Understanding bacteriophages through genomic characteristics

Elizabeth A. Lagesse
University of California, Santa Cruz

Introduction:

Bacteriophages are an exciting area of research for a number of reasons, and high-thoughput sequencing techniques have put their genomes within the reach of more investigators than ever before.

In this project I attempted to infer possible hosts for 9 novel phages by investigating the properties of their genomes.

Specifically, I attempt to learn which bacteria may serve as hosts for these viruses by comparing their GC content and relative codon biases.

Methods:

Using the Python programming language, I created a suite of tools to analyze the genomes of 9 novel bacteriophages previously collected from a 60°C pool with a pH of 9.

1. Each genome was analyzed separately:
   - Open reading frames (ORFs) with length greater than 100nt for each frame
   - Theoretical isoelectric point (pI) for each hypothetical protein
   - GC content of possible coding regions
   - Relative codon usage bias in putative genes

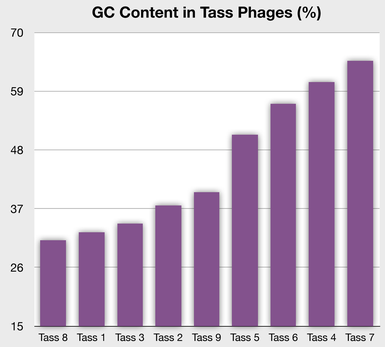
2. The 9 bacteriophages were compared to each other:
   - Average CG content
   - Average (highest/lowest) pI for each phage
   - Patterns in codon usage bias
   - Average ORF length
   - Number/length of ORFs for each phage

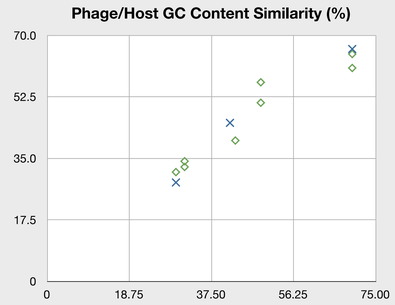
3. CG content and codon usage bias were compared to the literature in search of possible hosts among genera of bacteria known or suspected to exist at the sampling site (Thermus, Clostridium).

Results:

The first trait I examined was the GC content of the 9 new phages.

Here you can see that there is a wide range of values, which might indicate a wide range of possible hosts. (left)

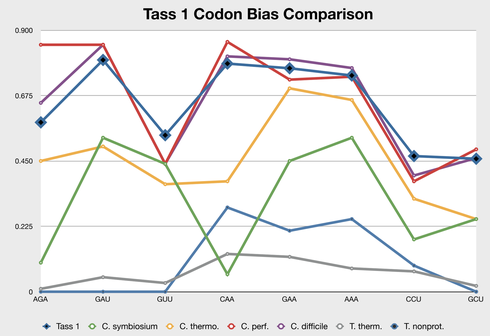


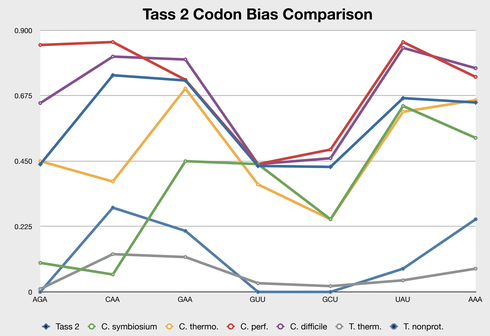


I've also included a graph (right) of the correlation between the GC content of the host genome (horizontal axis) and that of the phage genome (vertical axis), for known pairs as well as possible matches in my data. As you can see, the GC contents of the known phage/host pairs are very close. This agrees with the literature, where you often find very tight correlations.

The second trait I approached were the relative codon biases. Since bacteriophages are reliant on the host cell's machinery for translation, you might expect to find a correlation in codon preferences. For this comparison, I used several species in the *Clostridium* and *Thermus* genera, which are thought to be present in the sampling location.

Examples of plots generated for each phage: (horizontal axis is the fraction for each codon)





Acknowledgments:

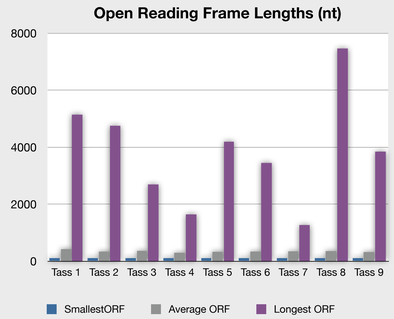
Thank you to Dr. David Bernick for being my advisor for this project, and for teaching me Python!

Conclusions/Next Steps:

This project is definitely a work in progress! It seems clear that there are some paterns in the data, which may give clues as to the hosts of these bacteriophages. (see codon bias and GC content charts)

In addition to the traits I've partially explored here, there are a number of additional ways to look at these genomes. Some ideas for future exploration:

- Hypothetical genes and their protein products



- Searches against known organisms to better understand phylogenetic relationships
- Implications of the varying isoelectronic points of the hypothetical protein products
- When possible hosts have been identified by computational means, wet-lab experiments to confirm these relationships should be performed