JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

img INTEGRATED MICROBIAL GENOMES

# Curation of Gene Models

**Igor Bogorad**

Genome Biology Program
Joint Genome Institute

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

## Table of Contents

- Introduction
  - Programs
  - ORFs
  - Annotations
  - Error Report
- 1. Overlaps
  - Different Types
  - Example
- 2. Short
- 3. Frameshift
- 4. Pseudogenes

# Table of Contents


- **Introduction**
  - Programs
  - ORFs
  - Annotations
  - Error Report
- 1. Overlaps
  - Different Types
  - Example
- 2. Short
- 3. Frameshift
- 4. Pseudogenes


# Programs Used

- **Artemis** – from the Sanger Institute is an annotation tool that allows visualization of sequence features
- **BLAST** - NCBI's Basic Local Alignment Search Tool, finds regions of local similarity between sequences

# Open Reading Frames

- There are six open reading frames (ORFs) in a genome
- This means that there are six possible amino acid sequences for the two DNA strands.
- The bottom DNA strand runs complementary to the top one. Thus when reading the bottom three frames sequence, read from right to left.
- In other words, read the sequence in the direction that the strand goes.

# Annotation

- Data from sequencers often have erroneous data that needs to have a human take a closer look.

- We're identifying 3 major types of annotation problems:
  1. **Overlaps-** Two genes shouldn't overlap. One may not be a real gene or have a wrong start site
  2. **Frameshifts-** sometimes there is a broken gene because of a polymorphism
  3. Search for **new features** such as missed genes and pseudogenes

# Error Report

- **Intragenic Regions**
  Two genes are separated by a colon. This represents a possible new gene between them.
- **Overlapping ORFs**
  Overlaps with more than 30 amino acids need to be changed.
- **Dubious Genes**
  These might not code for any proteins
- **5'prime problems**
  The genes flagged as TOO_SHORT may be extended. The ones marked as TOO_LONG might need to be cut.

# Table of Contents


- Introduction
  - Programs
  - ORFs
  - Annotations
  - Error Report
- **1. Overlaps**
  - Different Types
  - Example
- 2. Short
- 3. Frameshift
- 4. Pseudogenes

# 1. The Problem of Overlaps

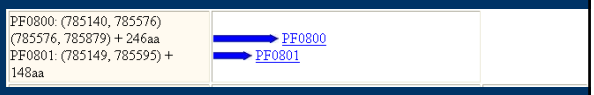
- **If the two lines shown below are genes, then:**
  - Both sequences are examined to see if they match any other homologs in other genomes.
  - These usually have been verified. Only one is viable.
  - When this problem is solved, we edit the existing information.

# Examples of Overlaps

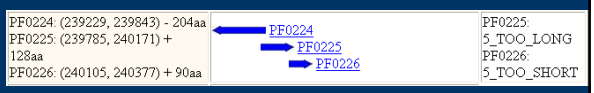
- Mark PF0707 as dubious since it is UNIQUE and overlaps

- Cut PF0597 because it is too long on the 5-prime end.

- BLAST both, see which one has better CDDs and homologs hits.

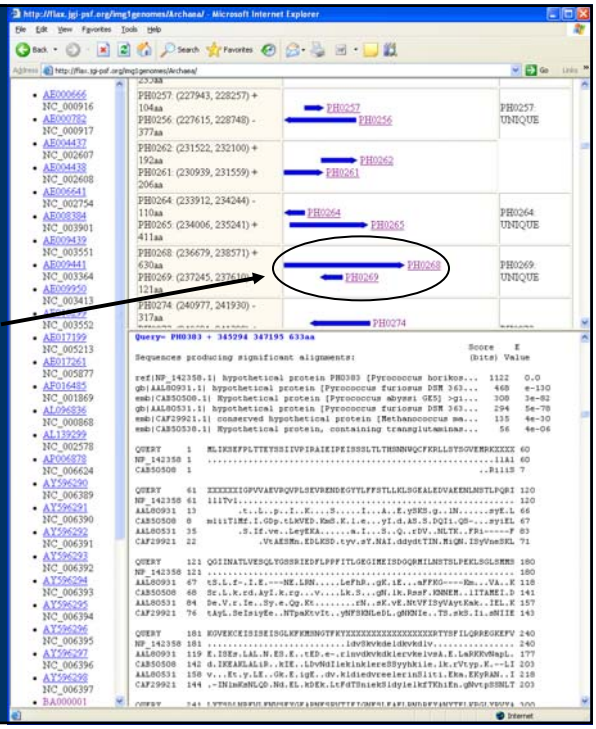



- After BLAST, mark PF0225 as dubious.



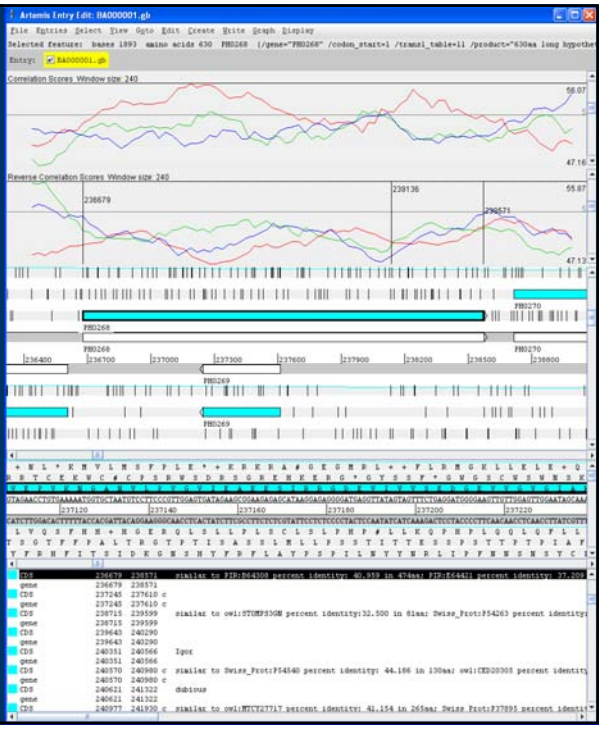
# Error Report

- In archaeal genomes, overlaps between two genes are very rare so one of the gene is dubious.
- This shows all errors, such as overlaps, that the computer was not able to fix when it finished sequencing the genome.
- The left frame shows a list of different genomes currently being edited.
- The top right is a visual representation of the errors.
- The bottom right is a BLAST of all the errors.
- The name of the gene (PH0268) is copied and then the location is found in another program called Artemis.



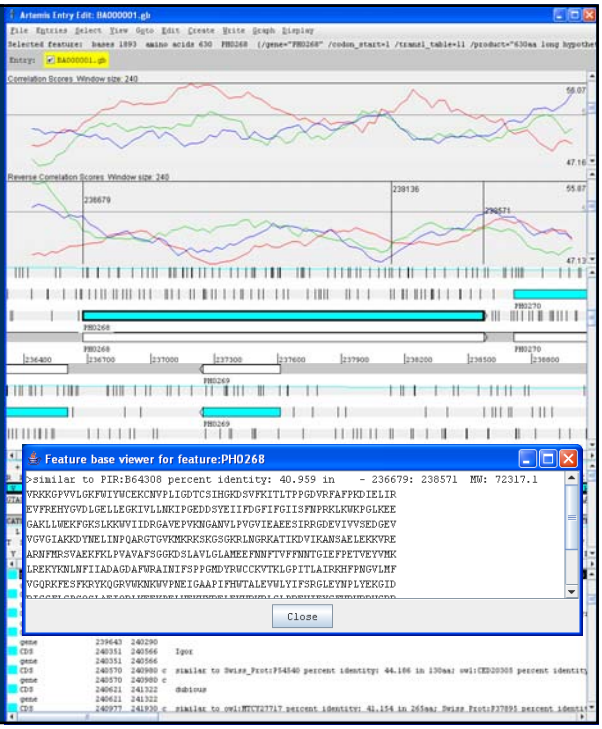
# Analyzing the Problem

- Artemis is used to look at the amino acid and DNA sequence of a genome.
- Correlation scores of the six reading frames shows the average protein distribution across the genome. If a reading frame is above the threshold line, the likelihood that it is a gene increases.
- We need to find out which one is probably the real one. Overall, the small genes with no hits are usually dubious. But there are exceptions.
- The blue highlighted areas are possible genes.
- Our gene is the blue bar that is outlined in black.



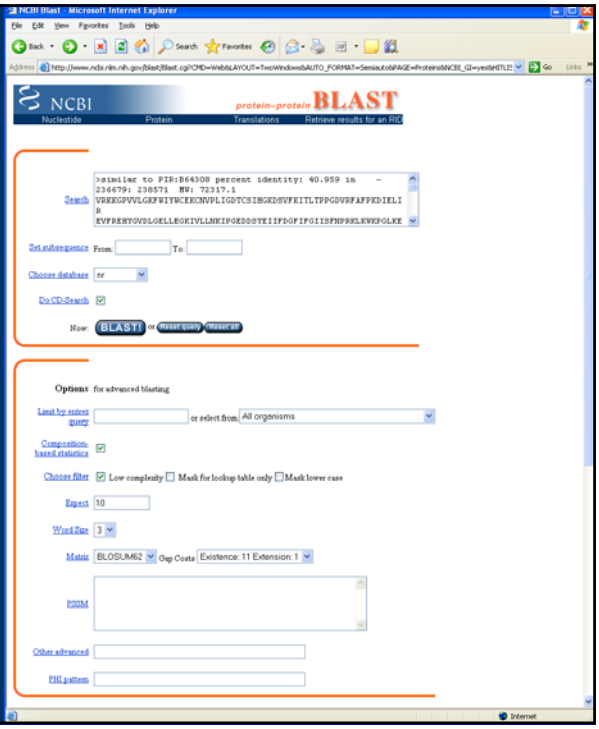
# Viewing the Sequence

- We're first going to look at the amino acid sequence of gene PH0268. This can be copied and analyzed in BLAST.
- PH0269 is selected and the amino acid sequence is shown in the box below. Select the sequence and go to NCBI's BLAST.



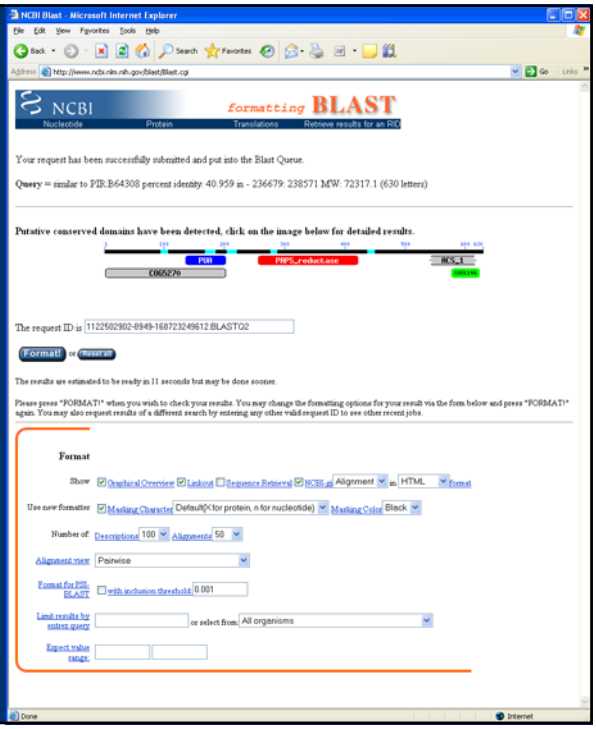

# BLAST the sequence

- Enter the gene sequence in the Search box.



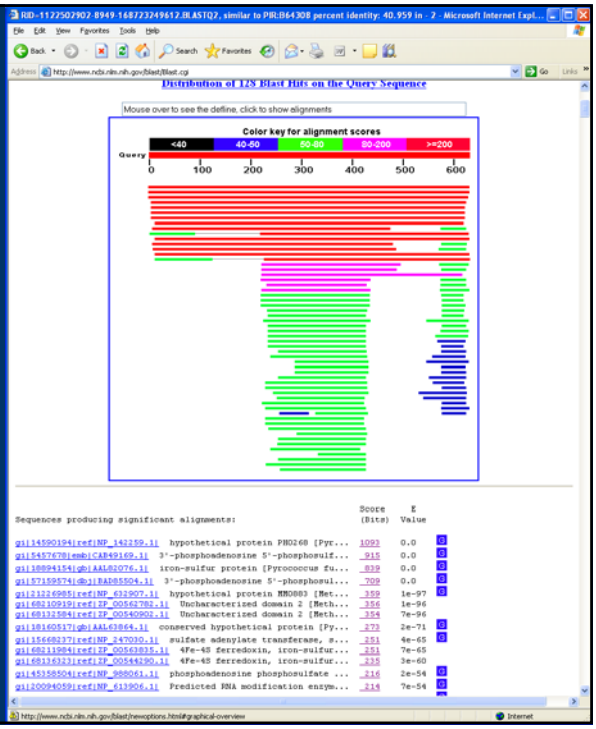
# CDD Image

- The results give good hits with many different protein domains. This means that within the amino acid sequences there are several domains that are similar to other organisms. These are called CDD hits.
- CDD stands for Conserved Domain Database. CDDs are basically classes of proteins. This broad functional search can reveal differences in genomic size and phylogenetic composition
- So of the two sequences, PH0268 is likely to be the real functional gene.



