading Frame Search. Click the link indicated by the arrow page.

1. Select "Text" and **Do not** add any nucleotides upstream or downstream to the sequence viewer (Figure 7.15).

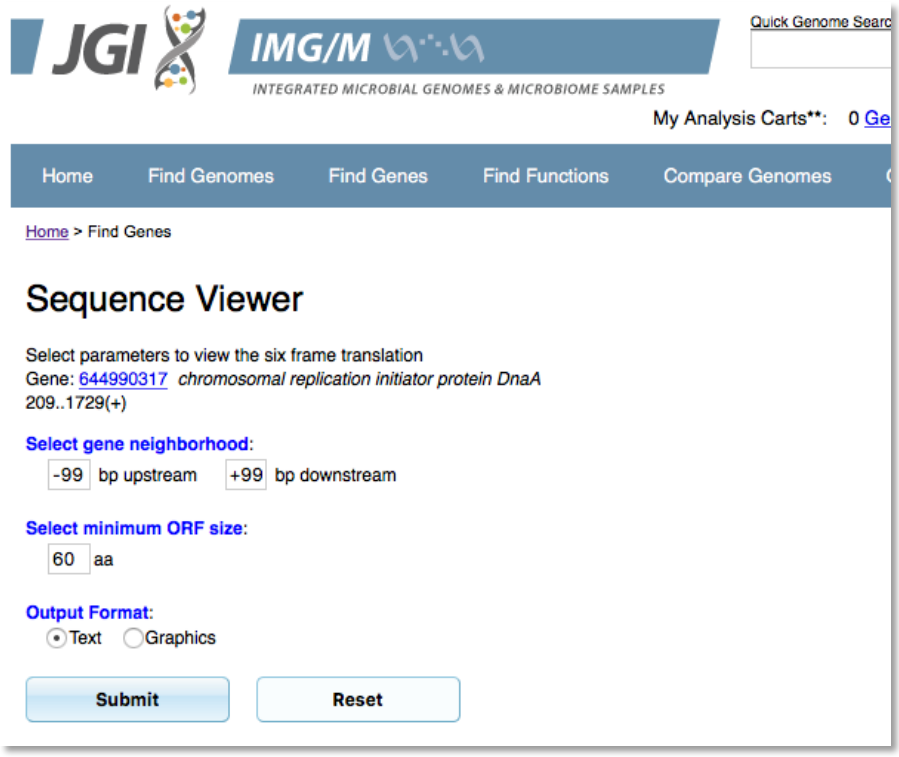


Figure 7.15. The sequence viewer for alternate open reading frame search. Unlike when looking for an alternate open reading frame in Module 5, no additional nucleotides should be added to the upstream or downstream neighborhood. Select the "Text" option and click "Submit"

2. The first reading frame in the results window will be the raw translation of the DNA sequence obtained during sequencing (Figure 7.16).

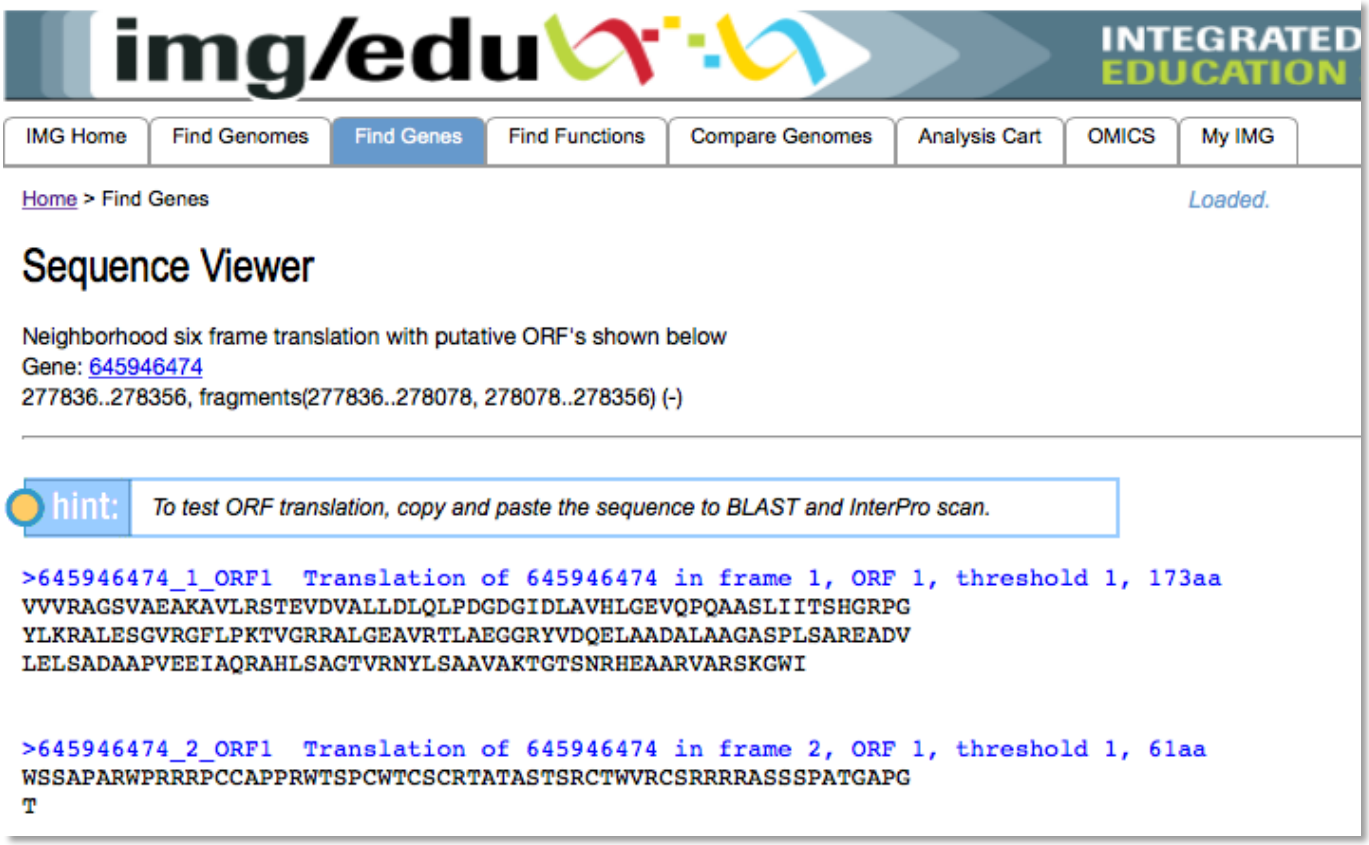


Figure 7.16. The raw nucleotide sequence translation result for Ksed_02850. Note that other reading frame translations were present but not shown in this figure. The first sequence on the page is the only one that is important in this analysis. Also

3. Copy the FASTA formatted translation from the sequence viewer page. We will now align the raw translation of the DNA sequence with the amino acid sequence record in Genbank (the one that is found on your gene information page.

