**Bioinformatics Practical 1 Sequence Manipulation and basic alignments** (the sequences you need for the practical are pasted at the end of this worksheet)

This is intended as an introduction or refresher depending on how much bioinformatics you have done in the past. We will pause several times during the practical for me to show you, answer the questions in the sheet and answer any queries you have.

Find sequence 1 (it is the first sequence at the end of this document)

In the first part of this practical you will use bioinformatics-based tools to obtain information on the sequence. There are many programs available on the internet which perform the same or similar types of sequence manipulation and analysis. For most of the procedures you have been given the web address of the site which is most straightforward for that particular type of analysis (but there are MANY others, all doing the same thing!).

In the second part of the practical you will use the second set of sequences to construct a multiple alignment and a phylogeny.

**Generating an ORF (Open Reading Frame) Map**

The sequence you have been provided with encodes a protein and thus the first task is to determine the amino acid sequence of the protein. Remember that a DNA sequence has six frames (three in the forward direction and three in the reverse direction). ORF finding software scans each frame of a DNA sequence for start codons and stop codon and displays all potential ORFs above some minimum threshold length. It does not necessary indicate which of the ORFs encodes a protein. In this practical though the longest ORF encodes the protein.

Go to: https://www.ncbi.nlm.nih.gov/orffinder/  
Copy and paste your DNA sequence into the box provided. Hit 'Submit'. The ORF maps

will be displayed graphically. Use the standard genetic code and other default parameters.

Question 1: Which frame contains the longest ORF?

* Frame 3: ORF11 (contains 2022/673)

Question 2: Where is the location of the ORF within the DNA sequence?

* Start from 480 to 2501

**Translating the Sequence**

Using the above information, you can now generate the amino acid (氨基酸) sequence of the protein. Translation programs often ask you what frame is to be translated and which genetic code it should use.

Go to: http://bio.lundberg.gu.se/edu/translat.html  
Translate the appropriate region identified from the ORF map above by pasting in the DNA sequence 1 into the box and using the standard genetic code. This site is a rather old translation tool but it contains the most important basic information.

Create a file on the desktop (in Word is fine) and save the protein sequence (there is no need to remove the numbers since most bioinformatics tools will recognise the sequence as it is and strip the numbers out.

Try out a couple of the other codes (e.g. mammalian mitochondria, yeast mitochondria) to see what impact they have on the translation.

Question 3:Give two examples of amino acid residues that are different between the standard code and the code you have selected. Hint**:** what does an \* represent in a translated protein sequence? Not all codes introduce or lose stop-codons though (see the first ten amino acids of the yeast mitochondria verses the standard code)

* \* refers to **the number of stop codons**. The standard code and non-standard code are different, giving a new meaning to any one codon can lead to confusion
* Pick a range because the only 1% of human DNA is coding with protein

See this link for a few examples, although there are quite a few other codes known that are not shown here: https://www.web-books.com/MoBio/Free/Ch3E1.htm

**Searching Sequence Databanks**

Having identified the protein sequence, you can now try to identify it. Search programs and databases were discussed in the lecture, but a brief description follows. Databases can be searched with both proteins and DNA. The programs you will use are the BLAST series:

blastn searches uses a DNA sequence to search a DNA database  
blastp uses a protein sequence to search a protein database.  
tblastn uses a protein sequence to search a DNA database.  
Blastx uses a DNA sequence to search a protein database – the DNA is translated in all possible ORFs and compared to proteins in the database.

TblastN uses protein to search a translated DNA database  
tBLASTX search translated nucleotide databases using a translated nucleotide

Question 4:Why would you want to do a tblastn search?

There are also various databases such as the ‘nr’ (non-redundant) database that contains finished DNA or protein sequences, swissprot that contains protein sequences and the EST databanks that contain mRNA sequences. Go to:

Go to: http://www.ncbi.nlm.nih.gov/blast/ Click on ‘protein blast' (blastp)

Paste in the protein sequence (don’t worry about the numbers) you have just created into the search box and run a blastp search against the swissprot database (scroll on the ‘database’ box). Click on the ‘show results in a new window box’

ALSO click on the algorithm parameters box and change ‘expect’ from 0.05 to 10

A new screen will appear – after a little while the ‘hits’ will appear. If you are ‘in-person live, do NOT keep hitting blast or refresh as it will just take longer (and NCBI might think we are a ‘bot’ and ban us-this HAS happened in a previous year and I had to write a grovelling apology, stating that I was an idiot for not controlling my class properly and that I would not do it again-it still took a couple of days for NCBI to lift the ban from most of SAF though ☺)

You can have a look at the graphic summary and the alignments as well if you like as it will help to answer Q8

Question 5: What are the scores and e (expect)-values of the closest and fourth closest matches?