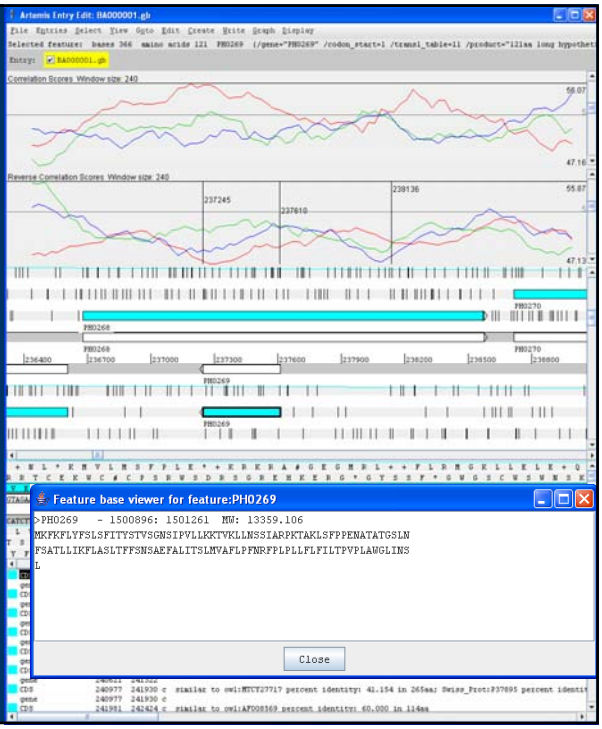

# Analyzing the BLAST

- Formatting allows us to see the homologs. The results give many different hits.
- In red, each line corresponds to a match that was found.
- The matches are listed with the best match (lowest e-value) first and the rest in descending order of their e-value.
- The protein is likely to be 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase



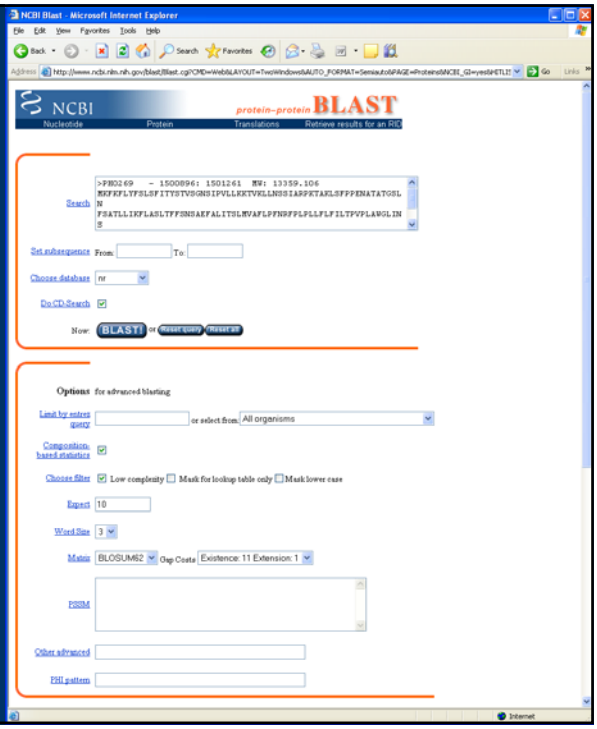
# Viewing the Sequence

- View the sequence of the smaller gene.



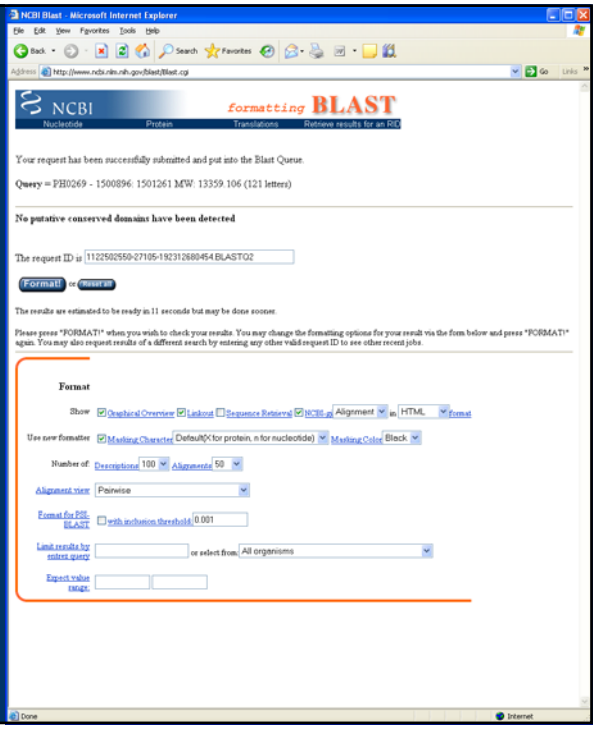

# BLAST the sequence

- Go to NCBI's BLAST webpage.
- Enter the sequence in the Search box.
- The sequence will be compared a database of all known genomes.



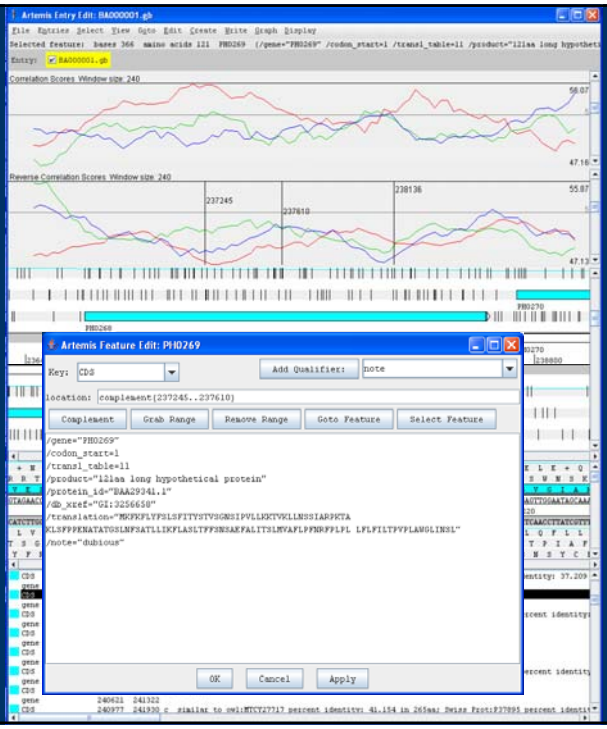
# Analyzing the BLAST

- No conserved domains indicate that the smaller gene does not have any similar homologs in different organisms




# Editing in Artemis

- Ctrl-E enables you to add qualifiers and identify new genes.

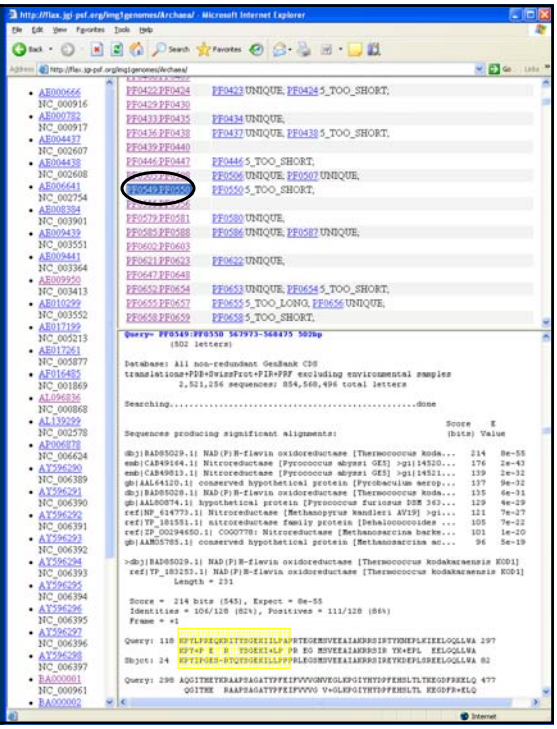


# Table of Contents


- Introduction
  - Programs
  - ORFs
  - Annotations
  - Error Report
- 1. Overlaps
  - Different Types
  - Example
- **2. Short**
- 3. Frameshift
- 4. Pseudogenes

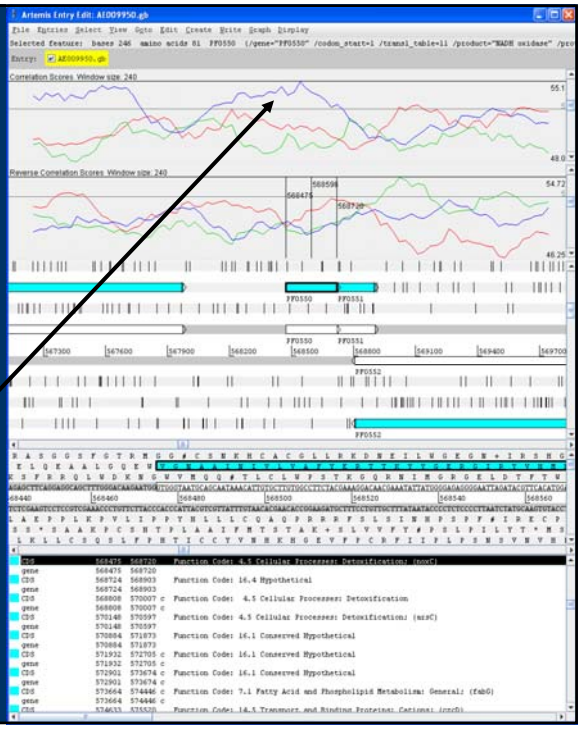

# Error Report

- The error report shows that there is an intragenic region possibly part of a gene.
- The sequence is given at the bottom.
- Copy the sequence KPYLPREQKRITYS GEKIILPAP to find it in Artemis.



# Analyzing the Problem

- In this example we will be working with the archaeum *Pyrococcus furiosus*.
- The selection shows gene PF0550.
- Several indications point that it needs to be extended.
  - Look at the correlation score for this frame (colored in blue). It looks as though there is good information upstream.
  - There are no stop codons in the large open section before the gene.
  - There would be no overlaps with nearby genes if it were extended.

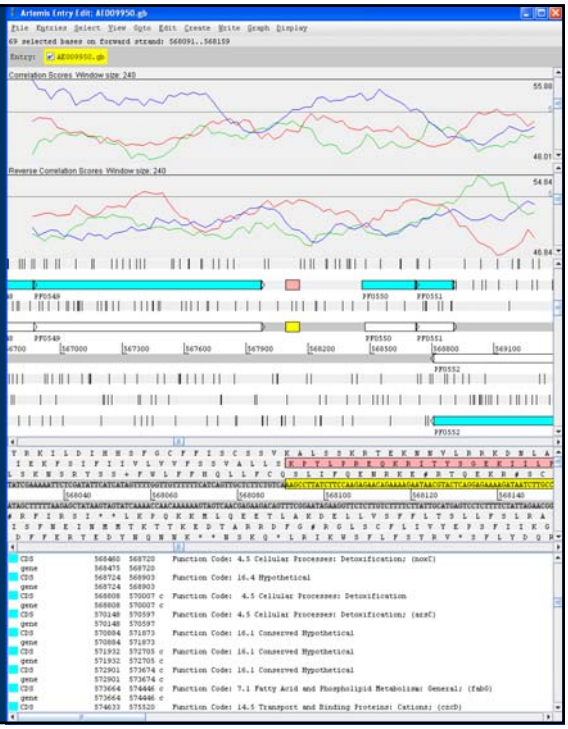


# Artemis Navigator

- Ctrl-G opens the Artemis Navigator.
- This tool enables you to easily find any desired gene, DNA or amino acid sequence, or notes that were added to genes.
- We enter the copied sequence (KPYLPREQKRITYSGEKIILPAP) in Find Amino Acid String.
- Press “Go to”

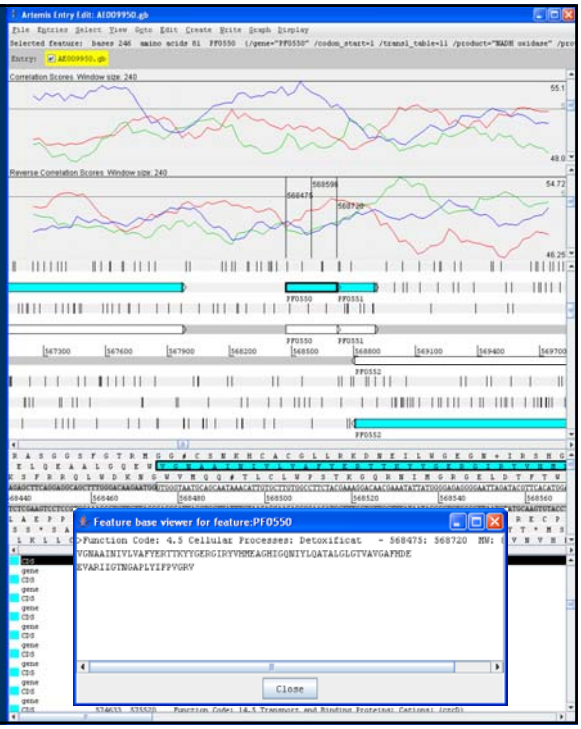
# Find the Missing Region

- Artemis selects the sequence and brings you to its start.
- We know that we can extend the gene if need.
- But first we need to make sure this is necessary, are any CDDs truncated?
- Now BLAST the gene PF0550 to find CDD alignments.



# Viewing the Sequence

- Copy the sequence.



# BLAST the sequence

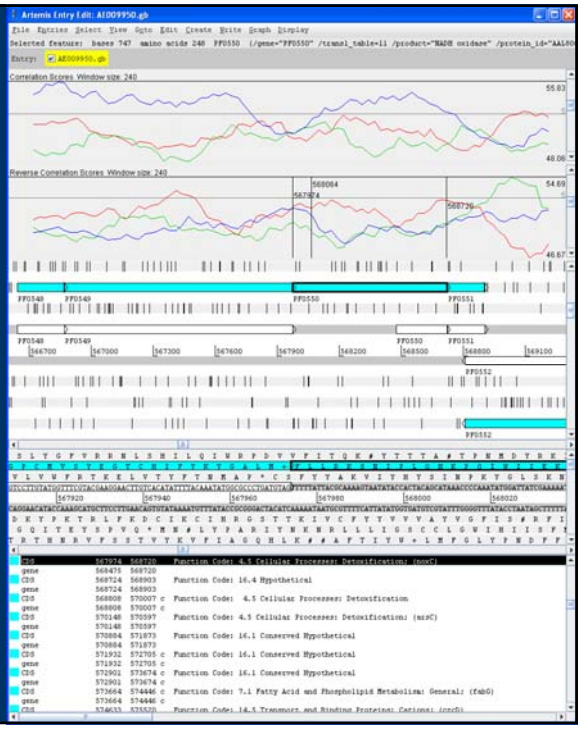
- Paste the sequence in the Search box.

# CDD Image

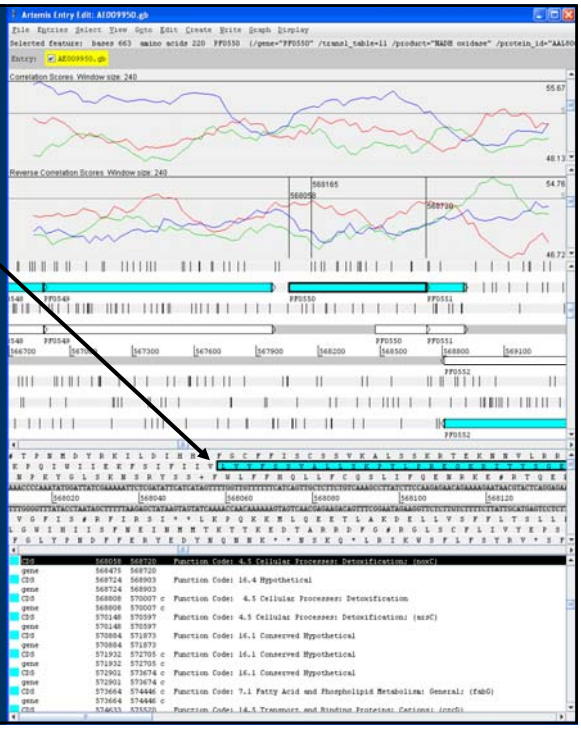
- The conserved domains shown are cut at the beginning.
- We need to extend this gene to include the missing portion.

# Change the Gene

- The first step to extend a gene is to bring it back to the last stop codon. The gene PF0550 is much larger now, but it begins on a Phenylalanine (F) which is not a start codon.
- We need to move the gene to the next start codon.



- The next start codon begins on a Leucine.
- Now BLAST the sequence to see if the change improved the CDD hit.



# View the Modified Gene

- Copy the sequence.

# BLAST the sequence

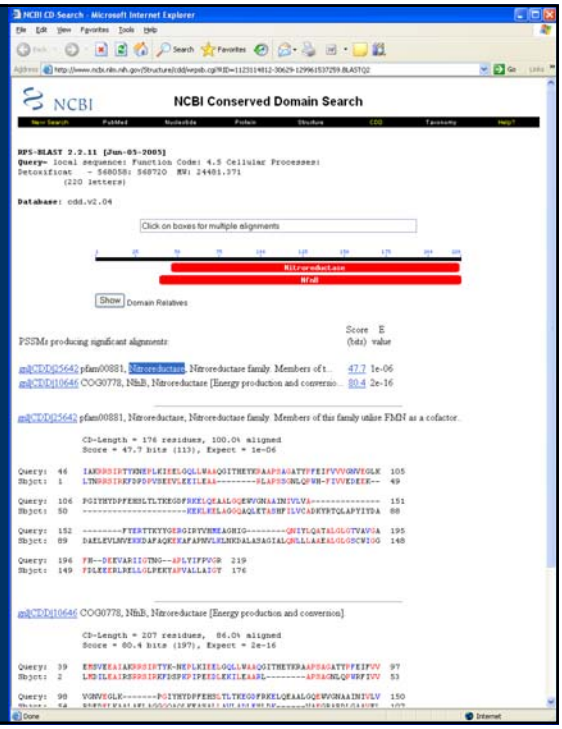
- Click on BLAST.

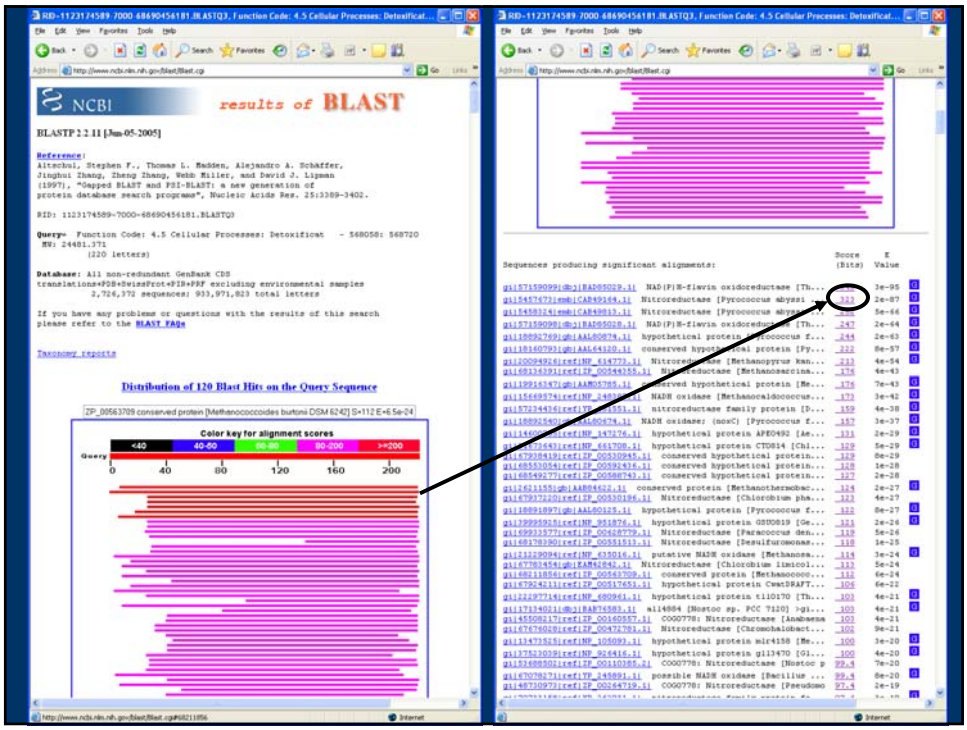
# CDD Image

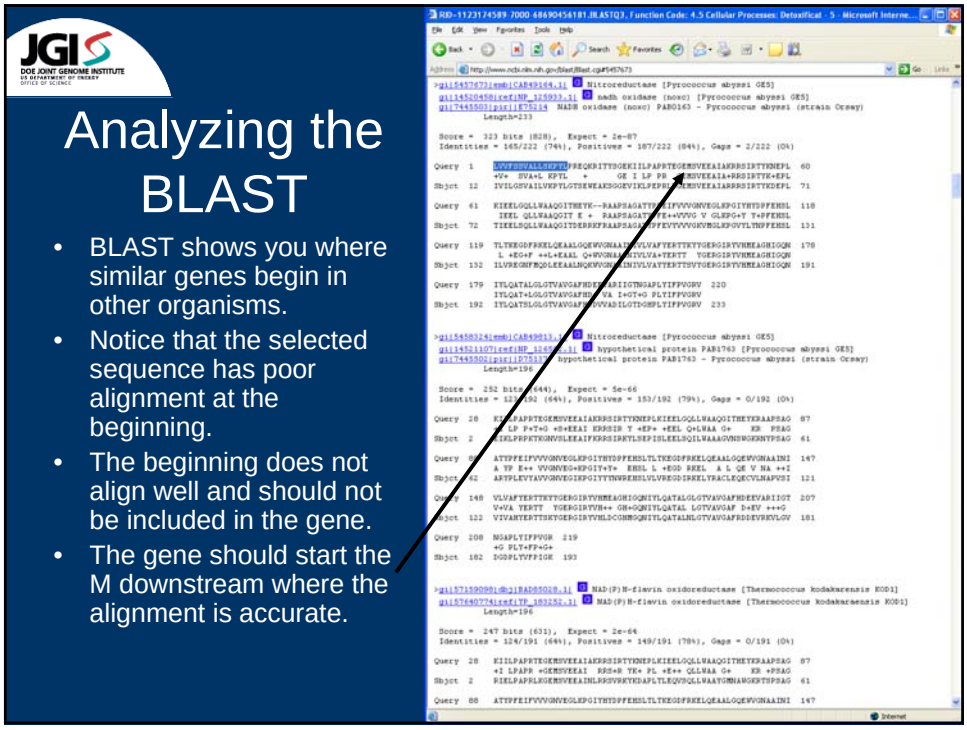
- Notice the Nitroreductase domain is now complete.
- Yet the CDD begins about 50 amino further down. Maybe we need to cut the gene.
- Clicking on the image will display other CDDs and their percent alignments.
- Clicking on “Format” will give the results to homolog in other organisms.
- Lets click on the image.

# Analyzing the BLAST

- The first CDD is pfam Nitroreducatase which has a 100.0% alignment.
- The other CDD also has a good hit (86.0%).
- Both represent similar families but pfam is a recent database compared to COGs.
- Go back and click on “Format” to see the homologs.





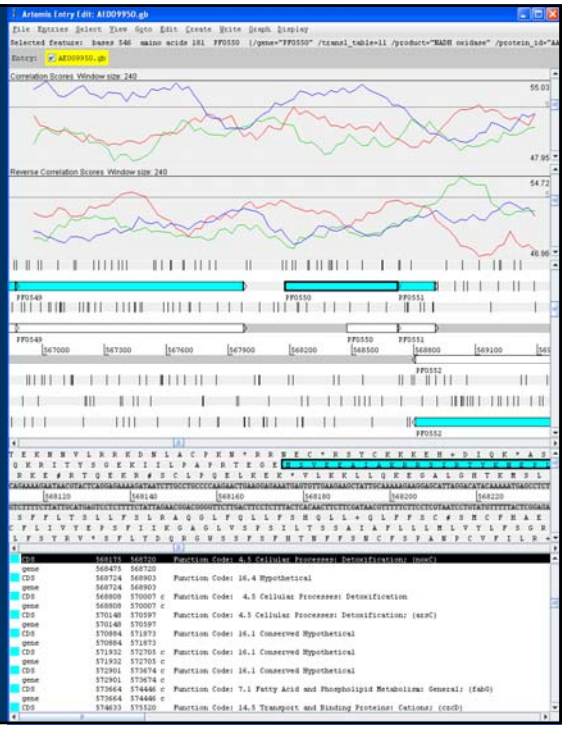




# Analyzing the BLAST

- BLAST shows you where similar genes begin in other organisms.
- Notice that the selected sequence has poor alignment at the beginning.
- The beginning does not align well and should not be included in the gene.
- The gene should start the M downstream where the alignment is accurate.