4. Navigate to the BLAST2SEQ page on NCBI at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=blasttab

5. Enter the FASTA formatted amino acid sequence of your gene sequence under Enter Query Sequence and the FASTA formatted translation from the sequence viewer page under Enter Subject Sequence (Figure 7.17) and then click "BLAST."

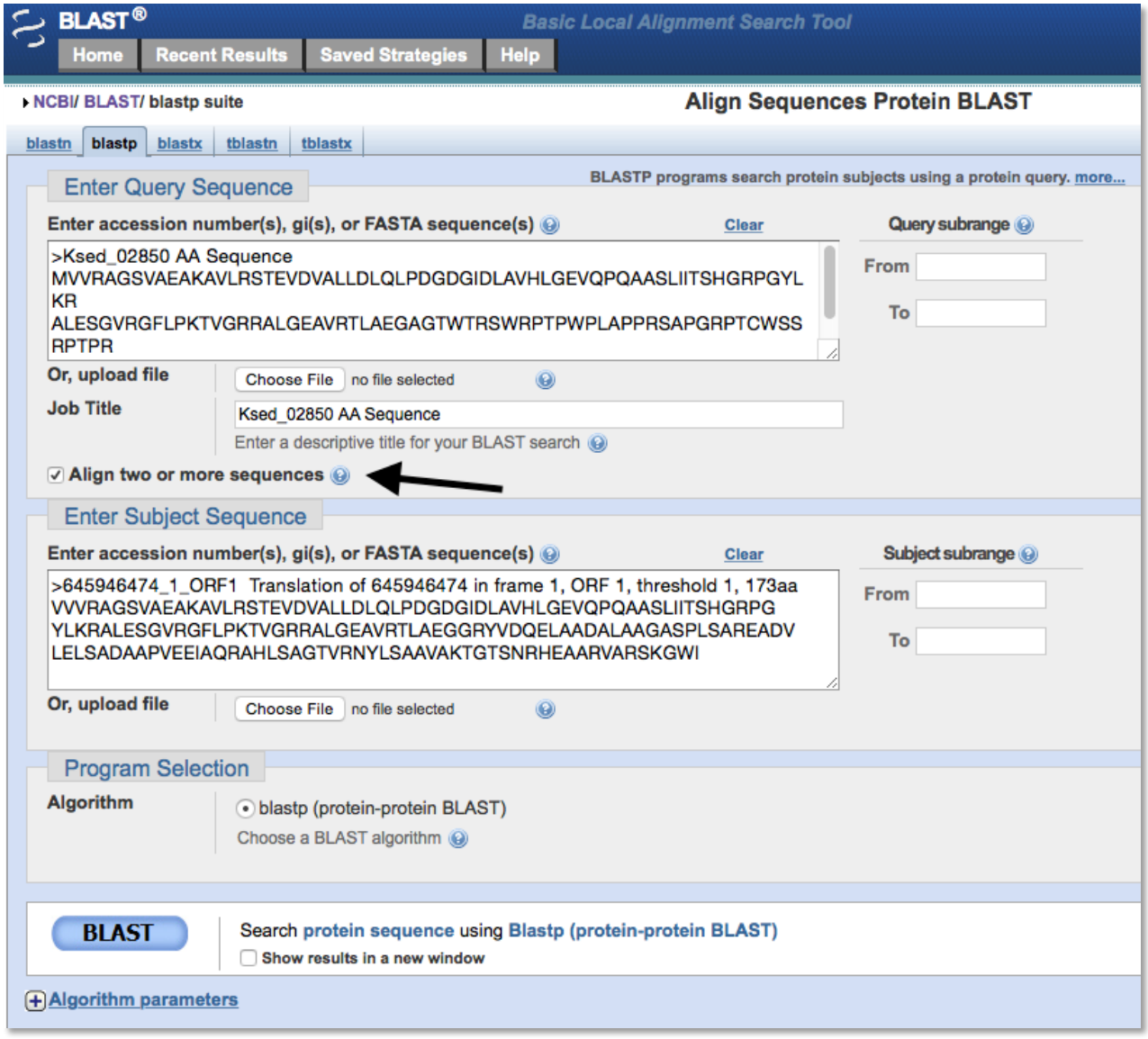


Figure 7.17. The BLAST2SEQ page. This page can be accessed from the blastp page at any time by selecting the "Align two or more sequences" box (arrow).

6. The BLAST results of the raw translation of the DNA coordinates encompassing Ksed_02850 and the amino acid sequence in Genbank for Ksed_02850 are shown in figure 7.18. If the two sequences are identical you will see a perfect match along the length of the alignment. However, the result for Ksed-02850 shows that the match is not perfect. The match is perfect up to the point where the arrowhead in Figure 7.14 indicates a frameshift has occurred.

