May 5, 2016

Made a slide for my annotation for gene 10.1 on Bacillus phage SPO1.

Overall summary:

The past few days, I was trying to do another standard annotation. I took a look at Eyuki, Bobb, Tsar_Bomba, and BCP78. I wasn't very successful though. I found valid genes (with no existing annotations) but couldn't find any papers. But I did do a challenge to an Enterobacteria phage HK97, gene 74.

April 27, 2016

In class:

I emailed the gonuts advisor to ask about my evidence code IGC. He told me to do two standard annotations and that I should look at the upstream gene. I looked at the PubMed and found the upstream gene, 9.1, is a protein for a tail sheath. He directed me to a reference 16, a paper that experimentally finds the function of some of the tail proteins for bacteriophage SP01. The first annotation would use the Uniprot ID for upstream gene, 9.1 because its function is known and use the evidence code IGC. The second annotation would use the evidence code IDA and refer to Figure 3 in the paper.

April 25, 2016

In class:

I wanted to do one more standard annotation and so I looked at the genbank files for Hakuna, Bobb, and Troll. Most of the genes in Hakuna and Bobb already had existing annotations but I found promising genes for Troll. I took a look at gene 74, a portal protein in Troll.

Here's a summary of the total annotations and challenges I have done:

Annotations:

Standard annotation for RNA Polymerase sigma factor in Bacillus phage SP01
Transfer annotation for RNA Polymerase sigma factor in Gene 82 Eyuki
Standard annotation for virus tail tube in Bacillus phage SP01
Transfer annotation for virus tail tube in Bacillus phage TsarBomba, gene 87
Transfer annotation for virus tail tube in Bacillus phage Troll, gene 85
Transfer annotation for virus tail tube in Bacillus phage Bobb, gene 186
Transfer annotation for virus tail tube in Bacillus phage BCP78, gene 0213

Challenges:

Challenge to a transfer annotation of Legionella pneumophila subsp. pneumophila and Hakuna_54
Challenge to annotation on Escherichia coli O145:H28 str. RM12581
Challenge to a standard annotation for Bacillus phage TsarBomba
Challenge to a standard annotation for Bacillus phage BCP78
Challenge to an annotation for Bacillus phage TsarBomba
Challenge to a transfer annotation for Legionella pneumophila subsp. pneumophila and Hakuna_54

April 24, 2016

A list of newly added phages:

Wolbachia and dengue
Phages:
Phrodo
Nigalana
Zuko
NotTheCreek
Vinny
Nemo
Juglone
DIGNKC
SageFayge
April 18, 2016

In class:

I wanted to work on an incorrect annotation, an annotation on TsarBomba_41. I searched up the file on GenBank and blasted it and found that it matched to Bacillus phage SP01 gene 2.11. UniProt linked me to a paper. This paper detailed SP01's entire genome. Seeing the paper, Allison recommended that I instead look at some of the heavily conserved genes like the portal protein, protease protein, and scaffolding protein. I read the paper and found that in the tail assembly genes section, it identified gene 10.1 as a tail tube subunits. I searched gonuts and found that there were no existing annotation for gene 10.1 and so I can use this paper as a standard annotation for 10.1!

Standard Annotations:

The information for 10.1 for the standard annotation (Annotations haven't opened up yet):

The paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666789/#!po=24.1667

- Specifically look at the tail assembly genes section which explains why 10.1 is a good candidate for a tail tube subunit protein
  - Specifically, 10.1 is the expected size for a tail tube subunit protein and other homologs, have corresponding positions.
  - Take a look at Figure 5
  - I think the evidence code might be IGC
  - No existing annotations exist in QuickGo or Gonuts

Transfer Annotations:

- I could also do a transfer annotation. Since I had first started with TsarBomba and I looked at the genbank file of TsarBomba and searched for tail tube. I blasted both of sequences together (TsarBomba_87 and SP01 10.1) to find 94% query cover, E-value of 7e-43, and identity of 46%.
  - Uniprot ID for SP01 10.1: B6V2P5
  - Uniprot ID for TsarBomba_87: A0A0K2D034
- I could also do another transfer annotation with the homologs of TsarBomba, for example: Bacillus phage Troll, specifically gene 85. I blasted Troll_85 and SP01 10.1 and got these results: Query cover is 97%, E-value of 6e-41, and identity of 44%.
  - Uniprot ID for Troll_85: S5YD00
- Bacillus phage Bobb is another homolog to TsarBomba and I could use it for a transfer annotation, specifically, gp_186. On Genbank, it is characterized as a structural protein. I blasted against SP01 10.1. The results: query cover is 97%, E-value of 2e-41, and identity of 44%.
  - Uniprot ID for Bobb gp_186: A0A076G711
- One more homolog to TsarBomba is Bacillus phage BCP78, specifically gene BCP78_0213. It is characterized as a putative uncharacterized protein on UniProt and a hypothetical protein on GenBank. Blasted against SP01 10.1 gave these results: query cover is 97%, E-value of 7e-43, and identity of 46%.
  - Uniprot ID for BCP78_0213: G9J213

To do:

- The only other thing I need to do is look at HHpred results for all of my transfer annotations.