# Changing the Gene

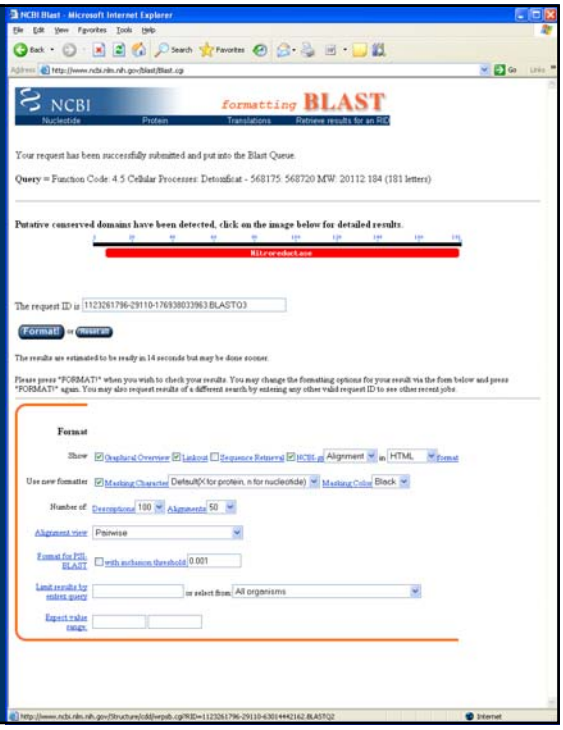
- Move the gene until it starts on M.



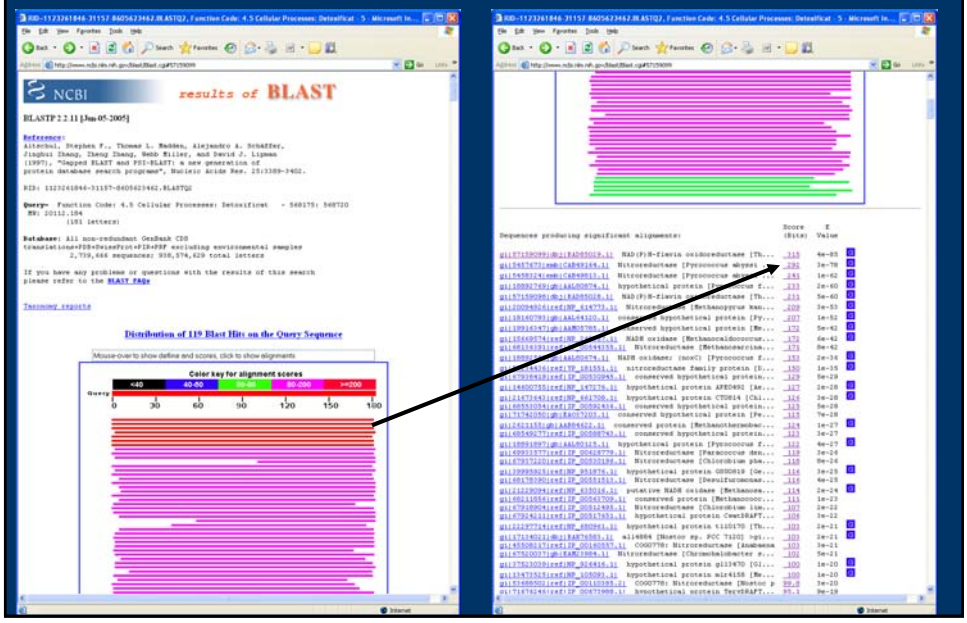
# BLAST the sequence

# CDD Image

- The CDD now begins at the start of the gene.
- Click format to see how the sequence aligns to other species.




# Analyzing the BLAST



# Analyzing the BLAST

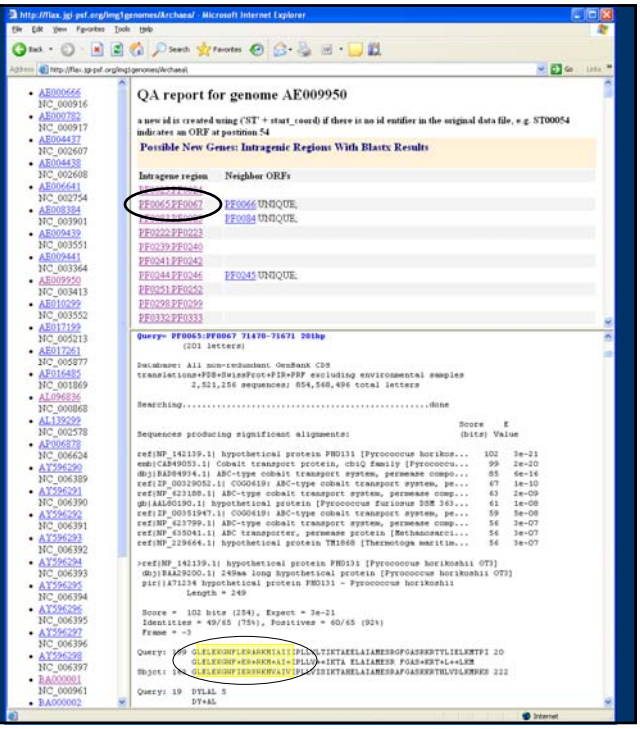
- All the homologs begin on M.
- The gene has been successfully changed so that the CDDs is complete and aligns well with homologs.
- No annotation needed.

# Table of Contents


- Introduction
  - Programs
  - ORFs
  - Annotations
  - Error Report
- 1. Overlaps
  - Different Types
  - Example
- 2. Short
- **3. Frameshift**
- 4. Pseudogenes

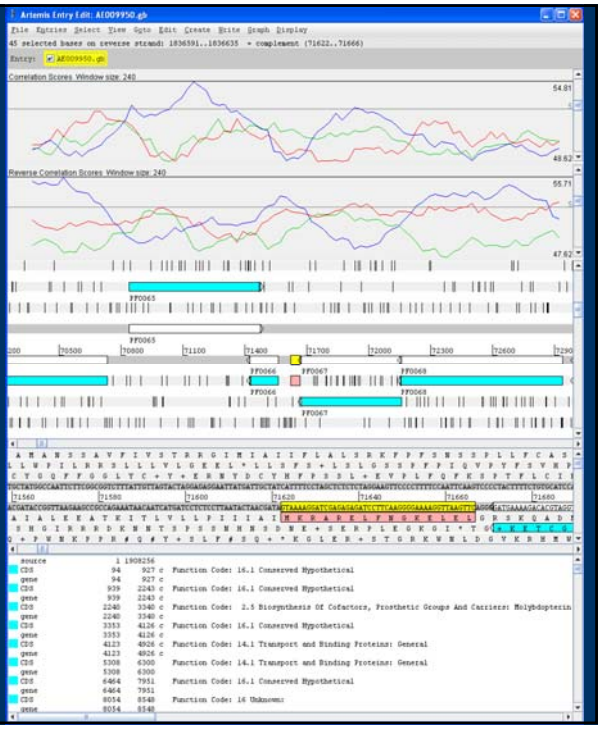
# Error Report

- There is another section of the error report.
- It shows pieces of DNA that were not flagged as a gene.
- PF0065:PF0067 indicates that there is data between the two genes that might be part of a gene.
- The sequence is copied and will be found in Artemis



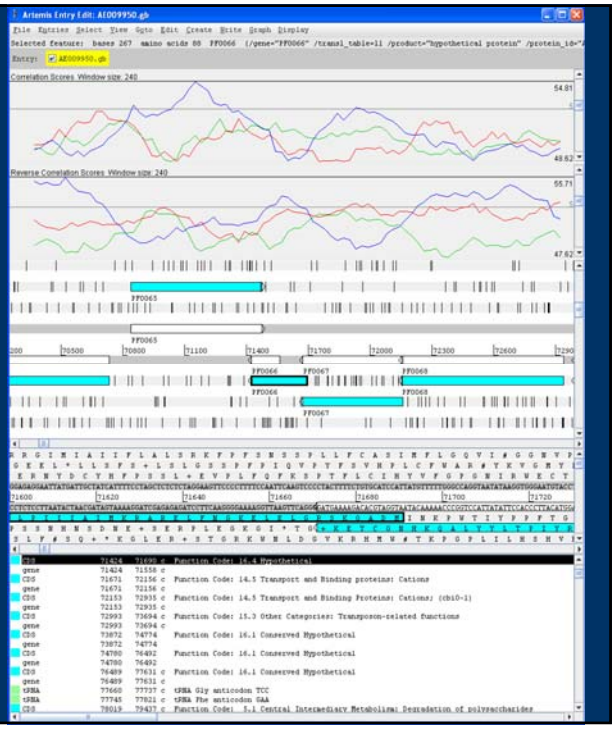

# Find the Missing Region

- The amino acid sequence LELEK… from the previous slide is found in Artemis.
- The blue gene (PF0066) likely includes the pink fragment.



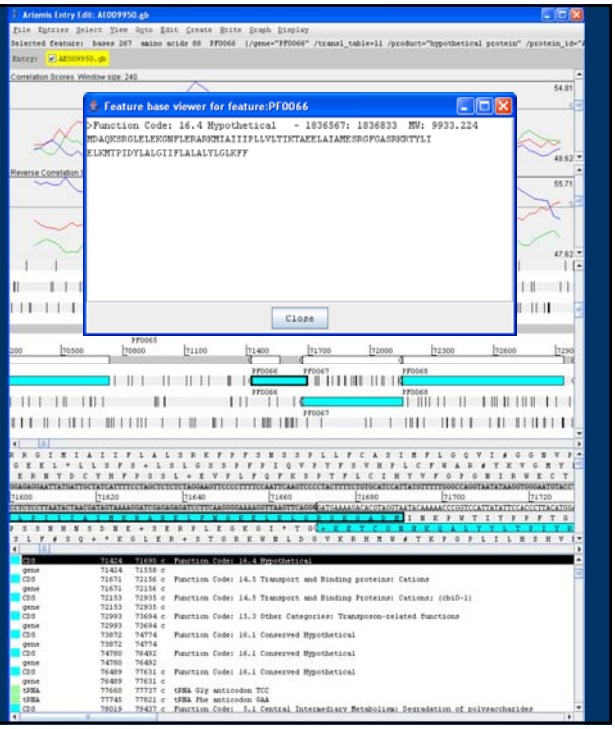
# Change the Gene

- By pressing Control-Q, the gene is push back to the last stop codon.
- Since genes can only begin on start codons Methionine (ATG), Valine (GTG), or Leucine (TTG) we must change the gene so it begin at an appropriate start codon. We cut the gene so it begins at the next available start codon by pressing Control-Y.
- Notice there is a small overlap with the nearby gene. Now lets BLAST the changed gene to see the results.



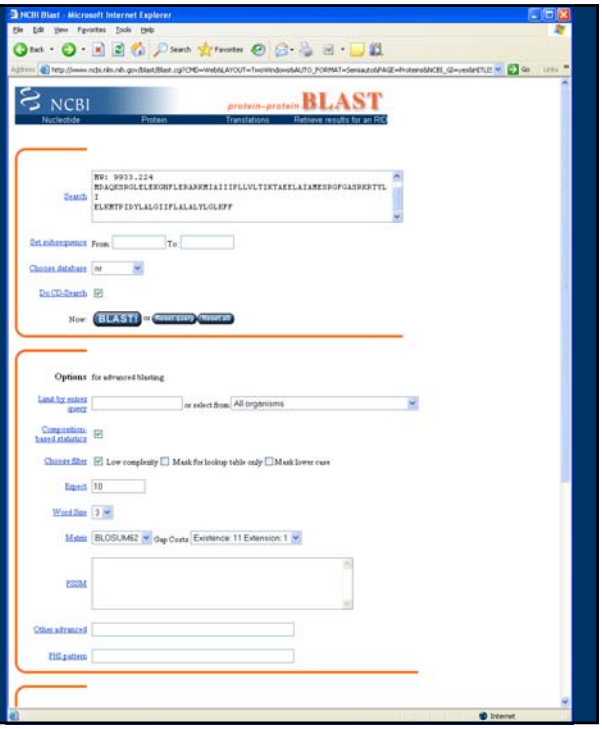
# View the Modified Gene

- The sequence is shown and copied.