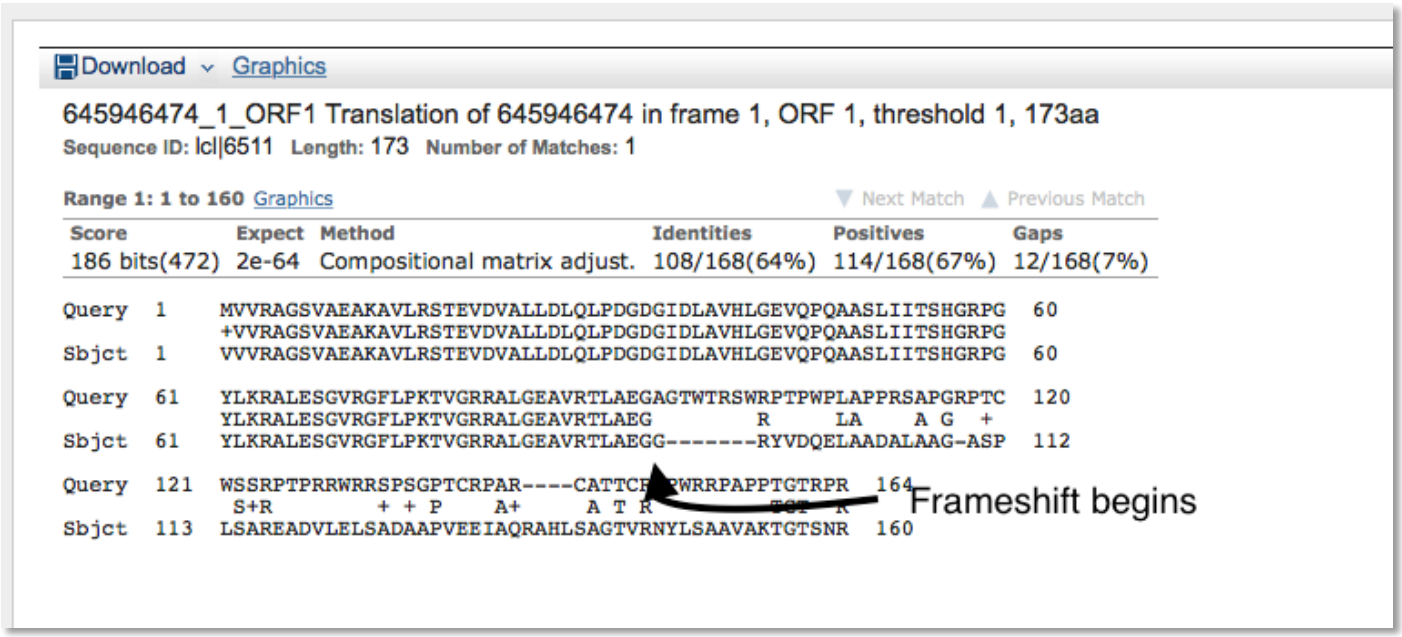
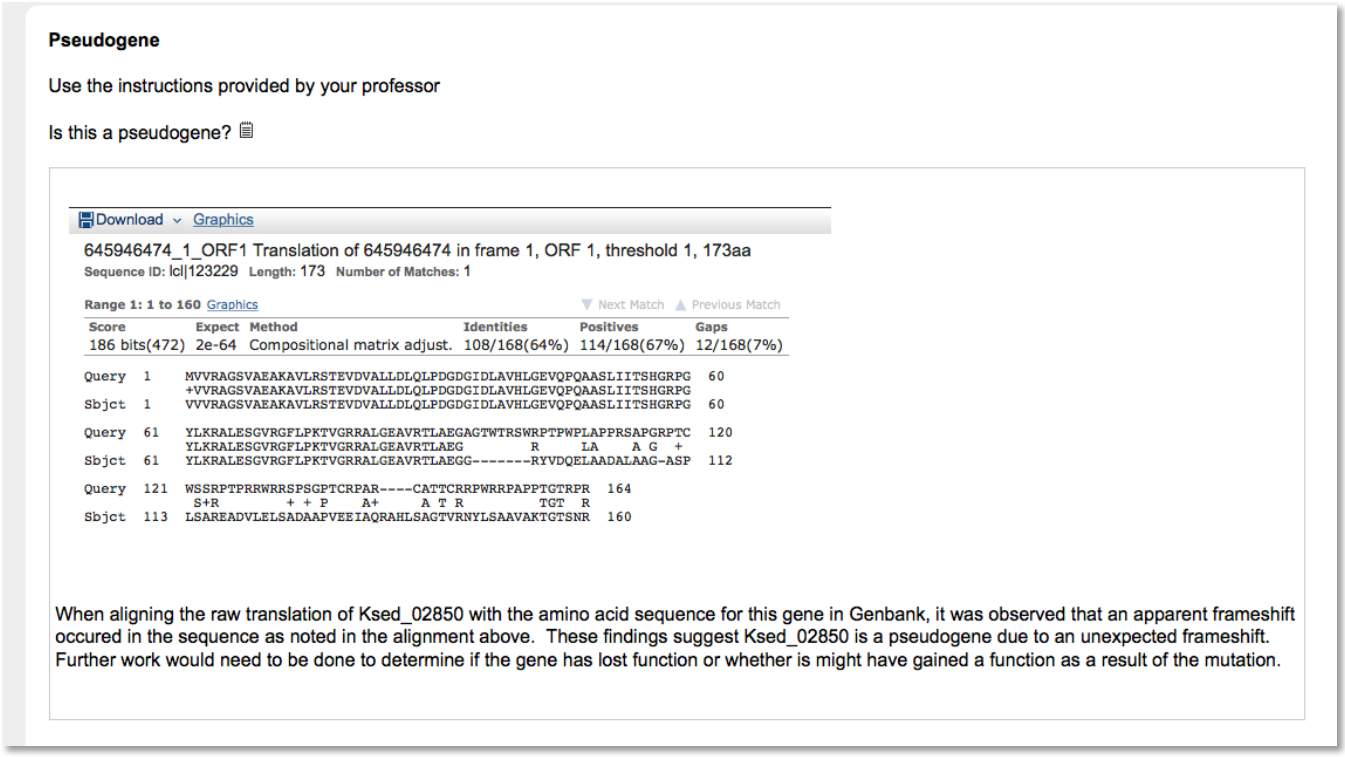


Figure 7.18. The results of a 2 sequence blast of the amino acid sequence in the Genbank record of Ksed_02850 with the amino acid sequence determined from the actual DNA sequence of Ksed_02850. The arrowhead points to the position in the alignment where an apparent frameshift mutation has occurred, thus suggesting Ksed_02850 is a pseudogene.

7. Make an entry in the lab notebook about your findings. Figure 7.19 illustrates the pseudogene notebook for Ksed_02850 with the alignment above entered along with some text to interpret the results. State that there is no evidence to support that gene you are working on is a pseudogene if you do not find evidence of a premature stop codon or frameshift in your protein. **Note that your gene may STILL be a pseudogene without showing signs of a premature stop codon or frameshift if it has lost any key amino acid residues due to point mutations changing one critical amino acid to another, or if there has been an insertion or deletion of amino acids in the same reading frame.** Such changes can be subtle and difficult to determine purely by computational tools (see below).