Question 6: Scroll down the screen to see how similar the input sequence is to (i) the closest match (ii) to the 4th closest match. Click on the 4th closest match and it will show you the alignment. Write down the percentage similarities of these matches across the aligned region. What do the + symbols between the sequences represent?

Question 7: Links to further information regarding the retrieved sequences are found on the right by clicking the ‘Accession link. Use these links to obtain further information from which the top match was derived. e.g. how long is the sequence in the databank?

Question 8: In question 5 above you wrote down the e-value of two of the matching sequences. What does the e (expect) value represent? Using this knowledge and the descriptions of the matching sequences what do you think is the first ‘random’ (i.e. non-homologous) sequence in the ‘Sequences producing significant alignments’ list? THINK about this and then justify your answer. (Knowing when a match is significant and when it is not is often important).

If you want, you can spend some time looking at the options

**Constructing sequence alignments**

You are now going to construct an alignment of the second (set of) sequences (shown at the end of the document). The taxa are listed below:

EIAV-an exogenous lentivirus infecting horses

FIV-an exogenous lentivirus infecting cats

HIV-an exogenous lentivirus infecting humans

MMTV-an exogenous and endogenous betaretrovirus in mice

SRV1-an exogenous betaretrovirus infecting monkeys

BLV- an exogenous deltaretrovirus infecting cattle

RSV-an alpharetrovirus infecting birds.

Note clustal can be very slow! Do not resubmit your sequences. If it is too slow, you can try to run the alignments on Wednesday and we can discuss on Thursday.

Sequence alignment programs use a ‘matrix’ or look up table to assign values between identical or similar residues or nucleotides. Because different sequences often have deletions or insertions when compared to one another these programs are able to inset gaps into one or more of the sequences in the alignment. The following web site looks quite complex, but all alignment programs tend to be like this.

<https://www.ebi.ac.uk/Tools/msa/>

We will use ‘clustal omega. Copy and paste the sequences at the end of this practical schedule into the box provided and add the sequence you found from your blast search. Set the alignment to DNA.

If you get an error message it means your input format is incorrect. Modify the format and try again.

Now repeat the alignment but this time set the alignment to protein. We are going to fool the program into giving us a bad alignment.

If you get an error message it means your input format is incorrect. Modify the format and try again.

Once it is working you will get the results on a new screen after a few minutes. You need to save 2 parts of the file (i) download the alignment file and (ii) copy the Tree data (newick file) from the phylogenetic tree box’ (it is the part with lots of brackets in it).

**Visualising your phylogenetic tree**

Now go to:

<http://iubioarchive.bio.net/treeapp/treeprint-form.html>

Upload the tree file, click on ‘tree diag’ and press submit

MMTV and SRV should cluster together, as should HIV, FIV and EIAV

Question 7: What is the difference between a good alignment and a bad alignment. Did the difference impact your tree?

**Sequence 1** (note this sequence is in courier font-try changing the font and see how ‘messy’ the sequence now looks-moral-always use a courier font to display DNA or protein sequences!!)