April 13, 2016

In class:

Our group won for our class! We had a total of 45 points. Today, we focused on doing challenges. Now that I've done a few annotation, I can better understand and find the errors of other annotations. This time, I found that those who used evidence codes didn't provide the necessary information. For example, one annotation used evidence code ISS, but did not provide Blastp results or only provided the E-value for the blastp. One annotation used an ISS code but did not provide any Blastp results. They provided an article but it linked to a completely different organism: Branhameill catarrhalis but they were trying to find a function/find a homolog for Bacillus phage TsarBomba. The Uniprot ID linked to TsarBomba_41 and Pseudomonas phage PaP3, but the notes section did not mention either proteins. I should perhaps do an annotation for this one.

April 6, 2016

In class:

I finally found the article to do a standard annotation! On Blastp, gene 182 matched gene 34 of Bacillus phage SPO1 with a 43% query coverage, E-value of 1E-09, and identity of 45%. (This is in itself is a transfer annotation). HHpred results were the similar for Bacillus phage SPO1 and Gene 182 in Eyuki. Phage SPO1 results were: probability of 99.9% and E-value of 1.4E-26. Gene 182 in Eyuki results were: probability of 99.9% and E-value of 6.3E-24. Both were matched to protein 1OR7, a RNA Polymerase Sigma-E factor.

I looked up gene 34 of Bacillus phage SPO1 on Uniprot and found that it was reviewed! This paper showed that the function of gene 33 and 34 of Bacillus phage were RNA Polymerase Sigma Binding Factors. Specifically, Figure 2 shows how the function was determined: A vector plasmid was used to help identify function. A solution of isolated RNA polymerases were inserted into lane 5 in the gel causing an aggregation of proteins at specific locations. I now have two annotations.

April 4, 2016

In class:

We challenged Jheins 2 Team Green B because their PubMed article did not mention their specified gene at all. It talked about its function, thymidylate synthase activity, but there is no gene mentioned in the paper. So, we suggested that the group find a PubMed article specifically about their gene M1Ia and connect to its function in the notes (which is not thymidylate synthase activity).

March 28, 2016
In class:

Today, we tried to find a tail assembly chaperone in the Megatron genbank file but it didn't show up. Instead, we searched up the GenBank file for Eyuki and located the tail assembly chaperone and there were two: gp104 and gp103. In order to locate the tail assembly chaperone on Megatron, we blasted the amino acid sequences for gp104 and gp103 individually to find that it linked to gp98 and gp97 respectively. In Megatron, both of these proteins were labeled as "hypothetical proteins." Then, we blasted gp98 and gp97 against phage lambda sequences and for both cases, there were no matches between the two sequences. In order to find these proteins on lambda, we followed the go-annotation guide. We looked at relevant papers to figure out what gene it is on lambda and we found it to be Gene G. We found the geneID on the genontus page.

