Using Genomic Characteristics to Propose Hosts for 9 Novel Bacteriophages

Elizabeth Lagesse

Chemistry Department: University of California, Santa Cruz

1 Introduction

The compositional characteristics of an organism's genome and proteome can give a great deal of information about the ways that organism makes a living. In the case of bacteriophages, their reliance on the host's translational equipment gives a valuable link between host and phage, which can be used to propose host candidates for novel phages.

Phage research in general has experience an increase in recent years. In part this is related to the advent of sequencing techniques that make it relatively easy to make new discoveries, but interest has also been spurred by the decreasing effectiveness of our antibiotic arsenal. It is hoped that some phages may provide suitable substitutes against human pathogens. In all of these efforts, computational methods for connecting a phage to its host can be valuable.

In this project, I examined the genomes of 9 bacteriophages isolated from a hot spring at the Tassajara Buddhist Retreat in California. The pool had a temperature of 60°C and a pH of 9, implying that they are somewhat thermophillic. Previous work had also indicated that members of the *Thermus* and *Clostridium* genera may be present in the area. With this in mind, I chose several members of those genera to consider as possible hosts. For out group comparison, I chose a halophilic organism (*Haloferax volcani*) and an ordinary, non-thermophillic bacterium from the *Escherichia* genus (*Escherichia Coli* 063).

2 Methods

Using the Python 3 [1]programming language, I created a suite of tools to analyze genomic data, including an open reading frame (ORF) finder, and tools for finding the isoelectronic point (pI) of hypothetical protein products, relative codon bias, and GC content.

Each genome was first analyzed separately:

- Open reading frames (ORFs) with length greater than 100nt for each frame
- Theoretical isoelectric point (pI) for each hypothetical protein product
- GC content of possible coding regions

[1]Working in Python 3 (rather than 2.7, which is still used regularly) will help insure forward compatibility of the tools developed.

- Relative codon usage bias in putative genes

And then 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

CG content and codon usage bias were then compared to the literature values for the potential hosts and comparison species.

3 Results

The final software package consisted of a module (SeqTools.py) containing the main methods used to analyze the genomes, and a program (named tassAnalysis.py) that compared the results and formatted them for printing.

These were used to generate the following data:

Genome	% GC Content (%)	Average pI	Number of ORFs	Average ORF Length
tass1	32.54	7.91	412	416.8
tass2	37.57	7.92	310	328.0
tass3	34.17	7.25	115	356.3
tass4	60.70	8.20	148	285.3
tass5	50.81	8.21	168	324.0
tass6	56.63	8.02	153	329.9
tass7	64.67	7.87	114	337.5
tass8	31.08	7.66	560	346.6
tass9	40.06	7.98	259	314.1

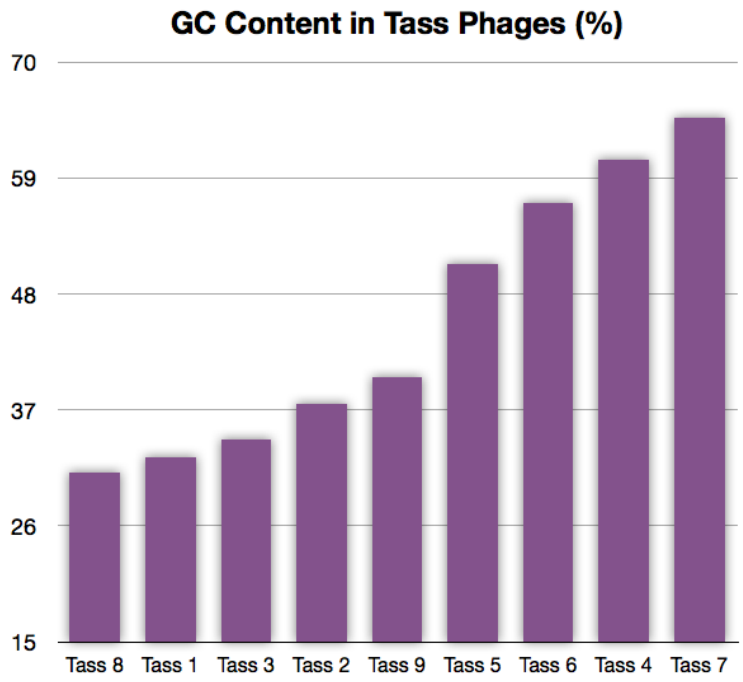

Figure 1: GC content of all 7 genomes

In addition to the above table, the program generated several other pieces of information about the individual genomes and the individual ORFs found in each. While this information was not used in the analysis described below, I believe it may be useful for future investigations. (I have included the entire output file as an appendix. Also attached are spreadsheets and text files containing work in progress, as well as the relevant statistics for the comparison organisms.)