Pseudogene

Use the instructions provided by your professor

Is this a pseudogene?

Download Graphics

645946474_1_ORF1 Translation of 645946474 in frame 1, ORF 1, threshold 1, 173aa

Sequence ID: lcl|123229 Length: 173 Number of Matches: 1

Range 1: 1 to 160 Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
186 bits(472)	2e-64	Compositional matrix adjust.	108/168(64%)	114/168(67%)	12/168(7%)

```
Query  1    MVVRAGSVAEAKAVLRSTEVDVALLDLQLPDGDGIDLAVHLGEVQPQAASLIITSHGRPG  60
            +VVRAGSVAEAKAVLRSTEVDVALLDLQLPDGDGIDLAVHLGEVQPQAASLIITSHGRPG
Sbjct  1    VVVRAGSVAEAKAVLRSTEVDVALLDLQLPDGDGIDLAVHLGEVQPQAASLIITSHGRPG  60

Query  61   YLKRALESGVRGFLPKTVGRRALGEAVRTLAEGAGTWTRSWRPTPWPLAPPRSAPGRPTC  120
            YLKRALESGVRGFLPKTVGRRALGEAVRTLAEG        R     LA    A G  +
Sbjct  61   YLKRALESGVRGFLPKTVGRRALGEAVRTLAEGG-------RYVDQELAADALAAG-ASP  112

Query  121  WSSRPTPRRWRRSPSGPTCRPAR----CATTCRRPWRRPAPPTGTRPR  164
             S+R        + + P    A+     A T R          TGT  R
Sbjct  113  LSAREADVLELSADAAPVEEIAQRAHLSAGTVRNYLSAAVAKTGTSNR  160
```

When aligning the raw translation of Ksed_02850 with the amino acid sequence for this gene in Genbank, it was observed that an apparent frameshift occured in the sequence as noted in the alignment above. These findings suggest Ksed_02850 is a pseudogene due to an unexpected frameshift. Further work would need to be done to determine if the gene has lost function or whether is might have gained a function as a result of the mutation.

Figure 7.19. The pseudogene GENI-ACT notebook page with an example entry for Ksed_02850. The alignment that suggests a frameshift has occurred has been added along with a brief explanation of the findings.

Advanced Investigation of Pseudogenes

Your instructor may tell you to skip this section. However, you should feel free to explore the criteria for calling your gene a pseudogene as described below. These exercises will take time and their interpretations are complex. They are presented verbatim as provided to the P.I. by Mr. Seth Axen while he was employed at the Joint Genome Institute in Walnut Creek, CA.

For the sake of identification in genomics, an open reading frame (ORF) is annotated as a pseudogene if it meets one of the following criteria:

1. The sequence is interrupted by more than one stop codon or frameshift so that it corresponds to a truncated Pfam less than 30% of the predicted profile.
2. The sequence is separated by another ORF
3. The sequence is missing key residues known to be required for functionality.

The first possible case is identified as given in the "Criterion 1" section below. The second case is identified through somewhat more complicated methods which are given in "Criterion 2" below. The third and final case requires using a new online resource as shown in "Criterion 3" below.

While annotating, one must keep in mind that there is some disagreement in the scientific community as to the technical definition of a pseudogene, and no consensus has yet been reached. Because of this confusion, many professional annotators improperly annotate some hypothetical proteins as pseudogenes with insubstantial evidence that the ORFs are, in fact, nonfunctional genes. As a student annotator, it is easy to fall into this trap of fallacious reasoning. For example, in a pilot program at UCLA, student annotators annotated a large number of features. Of the features annotated, 16 were predicted by the annotators to be pseudogenes. When the methods below were employed in identification, none of those predicted to be pseudogenes were revealed to be so. This example also demonstrates the fact that an actual pseudogene is usually very rare.

It should also be noted that the three criteria given above identify pseudogenes only from a theoretical genomics perspective. **Confirmation in a wet-lab is still required before a sequence can be known as a true pseudogene.** In fact, several sequences which have been annotated as pseudogenes have been verified to be functional in wet-lab experiments. The first and most famous example of a functional pseudogene was identified in 2003 by Hirotsune et al in yeast. They found that an untranslated RNA form of the pseudogene had a critical role in the regulation of the original copy of the gene. Since their research was published, several other such examples of functional pseudogenes have been presented, causing many scientists to question the assumption that pseudogenes are prime examples of "junk DNA."

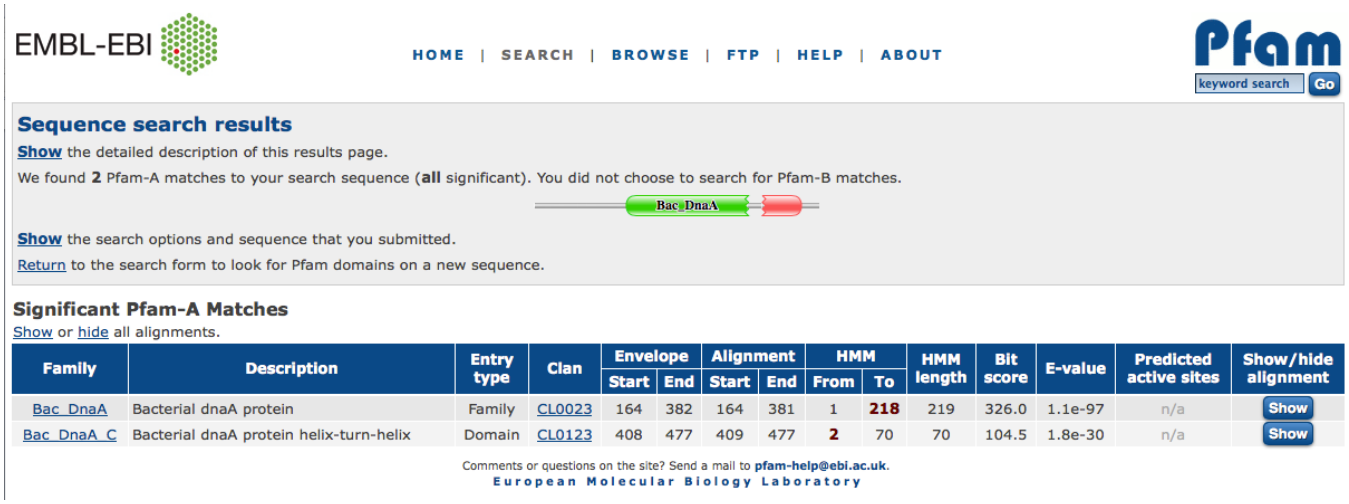
Methods:

Note: The tutorial was designed to be applicable to Mozilla Firefox. You may use any browser or operating system that you prefer. However, some of the following steps may not be performed under any other conditions.

