THE CASE OF THE FLORIDA DENTIST

This practical will introduce you to some of the online bioinformatic tools that are used by labs across the world to explore molecular evolution. If you get stuck, just ask for help. Feel free to go back and explore the websites that you meet if you finish.

bold = things to press or do
> CAPITAL = use the menu bar
→ just information
Question to answer

Background

This practical is based on a true story that took place in Southern Florida in the 1990s. The story involves a group of patients who claimed they contracted HIV from their dentist. If this was true, then we would expect the HIV virus isolated from the dentist to be more related to the patients virus than to other 'control' sequences from the wider population. To test this hypothesis we are going to compare HIV gene sequences isolated from the dentist, his patients and the wider population and use this alignment to build a phylogenetic tree.

We will use the *env* gene which codes for the outer coat of the HIV virus. The *env* sequences for this case have been placed into the *GenBank* database – an open access collection of all publicly available nucleotide and their protein translations.

Overview of practical

- Retrieve dentist, control and patient *env* gene sequences
- Compare these sequences using MAFFT into a *multiple sequence alignment*
- Build a *phylogenetic* tree of these sequences

Getting your sequences

First we need to get the *env* sequences. I have got most of them for you from the NCBI website (step 1.) but you still need to get one more in step 2.

1. **Open WordPad** and then **visit the following page**: https://thescienceteacher.co.uk/bioinformatics/
2. **Download the file** called HIV DNA Sequences and **save** a copy of this file to your desktop: call it HIV.txt

→ if you look at this file you can see the DNA sequences of the *env* gene from the patients A-H, dentist and the local controls are all in something called *FASTA format*. Each sequence has a > followed by the title, then the sequence on a new line. I got the sequences from the NCBI website to save time, now you are going to get the final one.

3. Go to NCBI web page http://www.ncbi.nlm.nih.gov/

4. Change the search to search **NUCLEOTIDE**
5. Now **enter** the following GenInfo Identification number **326847** into the box.
6. Press **SEARCH,** this will take you to the GenBank page for this sequence

→ have a look at the information on this page as it contains information on the organism, sequences (nucleotide and protein) and any relevant references. Answer the questions below:

Q1. How long is this sequence?
Q2. What is the last amino acid of the encoded protein sequence?

7. To retrieve the nucleotide sequence for this gene **scroll up** to the top of this page and **Click** on the **FASTA** button
   Select and **copy** this sequence and **paste** it into the HIV.txt file on your desktop that you created in 2.

8. **Rename** this sequence by adding the word **DENTIST** just after the > like I have done below and **SAVE** changes

>gi|326847|gb|M90848.1|HIVFLD1 Human immunodeficiency virus type 1, viral sample FLD1, V3 region

becomes…

>DENTISTgi|326847|gb|M90848.1|HIVFLD1 Human immunodeficiency virus type 1, viral sample FLD1, V3 region

Aligning your sequences

Now that you have all your sequences in one file (in FASTA format) we need to compare how similar the sequences are by carrying out a *multiple sequence alignment* using a program called MAFFT. First we will need to upload all the sequences from your HIV.txt file.

9. Visit the website: https://www.ebi.ac.uk/Tools/msa/mafft/
10. Select DNA
11. **Upload your sequences** from your HIV.txt file

Now we are going to get MAFFT to align the sequences so that each column (hopefully!) contains homologous positions (ones that have evolved from a common position in the ancestor)

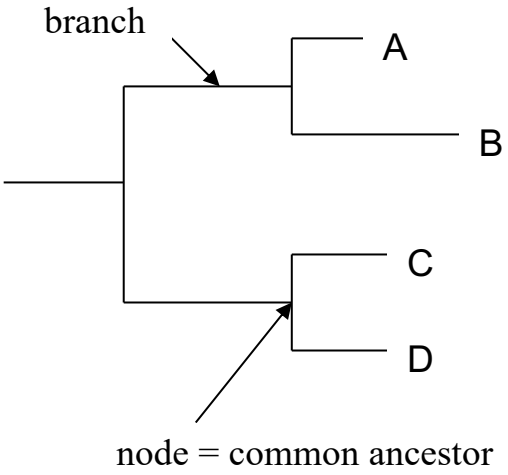
12. Set output format as **ClustalW**
13. **Press Submit**
14. Wait for your job to finish. Now look at the alignment.
    Q3. How many positions are identical?

→ * = identical residues, - in the alignment is a gap

Building a tree

1. Stay on the same page as the alignment, scroll up to the top
2. Press the tab that says phylogenetic tree – select branch length **Cladogram**

→ When looking at evolutionary trees the following tips may help:
   1. Every node on the tree represents a hypothetical ancestor.
   2. Relatedness is to do with common ancestry e.g. C is more closely related to D than it is to either A or B. This is because C shares a common ancestor with D more recently than it does with either A or B.



Look at your tree and look at the positions of the dentist, control and patient sequences. Was the dentist responsible for giving HIV to all, some or none of his patients? Complete the table below.

Patient	Likely source of infection?	Reasoning
A		
B		
C		
D		
E		
F		
G		
H		
I		