When compared to the published values for my comparison organisms some interesting clustering appears to occur. In particular, Tass 1 and 2 have GC contents and codon biases consistent with members of the *Clostridium* genera. Tass 4 and 7 seem to be consistent with members of the *Thermus* genus. (Analysis of the other phages, and further exploration of these, is a work in progress)

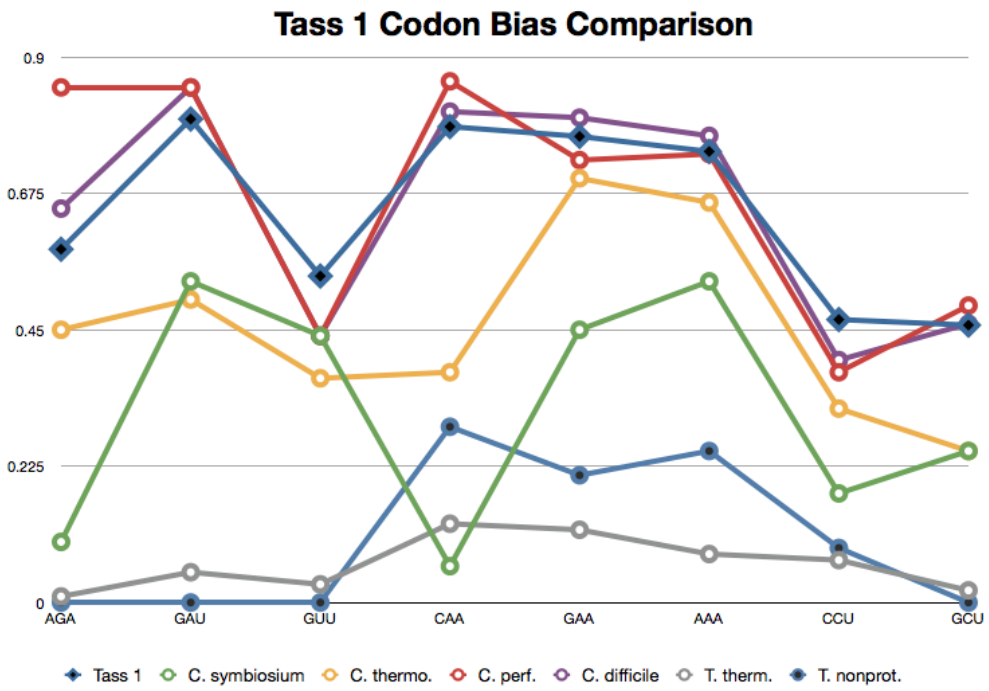

Figure 2: Comparison of tass 1 with various test species

4 Discussion/Conclusion

These programs can be useful any time you want to compare 2 or more genomes on the basis of any of the available factors. Since the main part of the code is written using an object-oriented style, the various parts should be easy to reuse for slightly different purposes as well.

The findings regarding the apparent clustering of the Tass phages into two groups are interesting, and merit further work. I plan to continue comparing my results to the comparison organisms, and to expand to considering more than just the GC content and codon bias. I will also perform statistical tests on the trends I believe I have found.

5 Acknowledgements

Thank you to Dr. David Bernick for advice on this project!

6 References

Thermus thermophilus Bacteriophage fYS40 Genome and Proteomic Characterization of Virions (Naryshkina, et. al)
Structure and Predicted Functions of Putative Thermus Phage TASS7s DNA (Unpublished - Patrick Mueller)
Genome Landscapes and Bacteriophage Codon Usage (Lucks, et. al).
The Genomic Structure of Thermus Bacteriophage uIN93 (Matsushita, et. al.)

Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species (Zheng, et. al)