<table>
<thead>
<tr>
<th>Dirty GenBank annotations</th>
<th>BlastP hits</th>
<th>Blastp conser ved domain</th>
<th>Best blastp hit</th>
<th>HHPred hits</th>
<th>Other notes</th>
<th>it as ser re su li</th>
</tr>
</thead>
<tbody>
<tr>
<td>113 4 0 1 (5 4)</td>
<td>baseplate protein, putative baseplate protein, baseplate lysozyme</td>
<td>basepl ate wedge subunit</td>
<td>baseplate protein Bacillus phage Eyuki</td>
<td>First hit is protein 2IA7, a crystal structure of putative tail lysozyme from Geobacter sulfurreducens. Probability is 99.9% and E-value of 3.2E-22. Second hit is 4HRZ, phage T4 Sheath Initiation Protein gp25. Probability is 99.8% and E-value is 5.3E-26.</td>
<td>The best matches are for Bacillus phages. There are other phages in blastp matches but their identity is very low. I would assume that this sequence of DNA is conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>114 2 4 1 (3 9)</td>
<td>Baseplate j protein, gp38, baseplate assembly protein J</td>
<td>Baseplate J-like protein</td>
<td>Baseplate j protein Bacillus phage Megatron</td>
<td>First hit is protein 5H0X, an in vitro assembled star-shaped hubless T4 baseplate with a probability of 100% and E-value of 5E-42. Second hit is 3H2T, a crystal structure of gene product 6, baseplate protein of bacterophage T4. It has a probability of 99.3% and an E-value of 3.5E-11.</td>
<td>Shows up for many Bacillus phages but in blastp, as you scroll down the list, there are also other phages such as Listeria phages and Staphylococcus or Enterococcus phages with high query covers (around 96%). The sequence might not be heavily conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>118 4 9 1 (5 3)</td>
<td>DNA Helicase I, gp42, recombinase helicase</td>
<td>Helicas e superfamily</td>
<td>DNA Helicase I Bacillus phage Megatron</td>
<td>First hit is 2OCA, the crystal structure of T4 UvsW, classified as a helicase. Probability is 100% with an E-value of 5.6E-40. Second hit is 5FMF, the P-lobe of RNA polymerase II pre-initiation complex. Probability is 100% and E-value is 1.8E-19.</td>
<td>This sequence of DNA shows up in Bacillus phages but also in Staphylococcus, Listeria, and Enterococcus phages. The sequence isn't very conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>119 3 0 8 8 4 3 (8)</td>
<td>HTH binding domain protein, DNA binding protein, transcriptional regulator</td>
<td>None</td>
<td>HTH binding domain protein Bacillus phage Eyuki</td>
<td>First hit is 2V79, crystal structure of the N-terminal domain of DNA primase from Bacillus subtilis. Probability is 99.2% and E-value is 1.5E-11. Second hit is 4RBR, a crystal structure of repressor of toxin, a central regulator staphylococcus aureus virulence. Probability is 99.2 and E-value is 1.7E-10.</td>
<td>Blastp hits are only Bacillus phages. The DNA sequence is conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>120 4 9 1 8 4 1 8 3 (1)</td>
<td>HeliX-turn-helix binding domain protein, putative transcriptional regulator, gp44</td>
<td>Helix-turn-helix</td>
<td>Helix-turn-helix binding domain protein Bacillus phage Megatron</td>
<td>First hit is 3C76, a crystal structure of transcription regulator from Pseudomonas syringae pv. Tomato str. DC3000. Probability is 99.8% and E-value of 1.5E-19. Second hit is 4P96, a FadR, Fatty Acid Responsive Transcription Factor from Vibrio cholera with a probability of 99.8% and E-value of 1.8E-19.</td>
<td>Best blastp matches are for Bacillus phages. The DNA sequence is conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>122 8 5 0 1 5 1 (1)</td>
<td>DNA helicase II, RecA-like domain, gp46</td>
<td>DNA helicas e C termin al domain</td>
<td>DNA helicase II Bacillus phage Megatron</td>
<td>First hit is 4MNH, an aquifex aeolicus replicative helicase (DNA) complexed with ADP. Probability is 100% and E-value is 2.6E-54. Second hit is 2R6A, a crystal form BH1. Probability is 100% and E-value is 8.1E-54.</td>
<td>Good matches are only to a small number of Bacillus phages. This might mean that the DNA sequences is conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>123 1 5 7 1 1 8 5 (8)</td>
<td>Hypothetical protein, gp47</td>
<td>None</td>
<td>Hypothetical protein Bacillus phage Hakura</td>
<td>First hit is CpMnP1 with Mannobiocide Bound, classification is a sugar binding protein. Probability is 97.6% and E-value is 3.1E-05. Second hit is 3O6I, a crystal structure of extracellular solute-binding protein from Blisobacterium longum subsp. Infantis. Probability is 97.5% and E-value is 5E-06.</td>
<td>-</td>
<td>1.1 3</td>
</tr>
<tr>
<td>124 2 4 8 3 (5 1)</td>
<td>Exonuclease II, gp48, phosphoesteras e, nuclease</td>
<td>Exonucle ase</td>
<td>Exonuclease II</td>
<td>First hit is 3THO, a crystal structure of MerR11-Rad50 in its ATP/ADP bound state. Classification is hydrolase/DNA binding protein. Probability is 100% and E-value is 2.9E-41. Second hit is 4LYT, a crystal structure of E.coli SsbD at 1.8 A resolution, classification is hydrolase. Probability is 100% and E-value is 3.4E-37.</td>
<td>Most matches are to bacillus phages though there are some staphylococcus phages. This sequence might be exclusive to bacillus phages.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>125 7 8 3 6 4 2 (4)</td>
<td>Exonuclease I, gp51, recombinase nuclease</td>
<td>DNA repair exonuclease</td>
<td>Exonuclease I</td>
<td>First hit is 3QF7, the MRE11-Rad50 complex forms an ATP dependent molecular clamp in DNA double strand break repair. Probability is 100% and E-value is 1.6E-37. Second hit is 3AU4, a crystal structure of Rad50 bound to ADP. Probability is 9.9E-37.</td>
<td>There are many Bacillus phages, some Listeria phages, and a few staphylococcus phages. Though the best matches are for Bacillus phages which means that the DNA sequence is conserved.</td>
<td>1.1 3</td>
</tr>
<tr>
<td>128 7 9 3 6 (3 6)</td>
<td>Hypothetical protein, gp52</td>
<td>None</td>
<td>Hypothetical protein Eyuki_128</td>
<td>First hit is 4G6D, a G1 ORF67, a Staphylococcus aureus sigmaA domain 4 complex, classification DNA binding protein. Probability is 100% and E-value and probability for second hit is too low.</td>
<td>-</td>
<td>1.1 3</td>
</tr>
<tr>
<td>130 3 1 7 7 5 1 (1 5)</td>
<td>DNA primase, gp54</td>
<td>DNA primase</td>
<td>DNA primase Bacillus phage Megatron</td>
<td>First hit is 2UA9, crystal structure of the aquifex aeolicus primase (Zinc Binding and RNA Polymerase Domains), or DNA primase. Probability is 100% and E-value is 1.7E-54. Second hit is 1DD9, the structure of the DNAG catalytic core. Probability is 100% and E-value is 1.8E-37.</td>
<td>Though there are some Staphylococcus phages and Listeria phages. The best matches are with the Bacillus phages. This shows that the DNA sequence is conserved within Bacillus phages.</td>
<td>1.1 3</td>
</tr>
</tbody>
</table>
Kida was exactly the same as DirtyBetty