GCACGTTGAGGGACGCATAAAAGAGCAGATCAACCGCGCCGGAGAAAAA

ATAGCCGCGGCTGAAGTGGAATCGGCACTGCTGCGTTTAGCGGAAGTGCA

GGATCCCGCGGTGGTCGCCGCGCCGGACACGCTGCTTGGCGAGCGGATTT

GCGCGTTTATCATCGCGCAGCAGGTGCCAACTGACTATCAGCAGTTGCGT

CAACAACTGACCCGTATGGGGCTCAGCGCGTGGAAAATTCCTGACCAAAT

CGAGTTTCTGGACCACTGGCCGCTCACCGCCGTCGGCAAGATAGACAAAA

AACGCCTGACGGCTCTCGCCGTCGACCGTTATCGCCATTCTGCCCAATAA

GCGCAAACCGACCCGAAACAGGTTGAAATAAACCCGTTTCGGGTAGCACC

ACTATTAGAAATAGTTATCATTTTCAATTCACCATTGTCGGTATTTTTGG

CGTTTCGCCGTCTTACAGGGACTCACAACAATGAAAATGACACGGCTTTA

TCCTCTGGCCTTGGGGGGATTATTGCTCCCCGCCATTGCTAATGCCCAGA

CTTCACAGCAAGACGAAAGCACGCTGGTGGTTACCGCCAGTAAACAATCT

TCCCGCTCGGCATCAGCCAACAACGTCTCGTCTACTGTTGTCAGCGCGCC

GGAATTAAGCGACGCCGGCGTCACCGCCAGCGACAAACTCCCCAGAGTCT

TGCCCGGGCTCAATATTGAAAATAGCGGCAACATGCTTTTTTCGACGATC

TCGCTACGCGGCGTCTCTTCAGCGCAGGACTTCTATAACCCCGCCGTCAC

CCTGTATGTCGATGGCGTCCCTCAGCTTTCCACCAACACCATCCAGGCGC

TTACCGATGTGCAAAGCGTGGAGTTGCTGCGAGGCCCACAGGGAACGTTA

TATGGCAAAAGCGCTCAGGGCGGGATCATCAACATCGTCACCCAGCAGCC

GGACAGCACGCCGCGCGGCTATATTGAAGGCGGCGTCAGTAGCCGCGACA

GTTATCGAAGTAAGTTCAACCTGAGCGGCCCCATTCAGGATGGCCTGCTG

TACGGCAGCGTCACCCTGTTACGCCAGGTTGATGACGGCGACATGATTAA

CCCCGCGACGGGAAGCGATGACTTAGGCGGCACCCGCGCCAGCATAGGGA

ATGTGAAACTGCGTCTGGCGCCGGACGATCAGCCCTGGGAAATGGGCTTT

GCCGCCTCACGCGAATGTACCCGCGCCACCCAGGACGCCTATGTGGGATG

GAATGATATTAAGGGCCGTAAGCTGTCGATCAGCGATGGTTCACCAGACC

CGTACATGCGGCGCTGCACTGACAGCCAGACCCTGAGTGGGAAATACACC

ACCGATGACTGGGTTTTCAACCTGATCAGCGCCTGGCAGCAGCAGCATTA

TTCGCGCACCTTCCCTTCCGGTTCGTTAATCGTCAATATGCCTCAGCGCT

GGAATCAGGATGTGCAGGAGCTGCGCGCCGCAACCCTGGGCGATGCGCGT

ACCGTTGATATGGTGTTTGGGCTGTACCGGCAGAACACCCGCGAGAAGTT

AAATTCAGCCTACGACATGCCGACAATGCCTTATTTAAGCAGTACCGGCT

ATACCACCGCTGAAACGCTGGCCGCATACAGTGACCTGACCTGGCATTTA

ACCGATCGTTTTGATATCGGCGGCGGCGTGCGCTTCTCGCATGATAAATC

CAGTACACAATATCACGGCAGCATGCTCGGCAACCCGTTTGGCGACCAGG

GTAAGAGCAATGACGATCAGGTGCTCGGGCAGCTATCCGCAGGCTATATG

CTGACCGATGACTGGAGAGTGTATACCCGTGTAGCCCAGGGATATAAACC

TTCCGGGTACAACATCGTGCCTACTGCGGGTCTTGATGCCAAACCGTTCG

TCGCCGAGAAATCCATCAACTATGAACTTGGCACCCGCTACGAAACCGCT

GACGTCACGCTGCAAGCCGCGACGTTTTATACCCACACCAAAGACATGCA

GCTTTACTCTGGCCCGGTCAGGATGCAGACATTAAGCAATGCGGGTAAAG

CCGACGCCACCGGCGTTGAGCTTGAAGCGAAGTGGCGGTTTGCGCCAGGC

TGGTCATGGGATATCAATGGCAACGTGATCCGTTCCGAATTCACCAATGA

CAGTGAGTTGTATCACGGTAACCGGGTGCCGTTCGTACCACGTTATGGCG