# BLAST the sequence

- Enter the sequence in the NCBI BLAST Search box.

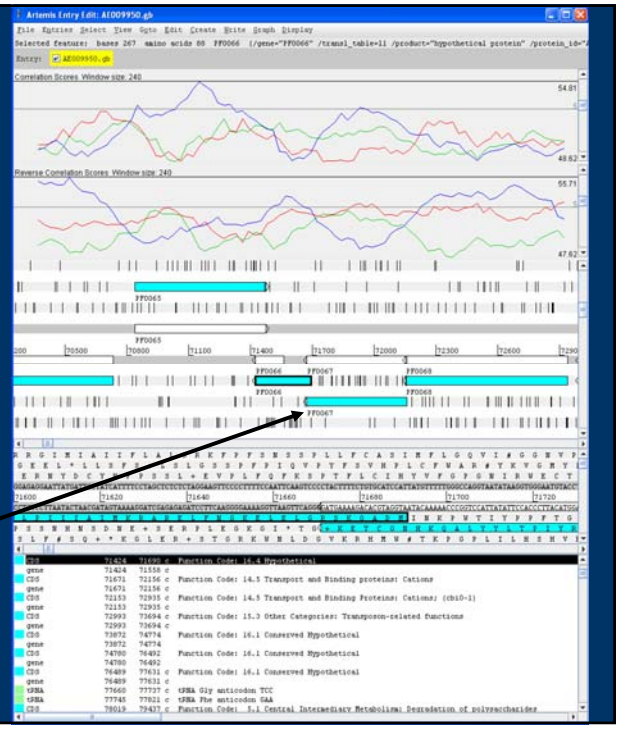



# CDD Image

- The results show that there is partial alignment to the CDD CbiQ.
- This is the ending fragment of the protein. We need to find its beginning.
- The beginning is cut off. We need to go back to Artemis and see where the missing part of the gene went.

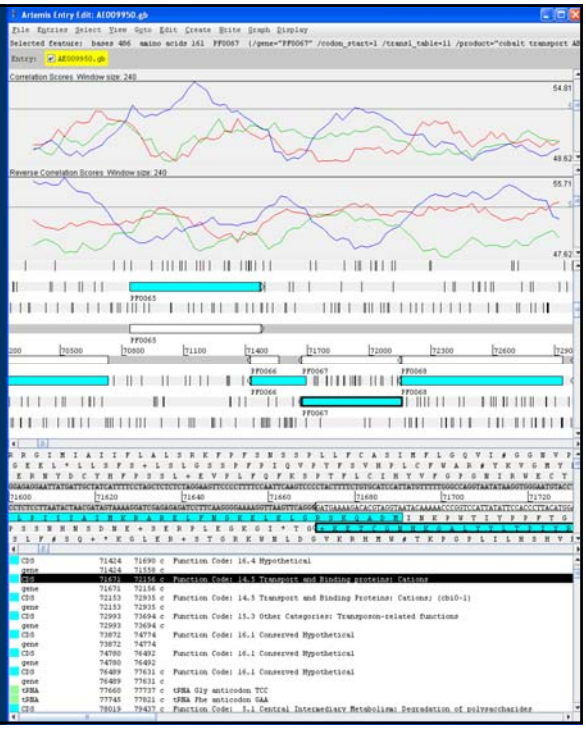
# Analyzing the Genome

- PF0067 looks as though it might contain the missing fragment because it is next to our gene with the cut beginning.
- If we BLAST the sequence of the nearby gene (PH0067) and find the missing CDD fragment, then we found a frameshift.



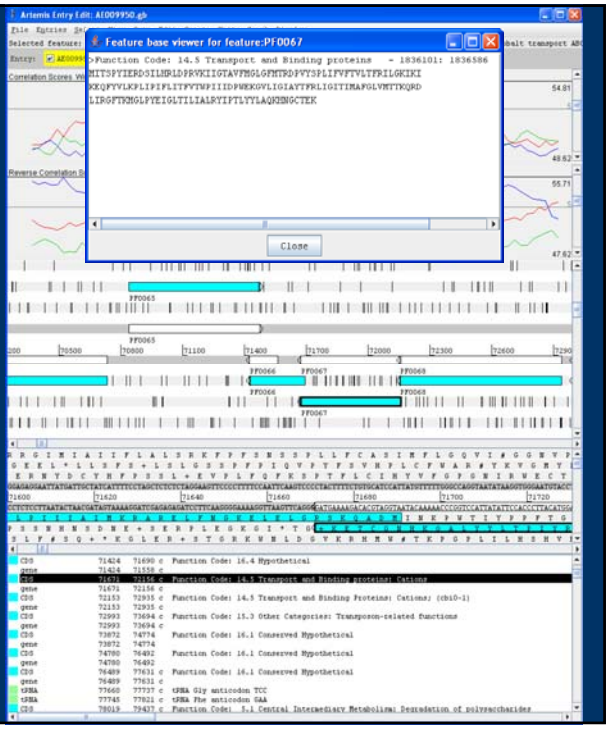
# Selecting Genes

- Click on PF0067 to select it.



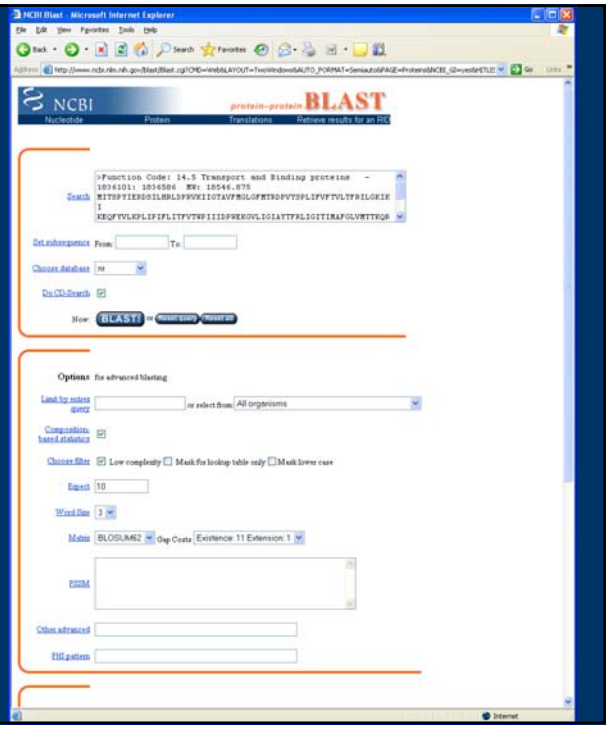
# View the Sequence

- The larger gene's sequence is copied.



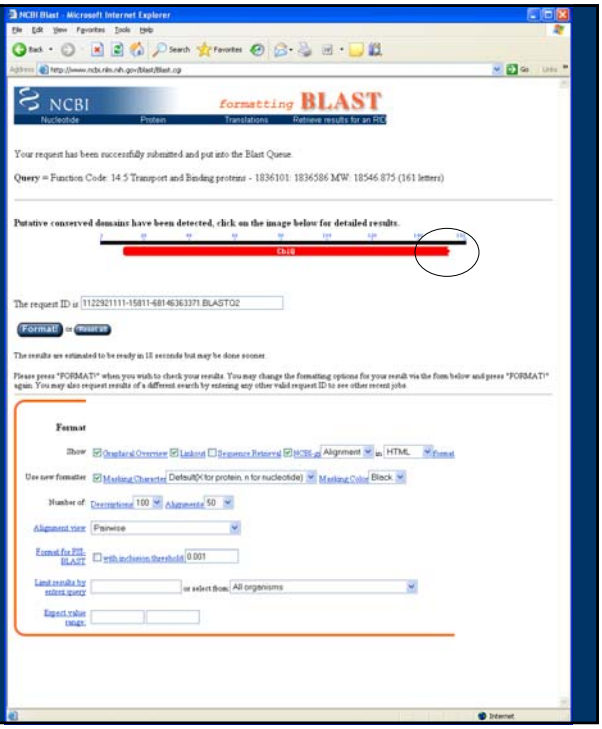

# BLAST the sequence

- Enter it in BLAST



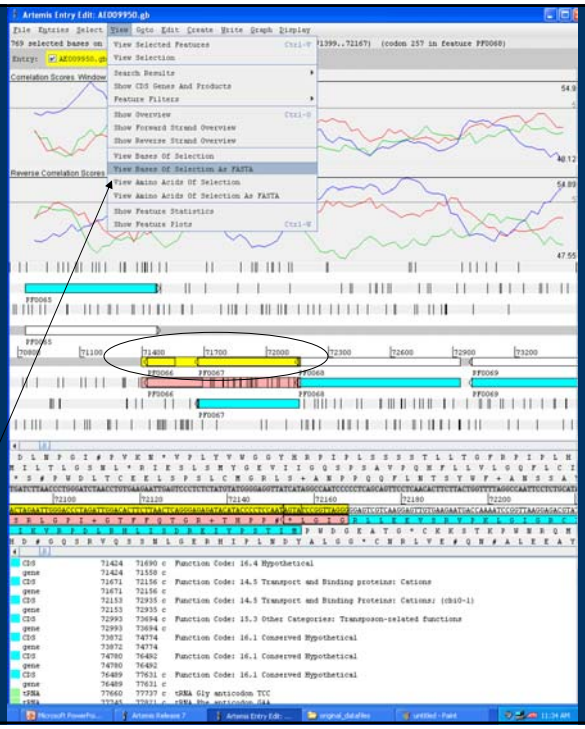
# CDD Image

- From the results we see that the larger gene is the beginning of the gene.



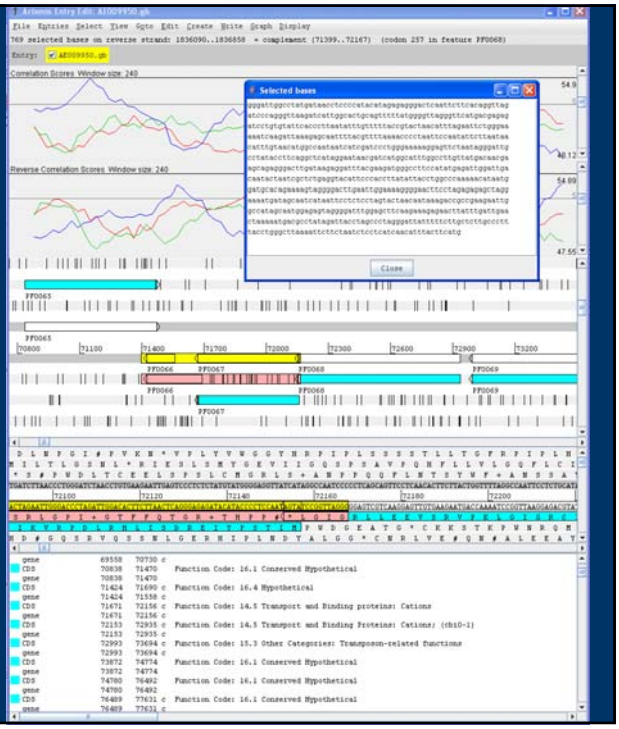

# Selecting the Nucleotide Sequence

- Click and drag the nucleotide sequence (Gray lines) to select the DNA sequence of both genes. It will change to yellow when you select it.
- To view the sequence, go to View Bases of Selection as FASTA



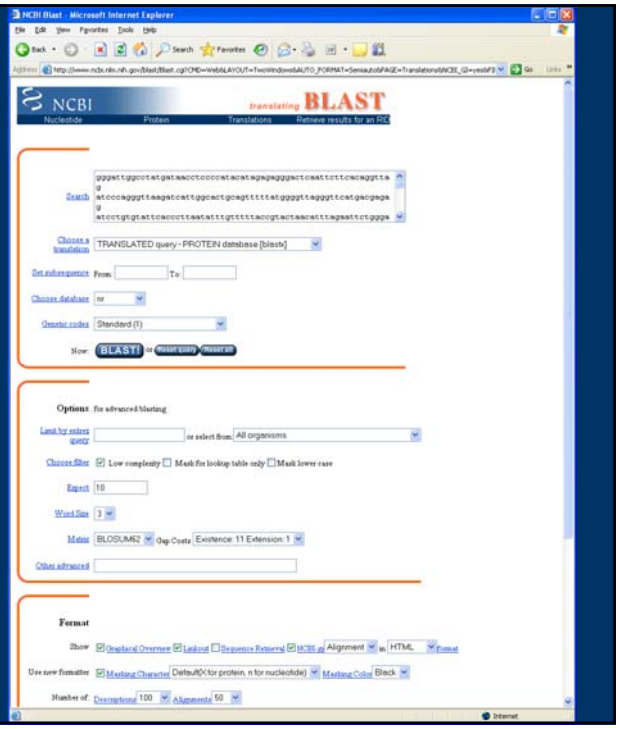
## View the Nucleotide Sequence

- The entire DNA sequence of the highlighted area is given in text format.
- Don't worry about it not being in FASTA form.