March 21, 2016

I found that protein 128 was interesting.

March 2, 2016

In class:

Confirmed annotations with Emily and Kelly. I started working on Kida and finished it. Kida went by quickly because it was identical to DirtyBetty.

Puthuveetil_112_136.dnam5
Kida_autoannotation_322016.dnam5

February 29, 2016

In class:

Finished all the genes. I checked some peculiar genes with HHpred and got some results.

DirtyBetty_autoannotation_22916.dnam5

February 24, 2016

In class:

Finished coding potential for genes 112-132. Only have four more to go! Some of the genes have functions such as DNA primase, dUTPase, and exonuclease I. I also used HHpred to check to see if the hypothetical proteins shown in blast were actually proteins on HHpred.

DirtyBetty_autoannotation_22416.dnam5

Interesting stuff:

- Gene 115 comes up with a conserved domain but the list of potential matches shows that it is a hypothetical protein. Also, the picture below is really strange. It seems that a big portion of the genome doesn’t match but the ends do.
- The start position for Gene 113 doesn't cover the coding potential.
- The start position for Gene 121 and 131 covers the black line but not the red line.

This is an interesting gene match for gene 115 on Blast.
February 23, 2016

At home:

On HHPred:
Annotated genes 113-124 today. I was extremely lucky, 8 of the genes so far have known functions. I have found an exonuclease protein, two baseplate proteins, two DNA helicase proteins, and one helix-turning-helix binding domain protein, one pyruvate formate-lyase protein, and maybe one tail protein. I have yet to finish the coding potential for the genes (It's a bit confusing for me to understand). I have not started on Kida yet.

DirtyBetty_autoannotation_22316.dnam5

February 12, 2016

At home:

I annotated Gene 3. It was a little bit difficult because it was on the reverse strand, thus, the start and stop was backward. The gene matched with Eyuki_4. It seems to me that DirtyBetty is very similar to Eyuki. I have yet to look at the coding potential for both Gene 2 and Gene 3. But, I have identified the SD score and blast match for Gene 3.

DirtyBetty_autoannotation_21216.dnam5

February 11, 2016

At home:

I worked on annotating Gene 2. For this gene, it was easy to identify which was the start codon because SD score, blast match, and ribosome binding site was favorable. I have yet to check the coding potential. Again, this gene matched with Eyuki, but this time Eyuki_3. I wonder what the difference between Eyuki_2 and Eyuki_3 is.

DirtyBetty_autoannotation_21116.dnam5

February 10, 2016

In class:

Today I annotated the first gene of DirtyBetty. We tried to find the where the start codons were using the coding potential, blast match, SD score, and its length. We decided that the start codon was at 1521 because of its 1:1 blast match, coding potential, and length.

DirtyBetty_autoannotation_21016.dnam5