**MIP 280A4: Computational Microbiology**

**Pairwise Sequence Alignments, In-class exercise questions**

1. Create the ‘best’ alignment for each pair of words. What ‘best alignment’ means is up to you, but you should think about how you will justify your choice. (2 pts)

**Pair 1**

groan & grown

**Pair 2**

manoeuvre & maneuver

**Pair 3**

camaraderie & cadre

**Pair 4**

Vermiform & formation

1. Invent a scoring system for DNA sequence alignments, and, *given that scoring system*, try to find the highest scoring alignment. (2 pts)

Sequence 1 A G C A A C T T

Sequence 2 A G G C A A C T

Describe your scoring system:

What is the best alignment you came up with given this scoring system?

**Random Alignments**

1. Generate two random DNA sequences of length 200 bp and copy and paste them into Geneious. Create a global alignment and a local alignment in Geneious. Make sure that “Automatically determine direction (slower)” is checked on. Record the percent identies and lengths of these alignments (1 pt):
   1. Global alignment pairwise % identity:
   2. Global alignment length (nt):
   3. Local alignment pairwise % identity:
   4. Local alignment length (nt):
2. Leading question: just because an alignment is produced between two sequences, can you conclude that the sequences are related evolutionarily? (1 pt)

**Affine gap penalties**

1. Given a gap open penalty of 12 and a gap extension penalty of 1, what is the cost (penalty) of the gap shown in the alignment below (1 pt)?



1. The gap in the above SARS-CoV-2 spike sequence alignment is 6 bases long. Why is the length of gaps in coding sequences usually a multiple of 3? (1 pt)

**Zika virus sequence alignment exercise**

1. Download into Geneious the two Zika virus genome sequences in the RefSeq database. Make sure you download the sequences so that annotation is included. Describe how you found the two sequences. (1 pt)
2. In Genious, perform a global alignment (with free end gaps) of the two Zika virus sequences. We are expecting these sequences to be pretty similar to each other, so set the Cost Matrix to: “70% Similarity (IUB)”. Leave other parameters at their default values.

What is the pairwise percent identity (% tab in Geneious) between these sequences? (1 pt)

1. Describe the scoring system that Geneious used to perform this alignment (1 pt)
   1. Match reward:
   2. Mismatch penalty:
   3. Gap open penalty:
   4. Gap extend penalty:
2. Consider the gap around position 1440 of the alignment. Geneious produced an alignment that splits this gap into two gaps. Does that seem like it reflects the likely event that produced this structural variation? Why did the aligner create two gaps instead of one? (2 pts)
3. If you wanted to force this to be one gap only, what parameter or parameters could you change in the alignment scoring scheme to make this happen? [play with the parameters until you make the gap one gap!] (1 pt)
4. Does this gap correspond to an insertion or a deletion in the newer Zika virus sequence? Assume that the older sequence represents the ancestral sequence (1 pt)
5. What is the coding impact of this gap? What Zika virus protein is impacted? (1 pt)