CRITERION 1: The sequence is interrupted by more than one stop codon or frameshift so that it corresponds to a truncated Pfam less than 30% of the predicted profile.

1. Navigate to the Pfam database at http://pfam.xfam.org/.
2. Search the amino acid sequence of your amino acid against the Pfam database as explained in the Pfam portion of the Structure Based Evidence module.

3. On the results page, note the domain graphic (Figure 7.20). If this is a pseudogene of Criterion 1, then, the domain graphic will show the last domain as truncated.



Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Bac_DnaA	Bacterial dnaA protein	Family	CL0023	164	382	164	381	1	218	219	326.0	1.1e-97	n/a	Show
Bac_DnaA_C	Bacterial dnaA protein helix-turn-helix	Domain	CL0123	408	477	409	477	2	70	70	104.5	1.8e-30	n/a	Show

Figure 7.20. Pfam results for Ksed_00010, as determined in the Structure Based Similarity Module.

4. Scroll down to the data table and observe the row for that sequence. Note the length of the HMM covered by the sequence given by under the HMM columns subtracting the "To" number from the "From" value and adding 1. In the example above, this would be 218-1+1 = 218 for the Bac_DnaA Pfam and 70-2+1 = 69 for the Bac_DnaA_C Pfam.
5. The total length of the HMM can be seen in the column titled HMM length (figure 7.20). For Bac_DnaA, this value is seen to be 219.
6. Divide the value recorded in step 4 by the value in step 5 and multiply the resulting number by 100% to obtain the percent coverage of the entered sequence. If this value is less than 30% and research of the literature indicates that the domain is necessary for protein functionality, then the protein is a pseudogene meeting Criterion 1. In the example above for Bac_DnaA, the value would be 218/219 x 100 = 99.5%. Thus Ksed_00010 would not be classified as a pseudogene by criterion 1.

CRITERION 2: The sequence is separated by another ORF

1. Criterion 2 is an extension of comparing the open reading frame of the raw sequence data to that predicted for that of your protein in Genbank that you performed above. If you found no evidence of a premature stop codon or frameshift from doing that analysis then this criterion would NOT apply to your protein.

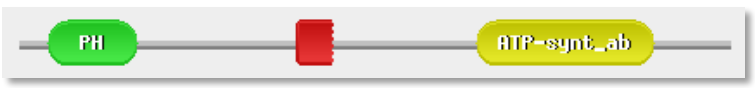


Figure 7.21. A Pfam domain graphic showing a possible interrupted reading frame of the domain shown in red.

2. Navigate to the Pfam database and search the amino acid sequence of your protein against the Pfam database as described above.
3. On the results page, note the domain graphic. If an inserted ORF maintains its reading frame and its stop codon is intact, a pseudogene of Criterion 2 will usually show one or more domains after a truncated domain of the predicted protein (Figure 7.21).
4. Navigate to IMG/M and access the Gene Details page for the gene on which you are working as you did at the start of this module.
5. Under "Evidence for Function Prediction," click on "Sequence Viewer for Alternate ORF Search."
6. If the truncation on the domain graphic in Pfam is on the right as in figure 7.18 above, on the "Sequence Viewer" page, change the value in the "bp downstream" box from "+0" to a number such as "+100." If the domain graphic showed the truncation symbol on the left, you would add the extra sequence to the bp upstream box.
7. Press "Submit." Locate the flanking nucleotide sequence that you added upstream or downstream of your gene sequence (which is colored green).
8. Paste the flanking sequence into a text file. Extra spaces between nucleotides may appear after pasting into the text file. These will be removed in step 10 below.
9. Copy the nucleotide sequence of your gene that was recorded in the basic information module and paste it in front of (if the flanking sequence is downstream of the gene) or behind (if the flanking sequence is downstream of the gene) the flanking DNA sequence in the text file.
10. Extra spaces can be removed from the sequence in the text file. Copy the combined from the text file and navigate to: http://www.bioinformatics.org/sms2/filter_dna.html. Paste the sequence into the text box and hitting submit.
11. Run a Pfam on the nucleotide sequence from step 10 by searching at http://pfam.sanger.ac.uk/search?tab=searchDnaBlock.
12. The results page will show domain graphics in all reading frames. If the inserted sequence maintains the reading frame of the protein throughout the sequence, the domain graphic will appear on one line as in figure 7.19 below. While the inserted sequence might not necessarily have a domain which is identified on Pfam, there will be a visible fragmentation of the domain from the original protein. What is important to gather here is that the second half of the fragmented domain is present in the flanking DNA (red domain in figure 7.22). If this is the case, then the feature is a pseudogene meeting Criterion 2.

Figure 7.12. Pfam domain graphic showing an inserted open reading frame with a Pfam domain (yellow) splitting the Pfam domain of the gene under investigation into two pieces (red). The red and yellow domains will only appear this way if the reading frame is maintained through the insertion of the second open reading frame.

13. If the open reading frame is not maintained throughout the sequence, the left and right parts of the Pfam domain graphic may appear in one reading frame and the rest in another reading frame.

14. If your gene has either the arrangement shown in step 12 or discussed in step 13, you should report that your gene might be a pseudogene based on criterion 2.

CRITERION 3: The protein lacks amino acids critical for its function.

ScanProsite is a database of protein domains, families, and most importantly for the detection of psuedogenes, functional sites. It contains a companion tool called ProRule that provides additional information about amino acids that are structurally or functionally critical in a fully functioning domain. It is, in a way, similar to TIGRFAM, Pfam and Cog searches that were performed in earlier modules. However, we will only use it to see if your protein possesses or lacks residues that are critical to the domain functions, not to re-identify domains that you have already documented. Your protein can be hypothesized to be a pseudogene if it has a mutation that results in a different amino acid replacing an amino acid predicted to be essential for the function of the domain present in your protein. You will use the amino acid sequence of your protein to search ScanProsite for functional domains that exist in your protein, and then use the resulting output to see if your protein lacks amino acids deemed critical for the function of that domain.

