Using Dot Plots for Comparative Genomic Analyses: Learning Activity Outline

## Authors

Andrew Kapinos, Canela Torres, Amanda Freise, Sean McClory, and Kathleen Cornely

##

## References

Krumsiek, J., et al., *Bioinformatics Applications Note* **23**:1026 (2007).

Smith, K. C., et al., *BMC Genomics* **14**:410 (2013).

## Introduction

The number of actinobacteriophages isolated and sequenced by students participating in the SEA PHAGES program has provided a wealth of information that can be used to examine differences in nucleotide sequence. Whole genome comparison of hundreds of sequenced phages has led to the assignment of phages with similar genomes to “clusters” and “subclusters”. Comparing phage genomes within and among clusters may reveal evolutionary relationships.

A simple way to compare DNA sequences is through the use of a “dotplot” which is a simple statistical tool in which a graphic display is constructed to allow a comparison of the frequencies within different categories. The application presented here involves the comparison of DNA sequences. In this activity, we will learn how to construct dotplots using the Gepard method and then we will use our dotplots to explore differences in nucleotide sequences among various phages.

While whole genome analysis is the method used to assign a newly discovered phage to a particular cluster, Smith et al. demonstrated that for a number of selected phages, comparison of just a single gene—the tape measure protein—allowed for the correct assignment of a phage to its appropriate cluster in most cases (Smith *et al.*, 2013). The ability to assign clusters based on the sequence of a single gene provides a way to cluster phages whose complete genomes have not yet been sequenced. In this way, the Gepard dotplot method can be used to compare whole genomes, or just one specific gene.

## Introduction to Dotplots

Dotplots are a graphical representation of the similarity between two sequences (either nucleotide or amino acid) and can be used to quickly analyze the whole genome relationships between large groups of phages.

In a dotplot comparison, sequences are not compared to one another altogether.

Instead, short chunks of nucleotides/amino acids, called ‘words,’ are compared. To generate a dotplot, the user will specify a ‘word length,’ which describes the stringency of the comparisons between the sequences. The ‘word length’ is, in essence, the number of nucleotides (or amino acids) which must be an exact match in order to place a dot on the plot. In the example here, the word length/size has been set to 1, meaning that each time any one nucleotide is found along both sequences, a dot is placed at that coordinate of the 2D plane. Note how, when the sequences are ***dissimilar***, there appears to be weak, scattered signals. However, when the sequences are ***similar***, there appears to be a thick, strong diagonal signal on the graphical output, indicating that the sequence has been well conserved.

The same goes for larger sequences and larger word sizes. The figure below includes a nucleotide comparison of four phage genomes, using a word size of 15. This means that, every time there is an exact match of 15 nucleotides on both sequences, a dot is placed on the graph. Also, in this case, we have merged the FASTA files containing the whole genome sequences of the phages being compared, meaning that each axis of the dotplot contains four full-length genomes that have been ‘glued’ together. By making the horizontal and vertical input sequences the same, we can generate a multi-FASTA dotplot that allows for informative comparisons between many phages at the same time. You can think of each ‘sector’ where genomes intersect as ‘mini-dotplots’ for just those two phages. Note that, where self-to-self comparisons are made (eg. 1 v 1 below), the signal is always strong, as the sequence being compared is identical at these regions.

## Learning Objectives

After completing this module, students will be able to…

* Prepare concatenated FASTA files in the correct order.
* Understand the purpose of different word lengths.
* Use Gepard to generate high quality dotplot images that are labeled correctly and annotated in an informative and aesthetically pleasing manner.
* Use dotplots to determine similarity or dissimilarity between two or more phages.
* Interpret the dotplot results and draw conclusions about genomic relationships.
* Identify repeat sequences and other features.
* Suggest cluster relationships based on dotplot results.
* Generate a hypothesis [make a prediction] to be tested with a dotplot.

## Student Activities

* Complete the Dotplot Pre-Activity worksheet provided by your instructor.
* Then follow the Learning Activity Instructions to install Gepard on your computer.
* Complete the Post-Activity assessment by using the Gepard program to construct and annotate a dotplot to answer your research question.

: