Annotating the Genome of a Bacteriophage - Part 1

The process of annotating a genome is a file-manipulation-rich endeavor. Understanding where your files are and what you are doing to them will be very helpful. The genomes of most phages range in length from 40,000 bp to 400,000 bp (40Kb – 400Kb), so being able to manipulate them using computer programs is essential.

A folder of all pertinent files needs created in the Master files. Please see Debbie to do this. Save each and every document as you go, labeling them carefully. Record how you label them in the annotation folder and your notebook. Also, create and Old Stuff folder in that phage folder. As you generate an updated version of a file, place the older version in the old stuff folder. This will avoid confusion in the future.

Once base 1 is established, e-mail the FASTA sequence file to Dr. Jeff Lawrence. He will post it on the PBI website at http://pbi.bio.pitt.edu/

Example of a simple FASTA file

```
> seq1 This is the description of my first sequence.
AGTACGTAGTAGCTGCTGCTACGTGCGCTAGCTAGTACGTCA
CGACGTAGATGCTAGCTGACTCGATGC
> seq2 This is a description of my second sequence.
CGATCGATCGTACGTCGACTGATCGTAGCTACGTCGTACGTAG
CATCGTCAGTTACTGCATGCTCG
```

FASTA is probably the simplest of formats for unaligned sequences. FASTA files are easily created in a text editor. Each sequence is preceded by a line starting with >. The first word on this line is the name of the sequence. The rest of the line is a description of the sequence (free format). The remaining lines contain the sequence itself. You can put as many letters on a sequence line as you want, and a FASTA file can have as many lines as necessary to include your phage genome sequence.

Blank lines in a FASTA file are ignored, and so are spaces or other gap symbols (dashes, underscores, periods) in a sequence. Any other non-amino or non-nucleic acid symbols in the sequence should produce an appropriately strident string of warnings on your terminal screen when you try to use the file.

Once a unit-length genome has been generated, the file needs to be formatted for use by DNA Master.

1. Open the Fasta-formatted file of your genome sequence
   Phage name.txt

2. DNA Master has special format for its input

- replace the Fasta description line (>blahblahblah) with two periods
  ..(return)
  ATGATCGGATTTGATGCGCGGATGACCTGGAGCTTTAA continuing to the end of the sequence
- Save this file as **Phage nameDNAmas.txt**

3. Open DNA Master, Then open your input file in DNA Master
   - **Open File ==> Degenerate DNA**

4. What is the length of the Genome? _____ bp

5. Check out the base composition
   - **DNA ==> DNA Composition ==> Table**

6. What is the percentage of GC in this genome? _____

7. What is the significance of that percentage? _____

Each program that is used detects specific features of the genome. The first program we will run is tRNA Scan. Like its name implies, it is used to detect putative tRNA producing sequences. Instead of making proteins, these are copied into tRNA molecules. This program will identify whether your genome has any tRNAs and give you output (printouts) of what each one looks like.

8. We'll search for any tRNA genes using the tRNAscan server:
   - http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
   - Notice that it only accepts up to 100,000 bases – If your phage genome is larger than 100Kbp, you'll need to copy/paste only the first 82 Kb or so (anything less than 100kb) region of your genome (When you cut the genome, be sure to overlap, so as not to miss any tRNAs. Also, RECORD exactly the base pairs you cut & paste so that you can calculate exactly where the tRNA is located when you are working with the entire genome.)
   - Select search mode other (browse for your FastA file)
   - Select **bacterial tRNA**
   - **Run** program

     tRNAs found_____

Send the FastA file to Craig Peebles and he will check for tRNAs and tmRNAs using a different program, Agarorn. cpeebles@pitt.edu

Another program you will need to use is Frameshift Programmed Frames. Written by Jun Xu, this program will identify areas of possible -1 frameshifts. This is based on his earlier work that found g and t ORFs directly preceding the tape measure gen in ___? The research found that the code of g sometimes undergoes a frame shift to make a protein gt. The ratio of g/gt can predict tail assemble efficiency. It is a web-based program that uses the FASTA sequence. Copy and paste the sequence in the appropriate place. Always change the output format to coordinates. Use L5 (or closest choice) as the coding potential model. The RBS(ribosomal binding site) model is M. tuberculosis (is annotating a Mycobacteriophage). All of these can be deselcted (and should be) and run. Print the output. This information will go into the final map and be explained in the analysis.

The next programs (**GeneMark** and **Glimmer**) try to predict where the protein-coding genes are ==> make sure to use the **FASTA sequence** (see step 1) as input. GeneMark is a program that highlights protein-coding regions in a graph-analysis fashion. The gene sequence is identified numerical with 'up-ticks' on the lines to denote starts and down-ticks' to denote stops. (Note: TTG is not used in this program as a start, so "leucine starts" can be missed.) The peaks on the graphs represent good coding potential predictions, according to a model organism, and the highlighted areas signify where this program might call genes. The GeneMark Program calls genes based on similarities to the closest species (which you specify).

9. Open the GeneMark server:
   - http://opal.biology.gatech.edu/GeneMark/genemark24.cgi
   - hit **Browse** to select your input file (you will use a FASTA file)
   - Select the closest species of organism or host as the model (*M. tuberculosis* for the Mycobacteriophages)
   - Under **graphics export options**, select everything except "generate postscript" & "mark putative exons". In the second column choose only 'list open reading frames' and 'list regions of interest'.
   - Run Genemark (Start)

10. You should see a text output of your GeneMark results
    - hit the **view PDF graphic** link
    - **save this file with a new name*****

Glimmer is another computer program used as a tool to predict genes. It calls the genes based on their coding potential. It gives you a text file that identifies the genes, using large predicted orfs in your phage genome as a model gene. The following are directions on how to use the program.

11. Glimmer is a command-line program, so your input file (your phage genome FASTA dile) needs to be in the same folder as the Glimmer program.
    - Place a copy of your input file in this folder:
      C:\Program Files\glimmer2.02_Win
    - **Open glimmer (Start ==> Programs ==> Biology)**
    - Run glimmer with the syntax:
      Run-glimmer2 [input file name]

    Go to command prompt
    Go to H drive
    Cd downloads
    Cd gliimer
    Cd glimmer 2.02
    Run glimmer2 [input file] (*.txt)

    Output comes back as XXX putative genes extracted, results will go into the folder of the original file named g2.cord. All records come back with that file name, so immediately go to the folder and change the name to your phage. Copy that file back in to your phage's folder.

12. Your output should be in the glimmer2.02_Win folder, called **g2.coord**
    - **Open** your results in Word
    - Adjust Margins and **Save** a copy
    - **Print** out a copy

Helpful HINTS!!

DNAMaster

Save DNA Master file often in case of crashes.
In the DNA Master frames, a full vertical line is a stop, a partial line is start.
Everytime you open DNA Master you must tell it that TTG is a start!
The frame number on DNA Master is not necessarily the same frame in Genemark or Glimmer.

Glimmer

In Glimmer, Gene 1 can be at the end of the list (if so, move it to the top).
Glimmer almost always calls the longest orf which is not necessarily the best choice.

GenMark

The Genemark output and graph shows no TTG starts.
An uptick on the Genemark graph is start (the longer uptick denotes ATG, and the smallerone denotes GTG).
A downtick on Genemark graph is a stop codon(TAG, TGA, TAA).

General

The numbering of Genemark and DNA Master protein sequence preditions includes stop codons, Glimmer does not include the number of base pairs for the stop codons.
A few extra genes may be in the Glimmer and Genemark outputs.
A few genes mayn be missed by Glimmer and Genemark.
Remember, genes are tightly packed in phage genomes, so there should be few gaps.
Small (a few amino acids) gene overlaps are OK. Big gene overlaps are not OK.
Often the stop codon and the start codon of adjacent genes overlap.
Annotating negative frame genes is somewhat more difficult, it may be helpful to look for the stop of the next gene before you call the start of the current gene.

Annotating the Genome of a Bacteriophage – Part 2

Armed with all the information you have from part 1 (you have those printouts in front of you, right?), you're ready to start calling genes. Stops are stops, so you can rest assured that when you run into an asterisk, you are at the end of that gene. (This is assuming that no nonsense suppressor or tRNAs are present.) However, picking gene start codons can be tricky. Gene startscalls are based on input from Glimmer, GeneMark, how closely the ends fit, the length of the gene, and the Shine Delgarno score. This score represents the nucleotide sequence (AGGAGG) that is present in the 5'-untranslated region of many prokaryotic mRNAs. This sequence serves as a binding site for ribosomes.

1. **Open** up your entire phage sequence in DNA Master (see part 1, step 3)

2. **Save** a copy of this file with today's date in your folder [Phage name Today's date.seq]

3. Make sure DNA Master recognizes all three of the possible phage start codons:
   - **Select** any piece of sequence
   - **DNA ==> ORF ==> Cues**
   - Make sure ATG, GTG, and TTG are selected as start codons
   - Hit the red check-mark to save changes

4. **Select** the first 5000 bases of your sequence, and display all six reading frames
    - **DNA ==> FRAMES**
    - In each reading frame, the long vertical lines represent stop codons, and the smaller vertical lines indicate start codons.

5. Consulting your Glimmer printout, Genemark printout and the Frame analysis (in DNAMaster) information, **select** your first gene
    - Clicking in the Frame Analysis window selects an open reading frame (green line from start codon to first stop).
    - Notice that the ORF coordinates show at the bottom left corner of the window.
    - To erase the green lines, hit the exclamation point icon in bottom right of the Frame Analysis window.

6. Once you've highlighted your first gene in the Frame Analysis window, switch back to your sequence view
    - **Window ==> Sequence**
    - Notice your ORF is highlighted

7. Now you have to choose the best start codon for this gene:
    - **DNA ==> ORF ==> CHOOSE START**
    - Using the scores here, coupled with your Glimmer and GeneMark results, choose the best start site. The usual rule of thumb is to choose a start site with a high schore that is early in the ORF and that minimizes gene overlaps.

8. Now you're ready to document this first gene (annotate)
    a. Make sure the correct ORF with the best start codon are selected in the Choose ORF Start window, and hit the **Document** button
    b. Document the Gene as '1"
    c. Document the Product as "gp1"

9. Also document (**highlight in color**) this gene on your 6-frame translation printout, Glimmer printout, and Genemark printout. Going from top to bottom on the six frame, use the following colors:
    - Fr1=purple
    - Fr2=pink
    - Fr3=orange
    - Fr-1=blue
    - Fr-2=yellow
    - Fr-3=green

10. Also put the number of this gene next to the highlighted info on all these printed documents

11. Go back to your Frame analysis window, and move on to your next gene.
    - **Note** that each gene can be in a different reading frame, but the genes will not usually overlap each other
    - **Repeat** steps 5-10 above, using the next consecutive number (2) for gene/product name

12. Call all the genes in your DNA segment

13. **Save** the DNA Master file. (Save often as you work!)

Annotating the Genome of a Bacteriophage – Part3

BLASTing the predicted gene products

Once you have called all of your phage genome genes, you may want to ask if you might be right. Without a gene product (an isolated protein or RNA), you are calling putative genes (ones you think that are there, based on all the factors you have learned). One way to check your prediction is to compare them to other genes that have been called in other genomes. Do they match base-pair for base-pair or nucleotide for nucleotide? How close of a match is it? Is the function of the gene known? What kind of genes does your phage genome have?

IN DNA MASTER:

parse the orfs
Select gene 1
Copy as translation
Go on the web to NCBI Blast Proteins, BlastP
Paste in the amino acid sequence
Blast and Record Results

Individually Blasting each predicted protein sequence one at a time is slow, and faster "batch" Blasting is a good way to proceed once you have mastered the slow one-by-one approach. Each technique has its own merits, you may Blast one gene when calling, but then Blast all protein coding sequences at the end to compare genomes.

Dotter Plots
Dotter plots are great ways to compare the genomes. The program allows a comparison of two genomes are as m any genomes as you can to string together. The time it takes to run is related to the square of the size of the files.

This program uses a sliding window to compare DNA code. We typically use a 25 bp window for 'sliding'.

You will need to move a copy of your FASTA file into the Dotter folder.
In the command prompt window, change directories to dotter
H:\dotter>dotter your file.txt your comparison file.txt
Unlike other DOS programs, even though the next prompt automatically appears, it DOES NOT mean that the program is completed.

Once done, you want to open the file. You can remove crossbar (right click on image) and adjust the grayscale. We typically print two versions, one light and one dark, to ensure the lack of manipulation of data. To print, be sure to deselect mail and select copy, then hit OK. Once you close the program you can no longer manipulate it. Be sure you do all the things you want to so, otherwise you will run it again.