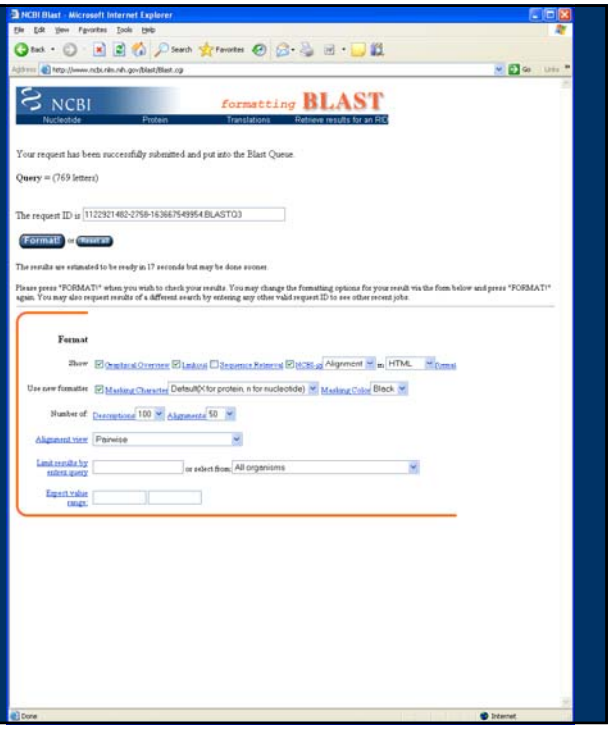


## Translations Page

- Since we want to compare DNA to DNA we click on Translations instead of Protein.
- Before we compare amino acid to amino acid sequences.
- But now that we want to combine two genes on different reading frames, we must use translations.

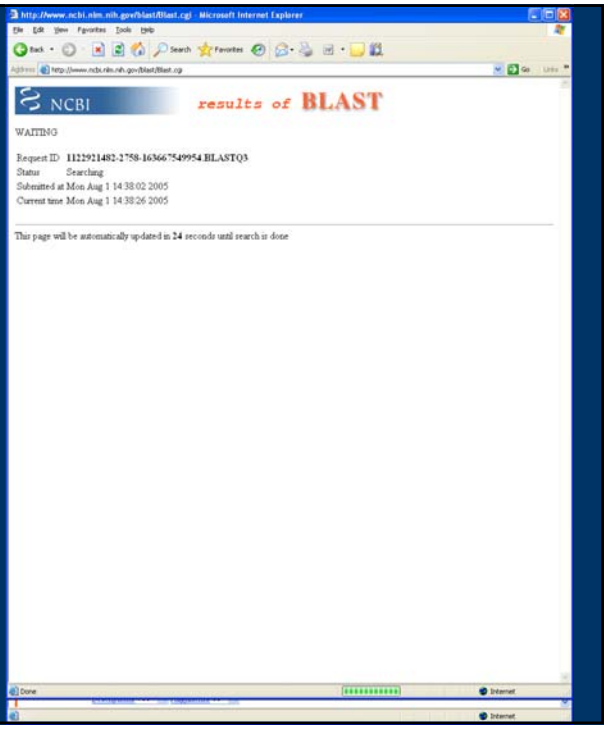


- Click on "Format"

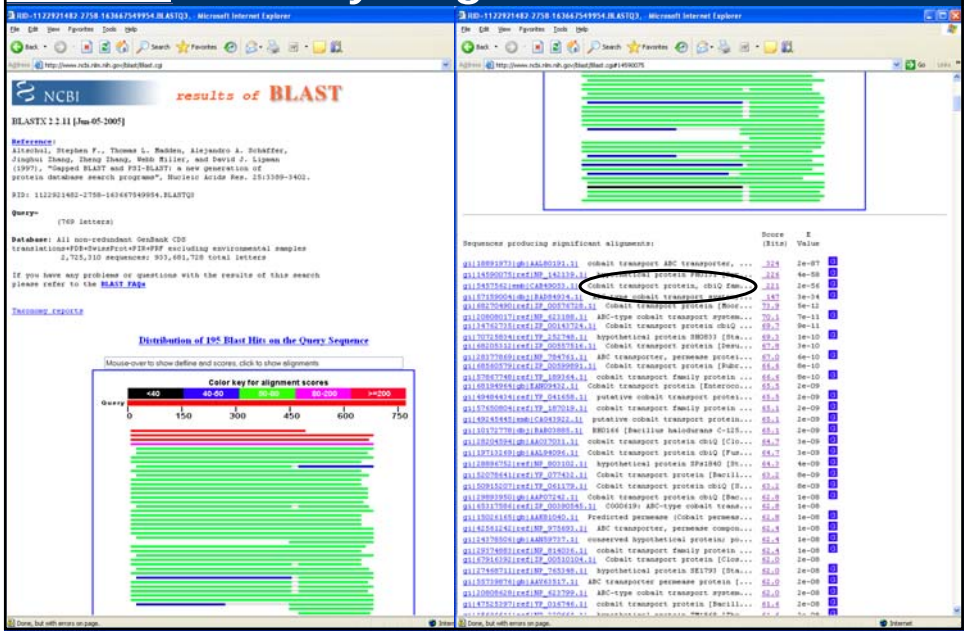



# Wait For Computation

- Patience is a skill highly valued in science.

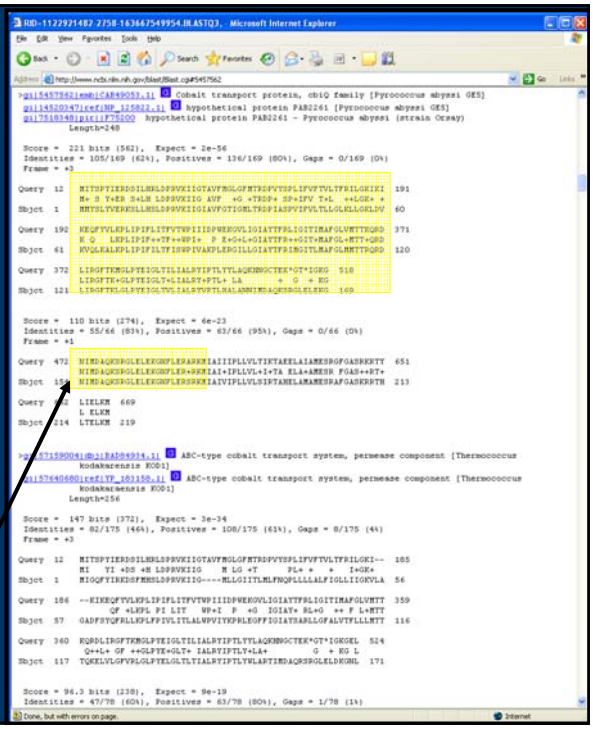


# Analyzing the BLAST

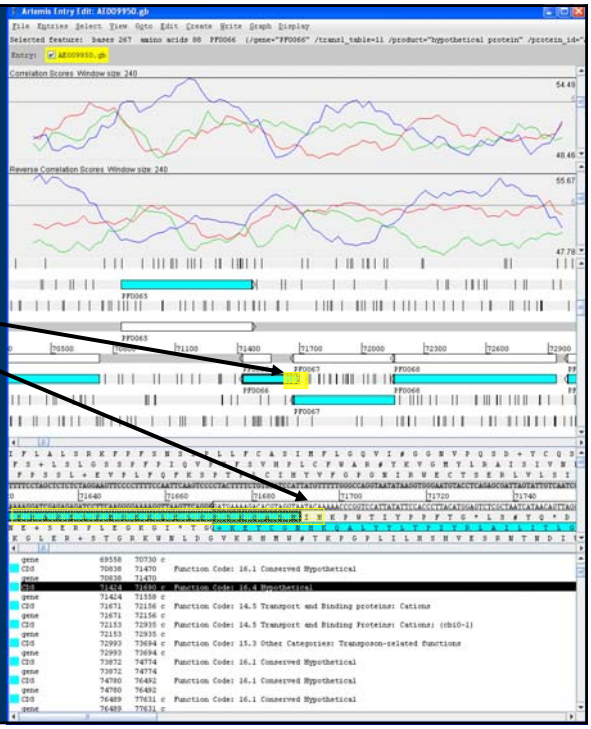



# Analyzing the BLAST

- This is the sequence of the larger gene.
- The first line represent the query.
- The second line shows the similarities to the compared organism.
- The third line is the sequence of a gene already present in NCBI's database.
- Remember the sequence: NIMDA

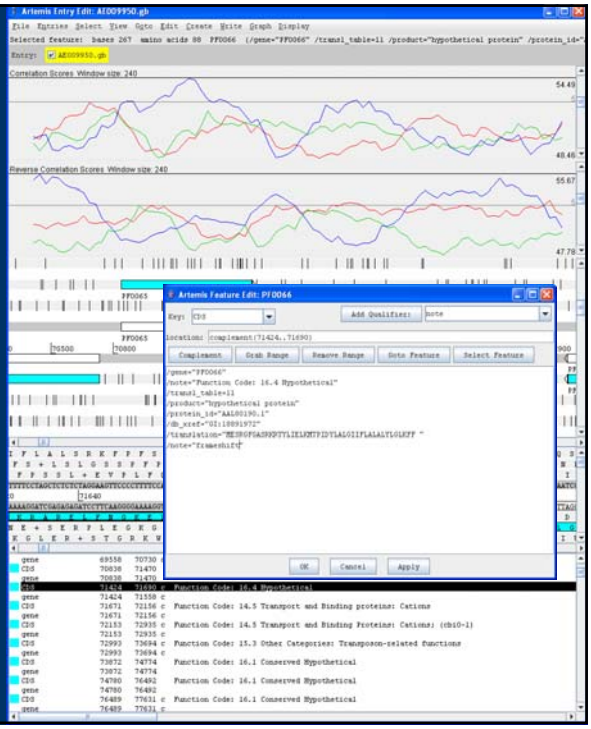


- Now we see the sequence NIMDA… in Artemis.
- The highlighted areas represent the beginning of the smaller gene.
- The BLAST result of the gene sequence of the smaller gene on a Asparagine (N) which is not a valid start codon.
- Thus we must move the gene to the next available start codon, which happens to be a Methionine (M).
- This can be done by pressing Ctrl-Y.



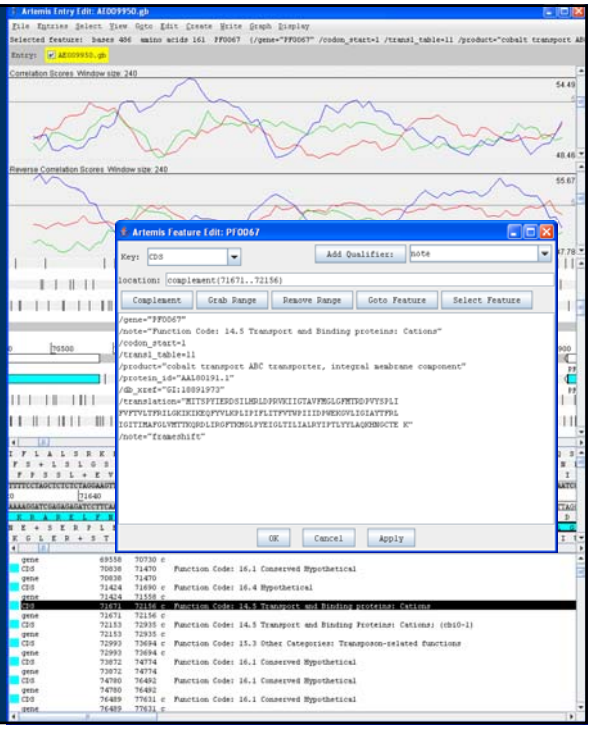
# Editing in Artemis

- Press Ctrl-E to open up the Artemis Feature Editor.
- Make a note by clicking “Add qualifier” and type frameshift between the apostrophes.