1. Navigate to the ScanProsite tool on Prosite at http://expasy.org/tools/scanprosite.

2. Enter the amino acid sequence of the protein in question into the box under "Step 1-Submit protein seqence" (Figure 7.23). Check the "Exclude motifs with a high probability of occurrence" box (leaving this unchecked results in a number of hits for protein post-translational modifications, which are not relevant to prokaryotes), check the "Show low level score" box, and click "START THE SCAN."

Figure 7.23. The ScanProsite tool start page. The amino acid sequence of Ksed_00010 has been pasted into the search window.

3. Figure 7.21 shows the top half of the results page from the search started from Figure 7.20. At the top is the sequence of Ksed_00010, followed by a legend that would identify features in the sequence. For example, if there was a known active site in Ksed_00010, the amino acids in the site would be colored red. Below is a domain graphic (profiles in this results page). Hovering over the graphic causes the sequence in Ksed_00010 to be highlighted yellow. In figure 7.24, the yellow highlighting comes from

moving the cursor over the graphic indicated by the arrow, illustrating on the amino acid sequence the region the domain covers. A full description of the results page can be found at: http://prosite.expasy.org/scanprosite/scanprosite_doc.html - of_miniprofiles.

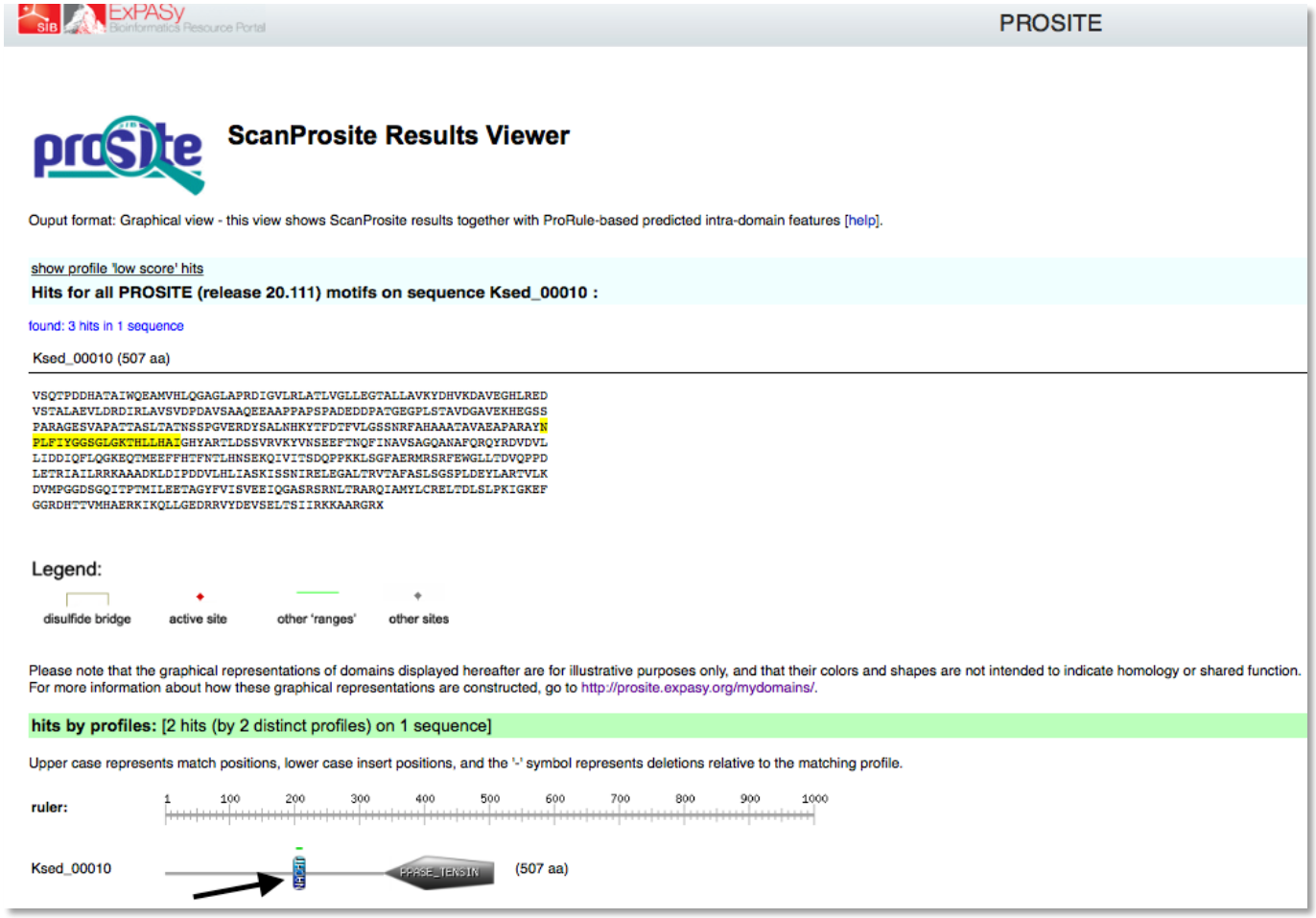


Figure 7.24. Part one of the search results for Ksed_00010 in ScanProsite. See text for a discussion of the results.

4. Figure 7.25 shows the bottom half of the results page for searching with Ksed_00010 in ScanProsite. There are two important notations to watch for when looking at the results for a ScanProsite search page. When hits for only one protein are shown, and if you have a Mozilla based web browser (Mozilla, FireFox,) you'll be able to see feature residues highlighted (green for predicted features, gray for absent features) on both the match and the full protein sequence (if shown) when you move your mouse cursor over a feature line. The two domain profiles in the graphic have corresponding text descriptions as indicated by the arrows in Figure 7.25. Also shown is a pattern hit and graphic for DnaA, a domain found previously in the Sequence and Structure Based Similarity Modules. The two profiles have a warning that they are low confidence level matches. The sore of -1 indicates a low confidence match, while a score of 0 indicates a higher confidence match (see DnaA pattern score). However the PPASE_TENSIN illustrates what might be seen if your gene was a pseudogene.

