Using Dot Plots for Comparative Genomic Analyses:

Pre-activity assignment

Dotplots are a great tool for visualizing similarities between sequences. In biology, the sequences we want to compare are of DNA, RNA (nucleotides) and protein (amino acids).

In the simplest example of a dotplot comparison, a dot is placed at the coordinate of the 2D plane each time that a given nucleotide is found in both sequences of two oligonucleotides being compared. When the sequences are dissimilar (as shown in the figure on the left), there appear to be weak, scattered signals. However, when sequences are similar, a thick, strong diagonal signal appears on the graphical output, indicating that the two sequences are conserved.



1. Practice by making a dot plot to compare the two 8-nucleotide DNA sequences shown below. Based on the example shown above, would you argue that these sequences have low similarity or high similarity? Why? Explain your answer to one of your peers using observations from the dot plot you generated.

Sequence 1: 5’-ACGTCAGA-3’
Sequence 2: 5’-CGTCACAC-3’



1. You can also use dot plots to compare protein sequences. Make a dot plot to compare the following two amino acid sequences. The sequences are listed from amino-to-carboxyl terminal, from left to right. One-letter amino acid codes are used. Do these sequences have low or high similarity? Why? Explain your answer to one of your peers.

Sequence 1: MAVHYPES

Sequence 2: MSLGNHHP



1. Dot plots can be used to compare multiple sequences by arranging them in order on the axis. Compare the following three DNA sequences using the template below. Which sequences are most similar? Compare your answer with one of your peers.

Sequence 1: 5’-GAGTAC-3’
Sequence 2: 5’-GTGACC-3’
Sequence 3: 5’-TCGAGA-3’



1. Dot plots are also a good tool to visualize evolutionary relationships between sequences.

Imagine two evolutionarily related bacteria: *Ann sestor* and *Evo lutian*. The species *Ann sestor* has a genetic makeup that is relatively unchanged from the last common ancestor, while several genetic changes occurred in *Evo lutian*. Let’s think about how specific genetic sequence changes might appear in a dot plot.

First let’s look at a short genetic sequence in these two populations that is 100% conserved. Both sequences are exactly the same: **5’-ACTGGA-3’**. The dot plot comparison would look like this:



Now, imagine some simple genetic changes that could have occurred in this sequence instead. On the empty plots below, compare the *Ann sestor* sequence above to the sequences below that contain the indicated mutations:

Base substitution at position 3: **5’-ACCGGA-3’**

Single nucleotide insertion at position 2 and deletion of nucleotide 5: **5’-AGCTGA-3’**

 

What visual indications on the dot plot show that there is a base substitution? What visual indications on the dot plot show that there is an insertion or deletion?

1. One challenge with making dot plots comparing DNA/RNA sequences is that the plots will have a lot of ‘noise’. When comparing similar sequences, many dots will be made on the diagonal (indicating sequence conservation), but many dots will also be made because of chance. When there are only 4 nucleotides that make up a sequence, many matches will be random, not meaningful, and this may hide the matches that indicate conservation.

One way to clear the noise is increase the **word length** compared in the two sequences. Instead of making a match between individual nucleotides, a dot will only be placed if the entire “word” (a string of two or more nucleotides) is identical between the sequences. This will greatly reduce the matches that arise from chance.

For example, on the dot plots below, sequences from two homologous genes are compared. In each dot plot, a different word length is used. Although the plots look different, they are comparing the same two sequences; however, with increasing word length, the ‘noise’ of the background decreases and the similarities become more obvious.



That doesn’t mean that a larger word size is always better. In the plots below, in which four genomes are compared to one other, you can see that at word size 15 there are some moderate similarities between the first phage (in the top row/column) and the other three phages. When word size is increased to 40 those similarities disappear, producing a ‘false negative’ that suggests there is no similarity at all. Making dot plots at several word sizes and comparing them to choose the most informative one is often a good approach.



Now it’s your turn. On the plots below, homologous sequences from phages Skippy and FrenchFry are compared. The plot on the left uses a word length of 1, and each T in FrenchFry is shaded for the T in Skippy. The plot on the right uses a word length of 2, and each TG in FrenchFry is shaded for the TG in Skippy (by shading the upper left corner of the 2 × 2 box made by the intersection on the chart).

Complete the plots below using a word length of 1 on the left and a word length of 2 on the right. (On the right plot, for each 2-nt word that matches, shade in only the upper left corner dot as shown. No need to draw the boxes.) On which chart is the relationship between Skippy and FrenchFry easier to see?

 

**Check your comprehension**

1. The dot plot below shows a comparison between phages Banjo and Skippy. On the dot plot, circle and label a region that shows:
2. A region where nucleotide substitutions may have occurred between the two genomes.
3. A region where an insertion/deletion may have occurred between the two genomes.



1. Which of the two plots below was made using a larger word length? Circle your answer. Defend your choice to one of your peers in the class.



1. Three phage genomes are compared using a dot plot as shown in the figure below. Use the data to answer the following questions in small groups. When you have completed the questions, get together with another group to compare your answers. Be ready to defend your choices.



1. Which two phages are most closely related to each other? Justify your answer using data from the dot plot.
2. On the graph above, circle and clearly label a region of highly similar sequence between Galactic and Yoshi.
3. On the graph above, circle and clearly label a region of dissimilar sequence between Jabbawokkie and Yoshi.
4. On the graph above, circle and clearly label a region where an insertion/deletion might have occurred between the genomes of Galactic and Jabbawokkie.
5. Explain why a linear diagonal line runs from the upper left corner of the plot to the lower right corner.