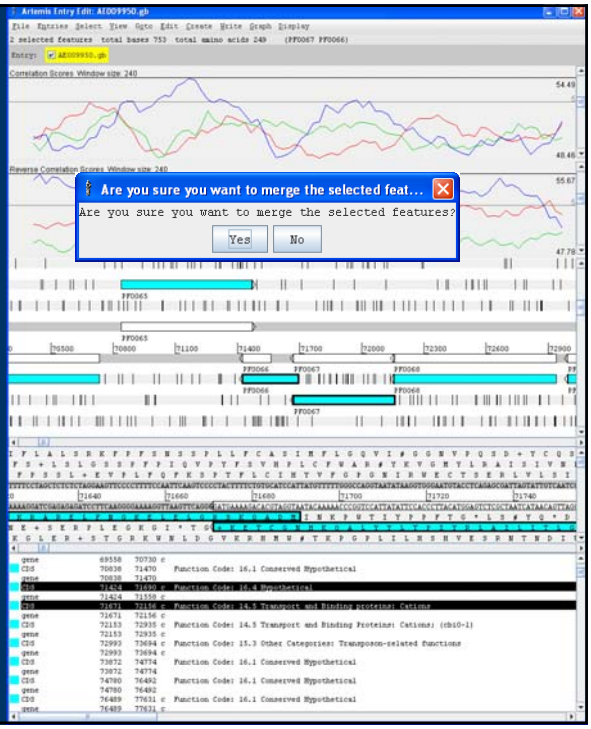
# Editing in Artemis

- Do the same to the larger gene



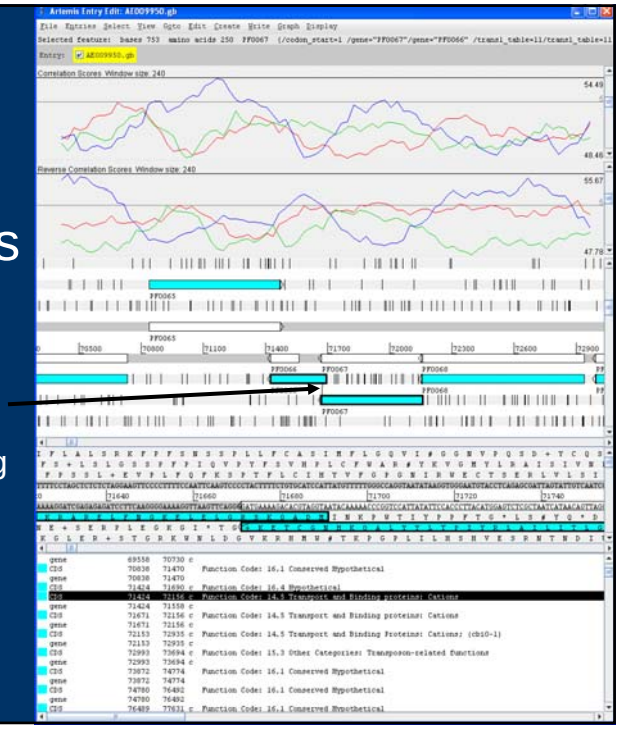

# Merging Genes

- By selecting both genes and pressing Ctrl-M, you can merge the two gene.
- The computer will combine the amino acid sequences of the two gene into one.



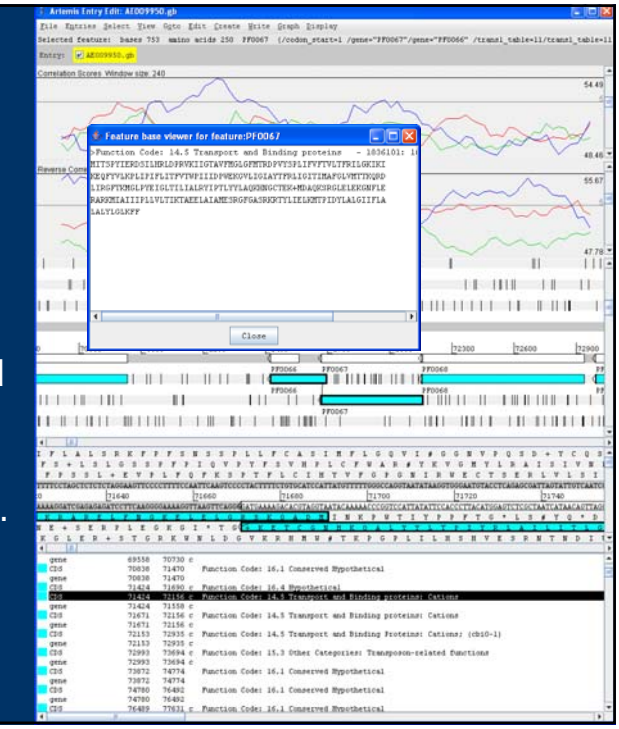
# Seeing Merged Genes

- Notice there is a small line connecting the ends of the two gene representing a merger.
- Artemis recognizes the two as one gene.



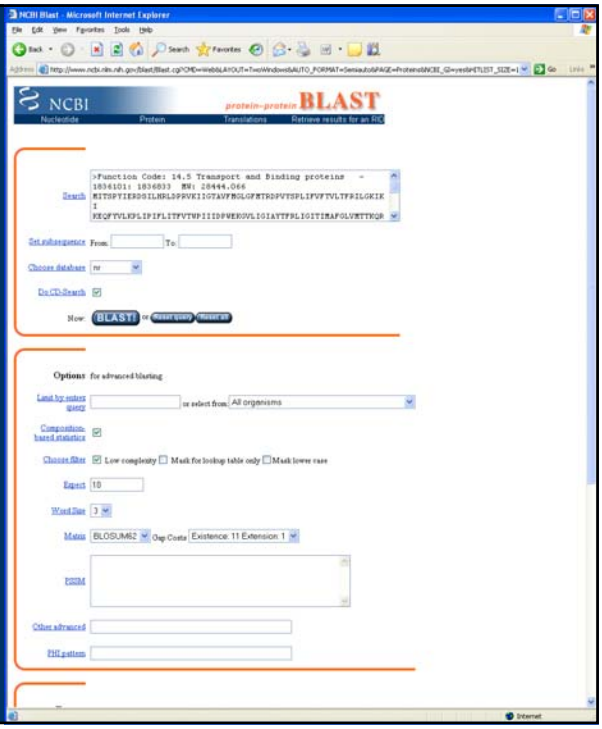
# Viewing the Combined Sequence

- Now lets BLAST the new combined gene to see if we get a complete unbroken CDD hit.
- The entire sequence of the gene is shown.



## BLAST the Merged Genes

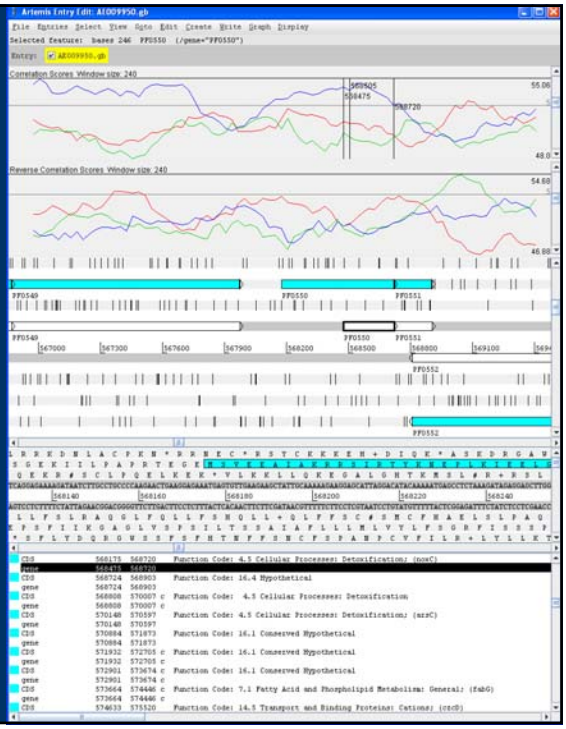
- Enter it into BLAST




## CDD Image of Merged Genes

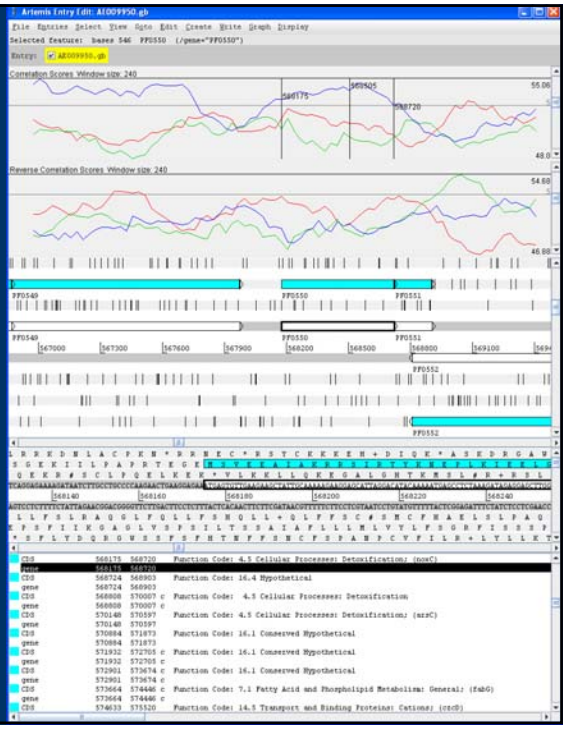
- The old CDD that was separated into two fragments is now one.
- We have successfully located a frameshift error.

- The blue region is called a CDS.
- Whenever a gene is changed, both the blue and white need to be the same.
- Move the white segment until it is on the same M.



# Finished

- The blue and white regions correspond.

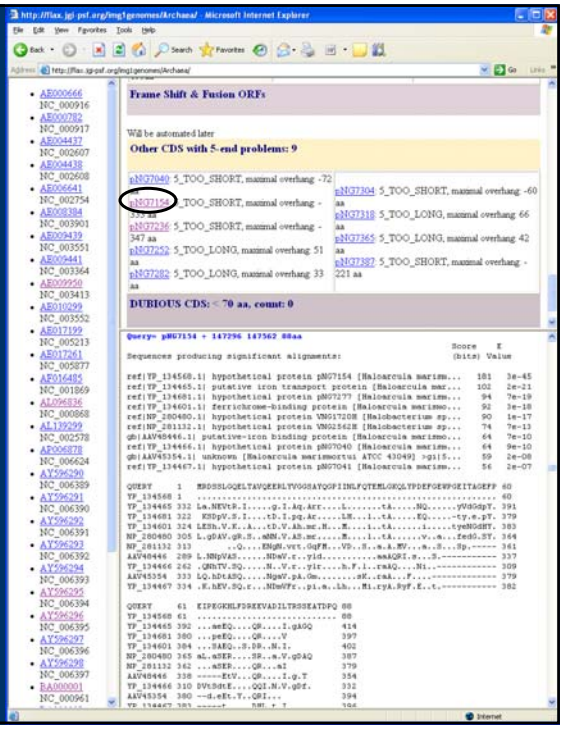


# Table of Contents

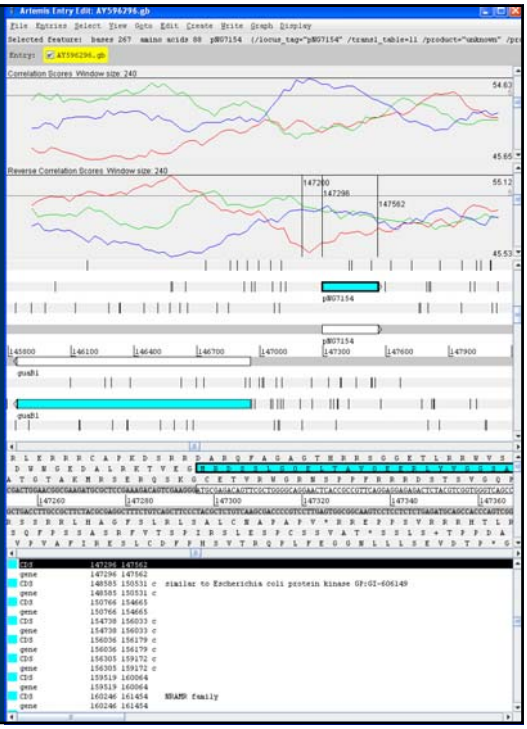

- Introduction
  - Programs
  - ORFs
  - Annotations
  - Error Report
- 1. Overlaps
  - Different Types
  - Example
- 2. Short
- 3. Frameshift
- **4. Pseudogenes**


# Error Report

- This gene was marked as too short.
- Lets see if we can extend it.