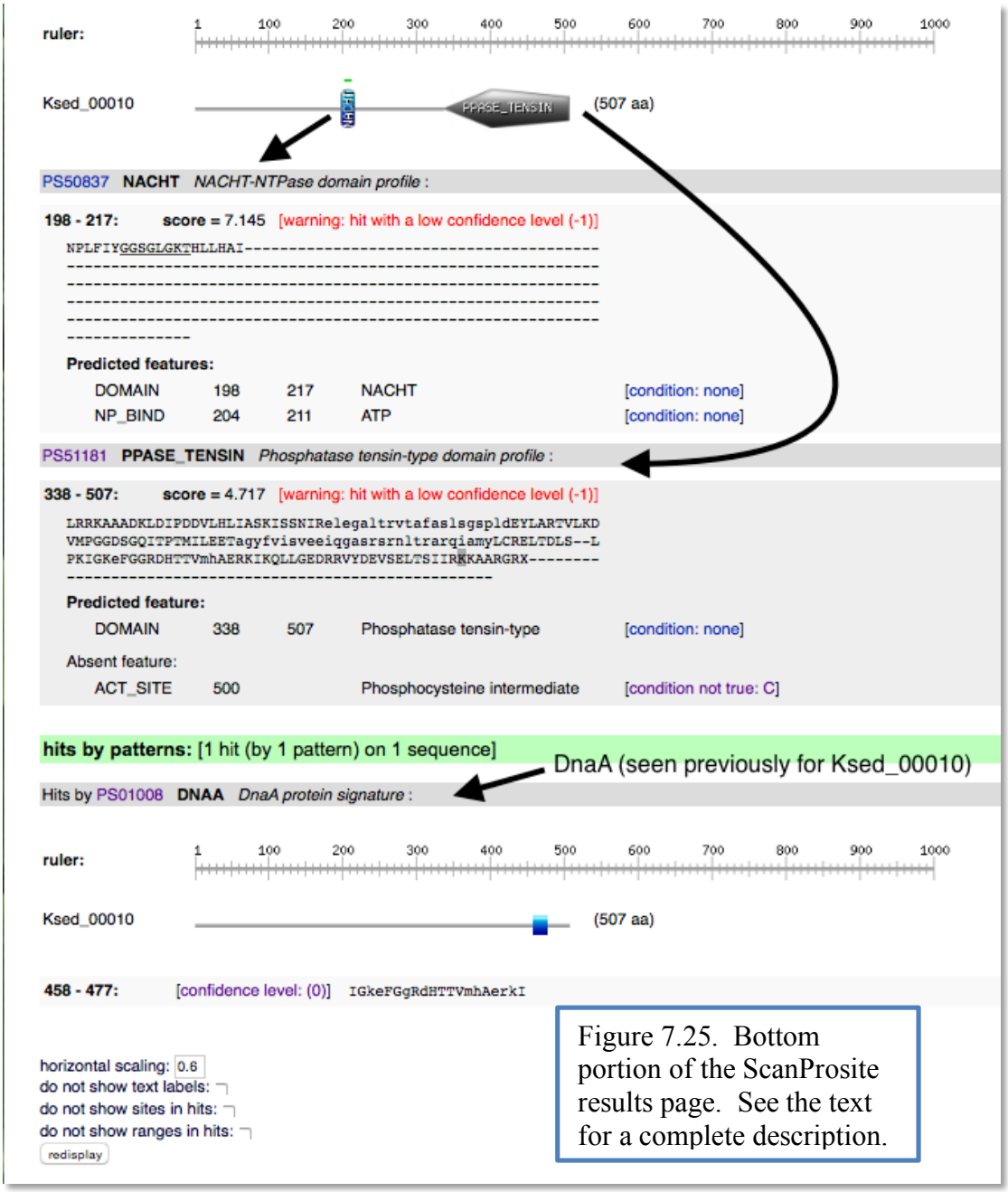
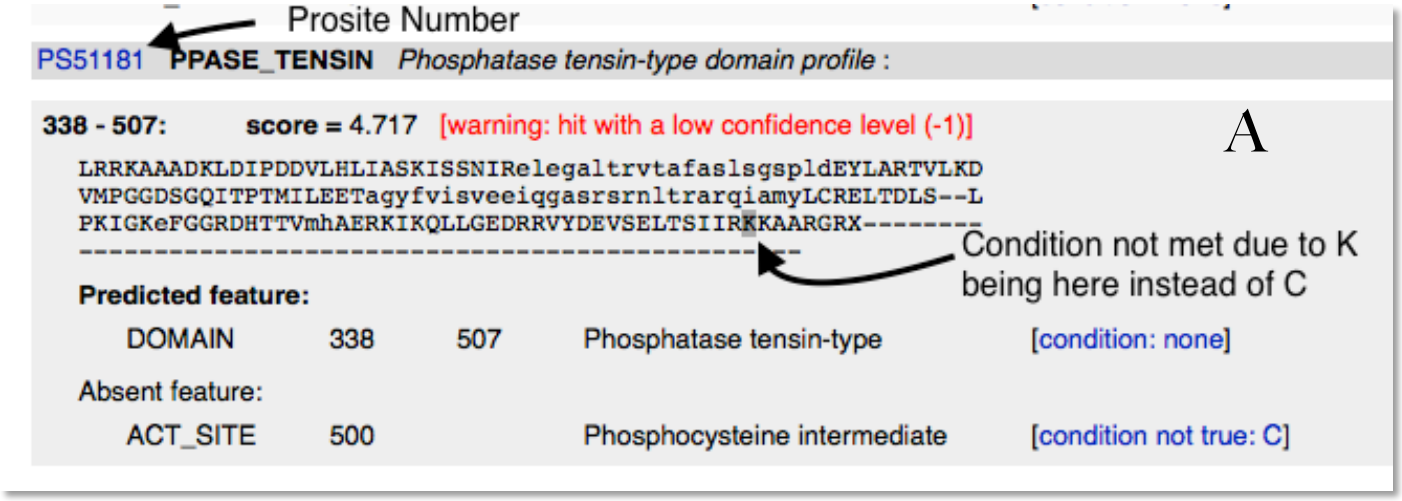


Figure 7.25. Bottom portion of the ScanProsite results page. See the text for a complete description.

