**MIP 280A4: Microbial Sequence Analysis**

**Dotplots, In-class exercise questions**

1. Create a self-dotplot using the word humpty dumpty and a word size of 2 (leave out the space between humpty and dumpty in the dot plot) (1 pt)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | HU |  |  |  |  |  |
| HU |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

1. Create a self-dotplot using the word racecar and a word size of 1 (1 pt)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | R |  |  |  |  |  |  |
| R |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

1. What property of the word racecar is evident in the dotplot pattern? (1 pt)
2. Create a dotplot between the words profiting and profiteering using a word size of 1 (1 pt):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | P |  |  |  |  |  |  |  |  |
| P |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

1. How do the letters eer in the word profiteering alter the diagonal pattern in the above dotplot? (1 pt)
2. In Geneious, use the NCBI/Nucleotide search function to download the reference sequences for the SARS-CoV and SARS-CoV-2 viruses, which have accessions NC\_004718 and NC\_045512, respectively. SARS-CoV caused the 2003 outbreak: don’t be confused by the strain name (Tor2): this is SARS-CoV-1 strain Tor2, not SARS-CoV-2. Move the sequences into a folder in Genious and select both sequences. Click the dotplot tab to view a dotplot of these two sequences.
   1. In the options on the right side of the dotplot, set the Data Source options to: Low Sensitivity/Fast and Word Size: 20. Are these sequences similar from end-to-end? (1 pt)
   2. Is there evidence in this dotplot of any large scale structural variation between these genomes (duplications, inversions)? (1 pt)
   3. Are these two genomes equally similar across their lengths? Which regions of their genomes are more or less similar than the rest? (1 pt)
   4. Now, set the word size to 8. Note that the diagonal line is more complete. Why is this the case? (1 pt)
   5. Note also with a word size of 8 there is more ‘fuzz’ off of the diagonal (compared to a word size of 20). Why does this happen? (1 pt)
   6. Perform a Needleman-Wunsch global alignment of these two sequences in Geneious. What is the global % identity of these genomes? (1 pt)
3. **From Genbank**, download **the first 100,000 bases** of two *E. coli* reference genomes: for E. coli K12 and O157:H7, which have accessions NC\_000913 and NC\_002695, respectively.

To do this, navigate in a web browser to the NCBI nucleotide database. Search for the above accessions. Use Change Region shown to 1-100,000. Use Send To-> File-> Format Genbank (full) to download the sequences in Genbank format. Drag and drop these into a folder in Geneious.

In the options on the right side of the dotplot, set the Data Source options to: Low Sensitivity/Fast and Word Size: 20.

* 1. Are these sequences *generally* similar from end-to-end? (1 pt)
  2. Is there evidence in this dotplot of any large-scale structural variation between these genome regions (duplications, inversions, insertions, deletions)? (1 pt)
  3. What type of structural variation is evident? (1 pt)
  4. Investigate the large structural variation evident centered around position 21,500 in the O157:H7 genome (NC\_002695). How would you describe this difference between the two genomes? (1 pt)
  5. Select this region of variation by clicking and dragging your cursor in the dotplot. Switch to Sequence View. What types of genes are present in the O157:H7 genome in this region that are absent in the K12 genome? (1 pt)
  6. Perform a Needleman-Wunsch global alignment of these two sequences in Geneious. What happens when you do this? (1 pt)
  7. Open the Geneious Preferences menu. Goto the Plugins and Features tab. Install the MAFFT plugin.
  8. While you are in Preferences, do two more things that will make Geneious more useful:
     1. Plugins and Features -> Customize Feature Set -> Enable Fasta View
     2. Appearance and Behavior -> Check on Show Dotplot view on single sequences (compare to self)
  9. Perform a MAFFT alignment of these two E. coli sequences in Geneious. What is the speed of the MAFFT aligner compared to the N-W global alignment? (1 pt)
  10. What is the *overall* global % identity between these two genome regions? (1 pt)
  11. What is the % identity between positions 1-2000 of this alignment? (1 pt)