**Tetranucleotide usage in mycobacteriophage genomes: alignment-free methods to cluster phage and infer evolutionary relationships**

Benjamin Siranosian1,2, Emma Herold2, Edward Williams2, Chen Ye2

1Center for Computational Molecular Biology, Brown University, Providence, RI

2Division of Biology and Medicine, Brown University, Providence, RI

**Abstract**

**Background**

The genomic sequences of phages isolated on mycobacterial hosts are diverse, mosaic and often share little nucleotide similarity. However, about 30 unique types have been isolated, allowing most phage to be grouped into clusters and further into subclusters [1]. Many tools for the analysis of mycobacteriophage genomes depend on sequence alignment or knowledge of gene content. These methods are computationally expensive, can require significant manual input (for example, gene annotation) and can be ineffective for significantly diverged sequences [2]. We evaluated tetranucleotide usage in mycobacteriophages as an alternative to alignment-based methods for genome analysis.

**Description**

We computed tetranucleotide usage deviation, the ratio of observed counts of 4-mers in a genome to the expected count under a null model [3]. Tetranucleotide usage deviation is comparable for members of the same phage subcluster and distinct between subclusters. Neighbor joining phylogenetic trees were constructed on pairwise Euclidean distances between all genomes in the mycobacteriophage database. In almost every case, phage were placed in a monophyletic clade with members of the same subcluster. With few exceptions, trees computed from tetranucleotide usage deviation accurately reconstruct trees based on gene content for a subset of the mycobacteriophage population (Fig.1).

We also evaluated the possibility of assigning clusters to unknown phage based on tetranucleotide usage deviation. Under a simple nearest neighbor classifier, cluster assignments were recovered at a frequency greater than 98%.

In addition, we looked for evidence of horizontal gene transfer by using tetranucleotide difference index, a measure of the deviation in tetranucleotide usage from the genomic mean in a sliding window across the genome [3]. Tetranucleotide difference index plots showed a strong spike at the end of cluster L mycobacteriophages, which could indicate horizontal gene transfer in the region.

**Conclusions**

Genome analysis based on tetranucleotide usage shows promise for evaluating host-parasite coevolution and gene exchange within the mycobacteriophage population. These methods are computationally inexpensive and independent of gene annotation, making them optimal candidates for further research aimed at clustering phage and determining evolutionary relationships. Code for genome analysis and data used in this project are freely available at [https://github.com/bsiranosian/tango\_final].

**Introduction**

1. Mycobacteriophages
   1. Diversity
      1. Clustering
      2. Phenotype
      3. Genotype
   2. Collection methods
      1. Phage Hunters
      2. How this research came about
2. Alignment methods
   1. How genomes are clustered
   2. Concept of phams
   3. Drawbacks of alignment-based
3. Alignment-free methods
   1. General overview
   2. Specific methods
   3. Advantages
   4. Previous research
      1. David Pride
      2. Bacteriophage in gut microbiome

Bacteriophages isolated on mycobacterial hosts represent an incredibly diverse and widespread pool of viruses. Part of the estimated 1031 phage particles present at any one time on earth, mycobacteriophages play a large role in the turnover of bacteria (ref). As of April 2014, 663 mycobacteriophage genomes were available on the database phagesdb.org (ref). Phenotypically, mycobacteriophages fall into the class siphoviridae or myoviridae. Short tails, long tails, capsid dimensions and tail dimensions. Applications for treatment of TB.