5. As shown the Figure 7.26A, and expanded version of the PPASE_TENSIN result from Figure 7.25, when any one of the amino acids in an active site or a predicted feature does not meet the condition required for assumed functionality, Prosite will no longer underline or color the residues but will change the condition section to say [condition not true: X]. Other messages are possible, and clicking on the highlighted condition will take you to the rule that explains the statement (Figure 7.26B). Placing the cursor over these conditions will highlight grey the altered residue(s) in the domain's sequence in the same box. However, it will always highlight the altered residue(s) in the complete sequence at the top of the page. When scrolling up, take care that the cursor does not hover over any of the other hits or those hits will be highlighted instead.



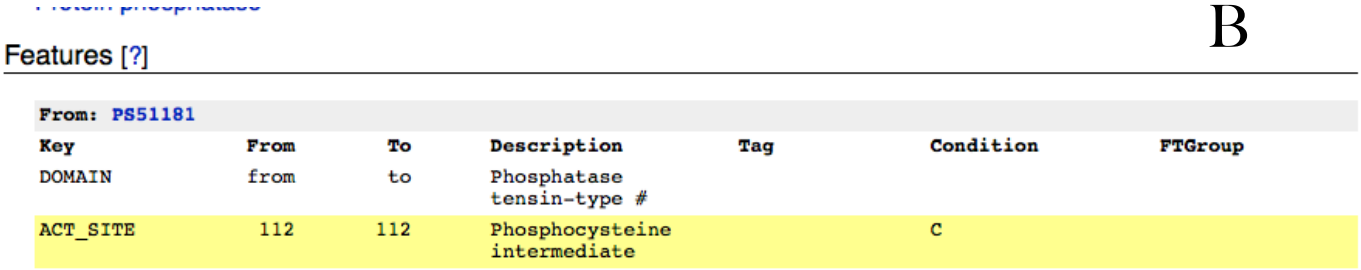


Figure 7.26. A. The enlarged PPASE_TENSIN domain profile from figure 7.22. It showw a highlighted K in the sequence for which the "condition not true :C" rule applies. B. Clicking on the hyperlink for "condition not true: C" in figure A yields the feature description in this figure. It states that a C is required in the active site at amino acid position 112.

6. If a condition is not met, that provides strong evidence that the protein is a pseudogene. However, one more step must be taken to confirm that exceptions to this condition are rare or absent. Click on the PS***** identification number hyperlink (arrow, Figure 7.26A) in the title line for the profile for which the condition is not met. This leads to the information page for that profile (Figure 7.27).

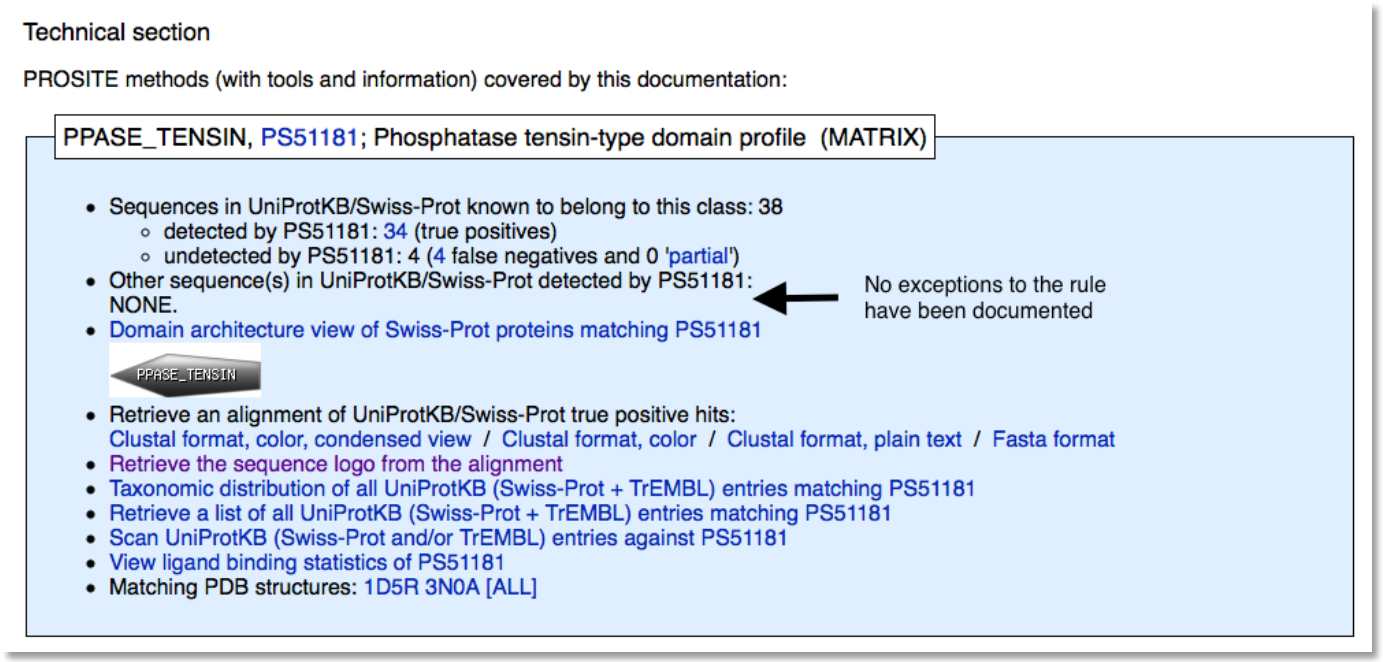


Figure 7.27. The description section from the page linked to the PS51181 hyperlink shown in Figure 7.26A. The arrow indicates that no exceptions to the rule have been documented. Thus, if the condition is not met in a protein domain matching this profile, the protein would be hypothesized not to be able to perform the function encoded by the profile, i.e. it would be a pseudogene.

7. On the profile page, be sure to scan the "Description section" toward the top for any key information concerning structure or catalytic activity. Scroll down to the "Technical section" and look for a data box that contains a grey graphic portraying the profile. Above this graphic is a box with a blue background that usually contains information concerning the condition in the profile. If multiple other sequences are detected in Swiss-Prot, indicating that there are multiple exceptions to the condition, then you cannot yet conclude that the feature is a pseudogene. If the data and the box give few or no exceptions to the condition, then it is safe to conclude that the condition is absolutely necessary for functionality of the protein, and the feature may be considered a pseudogene meeting Criterion 3.

8. Summarize your findings in your notebook and decide whether you have evidence for your gene being a pseudogene.