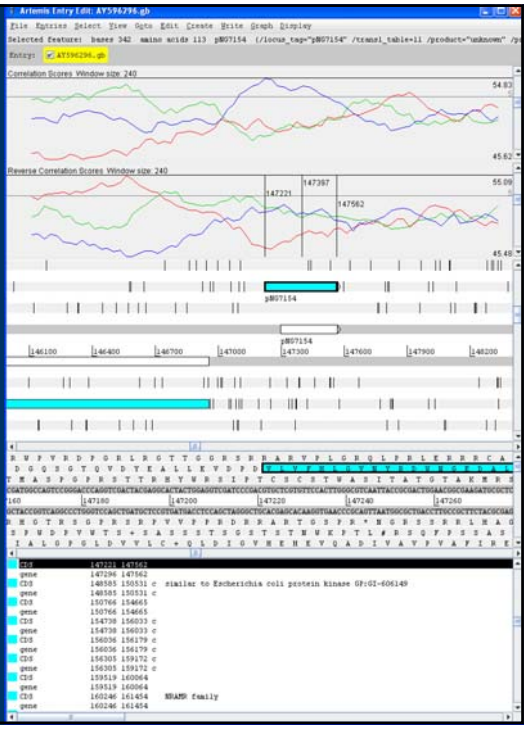


- The gene can be extended at the five prime end because there are no stop codons.
- Also there are no overlaps.

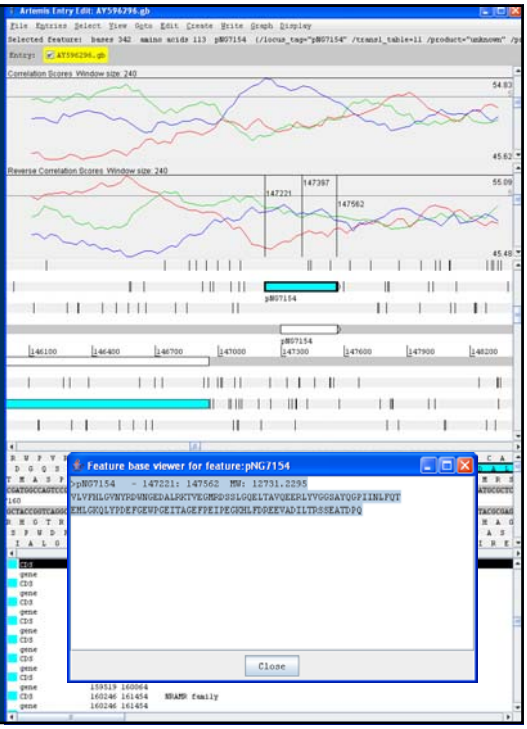


## Changing the Gene

- We extend the gene to first possible start codons.
- The sequence now begins on a Valine (GTG) instead of a Methionine (ATG).



# View the Sequence

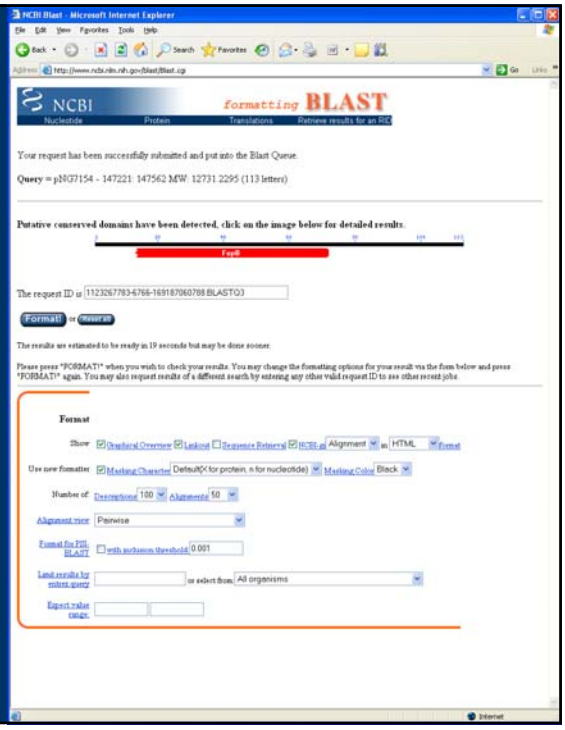



# BLAST the Sequence

- Enter the sequence.

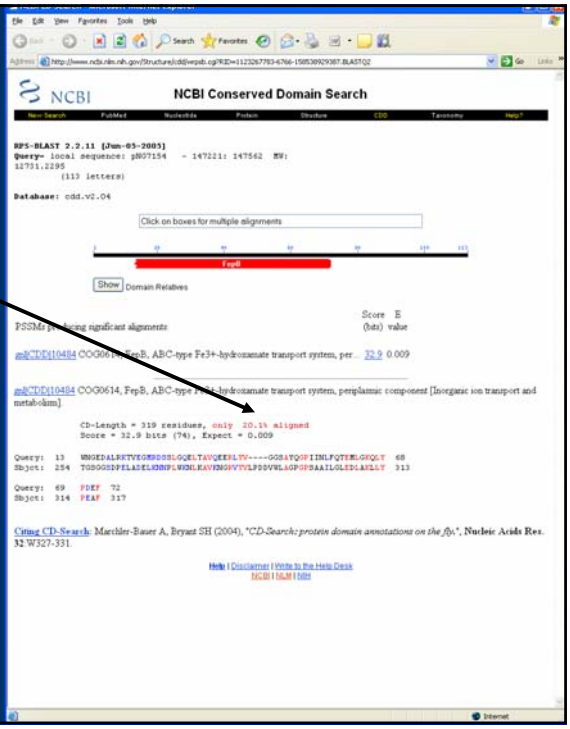
# CDD Image

- The 5’ end is jagged represent that the beginning of the CDD is missing.
- Click on the image to find the percent alignment



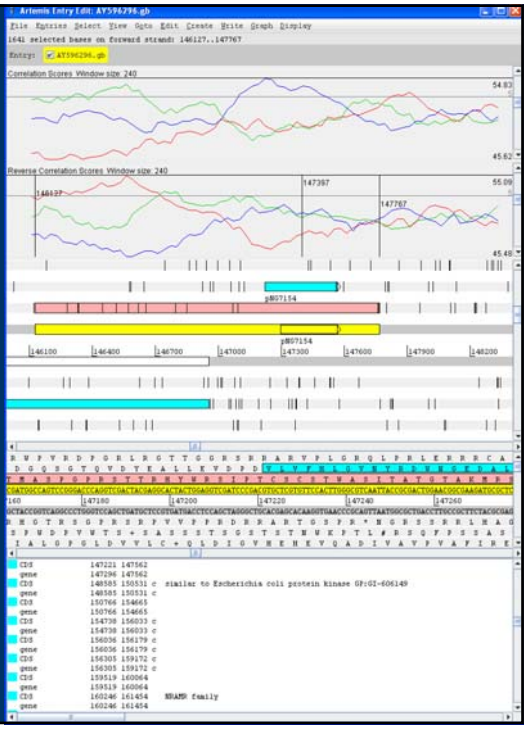

# Percent Alignment with CDDs

- The gene aligns with 20.1% of the CDD.
- Try to find the missing 79.9%.



# Selecting DNA

- Select a large portion of DNA upstream from the gene.
- If the missing part exists, it should come up in the BLAST results.

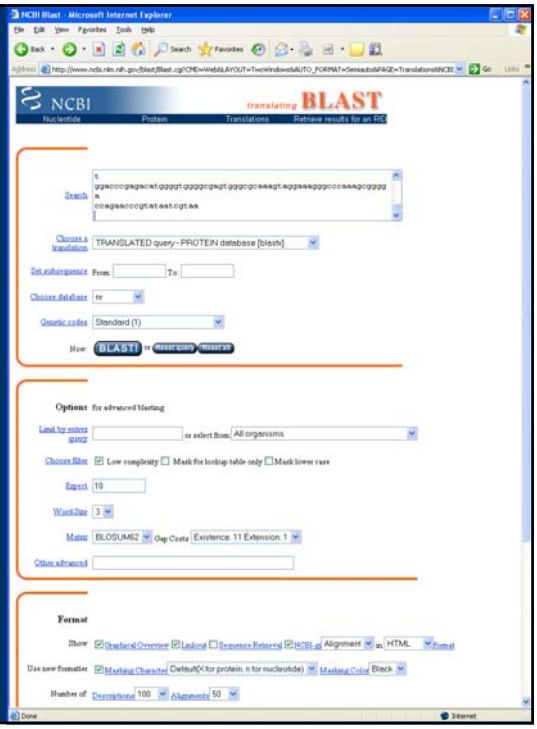


# Artemis Navigator

- View the DNA sequence of the selection.

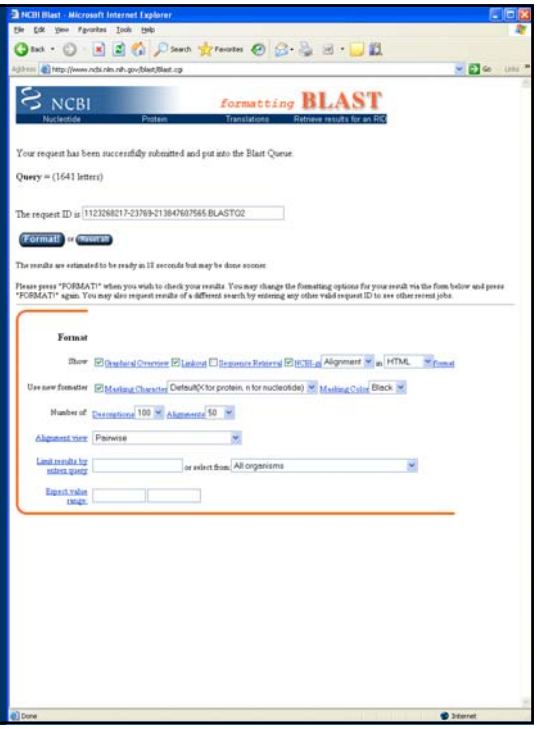
# BLAST the translation

- The DNA sequence will be “BLASTed” across all six reading frames.



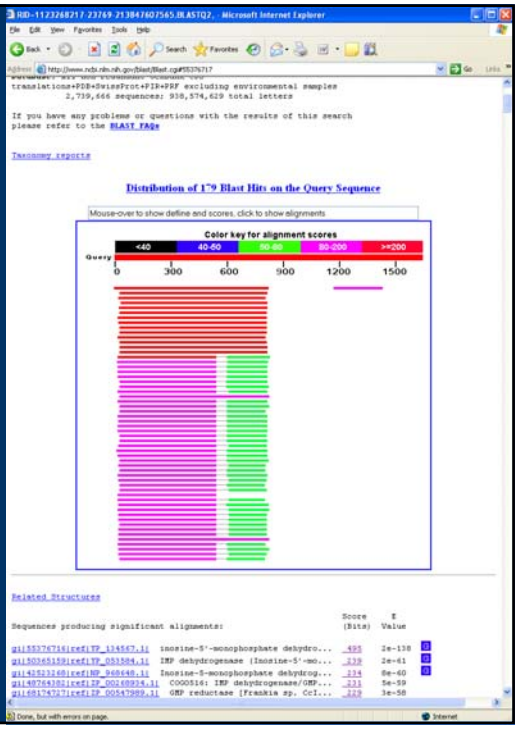

# Format

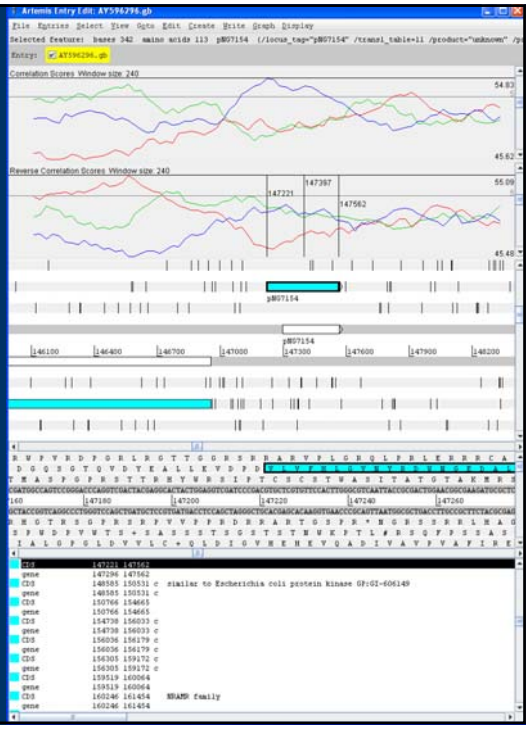
- Click on Format



# Analyzing the BLAST

- The first part of the BLAST has great hits.
- However this corresponds to another gene on the complementary DNA strand.
- The one hit at the right was our query.
- It has no homologs hits (remember that the first hit is always itself)




- Since the missing fragment cannot be found upstream, the CDD remains the same.
- CDD alignment less than 30% are considered pseudogenes.
- This could mean that a portion of the another gene was copied and inserted.
- The gene will be translated but it will not be a functional gene.