The genomic sequences of mycobacteriophages are mosaic and diverse. Global GC content ranges from 50.3% to 70%, averaging 63.9%. Genome lengths range from 41kb to 165kb and average 67kb. Greater than 50,000 genes are found within the population but can be sorted into almost 4,000 groups, called phamilies, based on their shared amino acid sequence (ref). The majority of genes found in mycobacteriophages are of unknown function and do not have homologs in other phage or bacteria (Hatful 2014 information, should spice up a bit so it’s not a direct copy…). Mention something about “biological dark matter”

Despite the large diversity present in the population, mycobacteriophages can be grouped into distinct clusters based on their morphologic and genetic features. Some clusters are large and further divided into subclusters (cluster A, for example, with 11 subclusters and 246 members) while other are small (cluster S with 2 members). Some isolated phage have no nearest neighbor to define a cluster and are defined as singletons until further sampling supports their inclusion in a cluster. Mycobacteriophages are traditionally clustered based on four methods: dot-plot comparisons, pairwise average nucleotide identities, pairwise genome map comparisons and gene content analysis. These methods require sequence alignment or genome annotation – two classes of methods that are effective but computationally expensive and may require significant manual input (automated gene calls are verified by hand before a mycobacteriophage genome is submitted to genbank, for example).

Studying taxonomy of viruses is especially difficult because they lack a common subsequence on which to compute alignment based statistics, such as 16S rRNA in bacteria (ref 13-15 from Pride 2006)

Alignment-based methods do have their advantages and have served mycobacteriophage researchers well in the past. However, no groups to date have looked at alternative methods for analyzing these genomes. Specifically, a class of methods known as alignment-free methods. These techniques typically use statistics based on the occurrence of oligonucleotides in a sequence to compare genomes. Several different methods have been proposed and developed, such as D2, D\*, etc. For an excellent review, see the Vinga review paper. Alignment-free techniques avoid many of the problems associated with alignment-based methods, such as EXAMPLE and EXAMPLE. They are also much easier and faster to compute – alignment-free statistics can be computed in seconds on a personal laptop for the entire sequenced mycobacteriophage database, while pairwise genome alignment takes days on a computing cluster.

There are some drawbacks to alignment-free methods for analyzing genomes, mostly with the interpretation of the statistics in an evolutionary context. It can be difficult to understand how oligonucleotide frequencies are modified in a population over time when selection usually takes place at the level of genes.

Despite these drawbacks, alignment-free methods have been used to study genomic sequences in a variety of contexts. For example, Pride et al. (2006) found tetranucleotide usage in bacteriophages and host organisms to be similar. More recently, Ogilvie et al. (2103) surveyed metagenomic sequencing datasets using a tetranucleotide usage based method and discovered several novel bacteroides phage, the majority of which could not be identified with alignment-based methods.

We examined tetranucleotide usage in the mycobacteriophage population as an alternative to alignment based techniques. Tetranucleotide usage supports the grouping of phage into subclusters similar to the published techniques. However, subclusters of the same parent cluster were not always related, leading to some surprising conclusions.

The number of sequenced mycobacteriophages has grown immensely in the past few years thanks to the “Phage Hunters” program administered through the SEA Education Alliance and Howard Hughes Medical Institute. This program allows first year undergraduate students to isolate and characterize a novel mycobacteriophage from the environment. It has also provided excellent opportunities for collaborative projects between undergraduates, resulting in the work you see here.

**Methods**

1. Phage sequences
   1. How we got
   2. Number
   3. Draft genomes
2. Oligonucleotide usage deviation
   1. Method
   2. 2-6
3. Comparisons of genomic windows

We obtained the genomic sequences of all 663 sequenced mycobacteriophages available on the website phagesdb.org as of April 2014. In addition to genomes that have been published in Genbank, this dataset contains unpublished mycobacteriophage sequences that have not yet been annotated for genes. There is no easy way to download the mycobacteriophage database in its entirety, so we used a simple python script to automate the process, available in our code [link].

To compare mycobacteriophage sequences in an alignment-free way, we first investigated the usage of *k*-mers (substrings of DNA of length *k*) in each genome. This is a composition vector approach, where the genome is summarized by a vector of frequencies of each k-mer (REF). To remove biases from the background nucleotide composition of the sequence, we normalized the frequencies of k-mers using a zero-order markov model. This method was examined by Pride et. al (2003, 2006) and found to be effective for analysis of prokaryotic genomes. Formally, k-mers are counted in the genome using a sliding window of size k and step size of 1. The expected number of a k-mer *W* given the background nucleotide distribution is given by:

Where A,T,C,G are the frequency of each nucleotide in the genome, a,t,c,g are the number of each nucleotide in the *k*-mer, and N is the length of the genome. Observed counts of each k-mer are then divided by the expected number to give the final composition vector.

