Viruses that infect bacteria are called phage. There are two distinct lifestyles of phages: lytic and lysogenic. A lytic lifestyle is when a phage infects a bacteria; replicates itself many times; lyses the host bacteria; and releases all the newly created phage. A lysogenic lifestyle is when a phage infects a bacteria; inserts its DNA into the bacteria genome; and replicates when the host bacteria divides.

The lifestyle of a phage is important to more accurately classify and understand phages and their environments. Shotgun sequencing now allows us to sequence entire environmental communities, where isolation of individual phages is impossible. Also of great importance are the fields of phage therapy and biocontrol. With the increase of antibiotic resistant bacteria and ever increasing food needs, both fields are seeing a resurgence. Since phages have the potential to infect antibiotics resistant bacteria, the need to control them is even greater.

Additionally, we are currently seeing a resurgence in the fields of phage therapy and biocontrol. With the increase of antibiotic resistant bacteria and ever increasing food needs, both fields are seeing a resurgence. Since phages have the potential to infect antibiotics resistant bacteria, the need to control them is even greater.

To test the accuracy of the predictions I sequentially removed each known phage from the database. Treating each removed phage as an unknown, I predicted the lifestyle of each phage. Out of the 228 known phages in my database I was able to correctly predict the lifestyle of 218 phages.

Seven of the phages that were incorrectly classified were the phages that had the lowest probability scores out of all 228 predictions. Subsequently the phage that was incorrectly classified with the highest probability score, was a lytic Lactococcus phage that was predicted to be a temperate phage. The reason for such a high certainty score was because this phage genome contains a functional integrase. Integrases are important proteins for temperate phage to incorporate themselves into host genomes. Cross over events like this are just one of the reasons that make genomic phage analysis difficult.

Incorporating the Random Forest algorithm, I calculate how important each protein is towards classifying phage lifestyles. I exclude any proteins that are less than two standard deviations above the mean. The Random Forest algorithm generates many classification trees from a training set of data. These trees are used to predict the class of a testing set by finding the mode of the predictions of all the individual trees.

Basic Computational Algorithm of PHACTS:

1. 50 random lytic phage and 50 random lysogenic phage are selected.
2. 300 important lytic proteins and 300 important lysogenic proteins are selected at random.
3. Each of the 600 proteins is then queried against each of the 100 random phage proteomes, and a the highest FASTA similarity value is returned.
4. These 100 vectors, of 600 similarity scores, become a training set for the Random Forest algorithm to create a classification scheme.
5. A testing set is created by querying each of the previous 600 proteins against the unknown phage proteome.
6. Using the classification scheme created by the Random Forest algorithm, the lifestyle of the unknown phage is predicted.

The output of the Random Forest prediction is a probability score for each class that is calculated by dividing the number of decision trees predicting that class by the total number of decision trees. To better account for error in the Random Forest model and random protein selection, I repeat steps 1-6 ten times and average the prediction probabilities.

Seven of the phages that were incorrectly classified were the phages that had the lowest probability scores out of all 228 predictions. Subsequently the phage that was incorrectly classified with the highest probability score, was a lytic Lactococcus phage that was predicted to be a temperate phage. The reason for such a high certainty score was because this phage genome contains a functional integrase. Integrases are important proteins for temperate phage to incorporate themselves into host genomes. Cross over events like this are just one of the reasons that make genomic phage analysis difficult.

REFERENCES

1.) The SEED Database www.theseed.org
2.) R programming language www.r-project.org
3.) Random Forest Code www.berkeley.edu/users/breiman/RandomForests/
4.) FASTA http://fasta.bioch.virginia.edu/