We computed the composition vector for k in the range of 2 to 6. We chose to conduct most of our further analyses on composition vectors of k=4 because of …. (REF, EXPAND). This is equivalent to the tetranucleotide usage departures from expectation measure proposed by Pride et al (2003). However, we note that many of our results are consistent for varying values of k.

Comparison of phage genomes

To compare phage genomes in terms of tetranucleotide usage, we calculated the Euclidean distance between composition vectors. Done for the 663 sequences we analyzed, this created a distance matrix capturing relationships between entire phage genomes.

For analysis of the subset of 60 phage in Hatful et al. (2010), we used the splitstree program (REF) to construct neighbor joining phylogenetic trees. This was done to facilitate easy comparisons between trees published previously and our alignment-free trees. Splitstree was used to compute the neighbor joining trees from distance matrices on tetranucleotide usage. Analysis of larger sets of phage genomes proved too time consuming for splitstree to handle, so we used hierarchical clustering within R (version 3.1.0) do construct dendrograms. To compute bootstrap values for our cluster dendrograms, we used the package ‘pvclust’ version 1.3-0 (REF), which computes the ‘approximately unbiased’ and ‘bootstrap probability’ values for splits within the dendrogram.

RESULTS

1. Tetranucelotide usage reconstructs alignment-based phylogeny
   1. Comparison with hatful subset
   2. Supplement – k 2-
   3. Whole database phylogeny / clustering
2. Tetranucleotide usage in windows highlights genome structure
   1. Areas of self similarity and difference
   2. In depth analysis of certain genomes
3. Exceptional k-mers
   1. B3 GATC

We were first interested if relationships between mycobacteriophage genomes constructed from tetranucleotide usage could accurately recreate relationships from alignment-based statistics. In particular, does a clustering scheme based on tetranucleotide usage agree with clusters previously assigned to the phage? To test this hypothesis, we examined a subset of 60 mycobacteriophages first analyzed by Hatful et al. (2010). In this first paper, a clustering scheme is proposed based on dot-plot comparisons, pairwise average nucleotide identities, pairwise genome maps and gene content analysis. However, it should be noted that the clustering scheme proposed for these phage mainly serves to identify elements of similarity and horizontal gene transfer. The clustering scheme, and our methods of grouping based on tetranucleotide usage, are not true taxonomic representations of the mycobacteriophage population (Lima-Mendez et al. 2008) (rephrase this to sound less like the Hatful paper).

We used the splitstree program (REF) to construct a neighbor joining tree from tetranucleotide usage distances between the subset of phage (fig 1a). There are several points of similarity between this tree and the tree

DISCUSSION / CONCLUSION

Application to metagenomics sequencing and novel virus discovery

REFERENCES

1. Hatfull GF: **Mycobacteriophages: Windows into Tuberculosis**. *PLoS Pathog* 2014, **10**:e1003953.

2. Vinga S, Almeida J: **Alignment-free sequence comparison—a review**. *Bioinformatics* 2003, **19**:513–523.

3. Pride DT, Wassenaar TM, Ghose C, Blaser MJ: **Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses**. *BMC Genomics* 2006, **7**:8.

4. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko C-C, Weber RJ, Patel MC, Germane KL, Edgar RH, Hoyte NN, Bowman CA, Tantoco AT, Paladin EC, Myers MS, Smith AL, Grace MS, Pham TT, O’Brien MB, Vogelsberger AM, Hryckowian AJ, Wynalek JL, Donis-Keller H, Bogel MW, Peebles CL, Cresawn SG, Hendrix RW: **Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size**. *Journal of Molecular Biology* 2010, **397**:119–143.