CGGGAAGCAGCGTGAACGGCGTGATTGATACGCGCTATGGCGCACTGATG

CCCCGACTGGCGGTTAATCTGGTCGGGCCGCATTATTTCGATGGCGACAA

CCAGTTGCGGCAAGGCACCTATGCCACCCTGGACAGCAGCCTGGGCTGGC

AGGCGACTGAACGGATGAACATTTCCGTCTATGTCGATAACCTGTTCGAC

CGTCGTTACCGTACCTATGGCTACATGAACGGCAGCAGCGCCGTCGCGCA

GGTCAATATGGGTCGCACCGTCGGTATCAATACGCGAATTGATTTCTTCT

GATTATTGTAAAAGGGATACCGAAAAGGTATCCCTTTTACACCACTAGTT

AAAACCAGTAACTCAGCAGAGTCGCAAAAAATATTAATCCATAGTGATTA

TTTAAACAATGAAATTGCGATTAGGACAAATAGATTTAACTTTCTCGTTC

CTTTCTCTCCTTATACTAAAGAAATAATCATATCAAAATAAAAATTCACA

ACAGTGCAACATTAAAAATACAACCAACAAACAATCCTATATACAAGGCA

CATCTCCAGAATATAAAAGCACAGACAAACAACCTAAAAAAAACAACCCG

AATTAATAAAACCTTTACAATTACACACCCTCAACTCAAAACAATTTCGA

AAACTCAAAGATTTCATCGGCAAAAACAGTCATTAACACATCTATTTTTT

TGAAATTCTGTAGAGTAAAACTGATATAAAGCATTTATCATATTAACATA

TCAATAAGTGCAAACTTAAAAATCAAAAGTTAGGGTTCAGTAAAACCAAT

TCGCCACAAAAAAAACCACCCCATACATATAAATTATTTTATAGGTAAAA

TAGATTATATATTCTCATAGCTACCACAAACTATACTAGCCTGAACTATA

TTTATTCTGCTGCAATCAATGCATACATAACACAAATATCACTCAGGTAC

ACTACTCAAACCACGCTGGGATTTTTCCTCAAGTTATAATCATCCCCCCG

AAAATCATTCGGCATTTACTCATTAAATAGTCACCCCATAGGCCTGTACA

TGTTCACTCAGAAATATACATCCTTTTCTCTGTCATAAACCCTCTGATTA

ATCATAAATAAATACTTGTGACACCAATCTTTTTCCTTAACGGAACGAAT

TGTTGTGTAGAAGGAGATAATATTATGTCAAAAAAATATCAGCCATTGCT

TATAACTCATTATATGTCAACATGGGTCACTATAACGGAAGCAGTTGAGA

TCACTACCAAAGCCATAAAACAAAAAATTACTCCTAGCGATATTTATCGT

CATGCCTTGAGTGGCAACATCCTACTATCGGTCTATTTCCAGTCTCCTGT

GATACTTAAGAAGATACAAACTTTTAATGGATCCATAAAATTCAGGCAGT

TTGAGGGAGACCTACTTGATAAACTATGCATGCTTGAC

**Sequence set for the multiple alignment**

>BLV

CAATGGCCCTTTAAACTAGAACGCCTCCAGGCCCTTCAAGACCTGGTCCATCGCTCTCTGGAGGCAGGCTATATCTCCCCCTGGGACGGGCCAGGCAATAATCCAGTATTCCCGGTACGGAAACCAAATGGCACCTGGAGGTTTGTGCATGATCTACGAGCTACAAATGCTCTTACAAAGCCCATCCCGGCGCTCTCCCCCGGACCGCCAGACCTTACCGCTATCCCTACACACCTTCCACATATCATTTGCCTAGATCTCAAAGATGCCTTCTTCCAGATTCCAGTCGAAGACCGCTTCCGCTCCTATTTTGCTTTTACCCTCCCTACCCCCGGGGGACTCCAACCTCACAGACGCTTTGCCTGGCGGGTCCTACCTCAAGGCTTCATTAACAGCCCAGCTCTTTTCGAACGGGCACTACAGGAACCCCTTCGCCAAGTTTCCGCCGCCTTCTCCCAGTCTCTTCTGGTGTCCTATATGGACGAT

>EIAV

CAATGGCCACTTACTAAGGAGAAGTTATTAGGTGCTAAAGAAATAGTCAAAAAATTGTTGGATGAAGGTAAAATATCAGAAGCCAGTGATGATAATCCTTATAATTCTCCTATATTTGTAATAAAAAAGAAATCTGGGAAGTGGAGATTATTGCAAGATTTAAGAGAGTTAAATAAGGTGGTACAAGTAGGAACTGAAATATCCAGAGGGTTACCTCATCCAGGGGGATTAATTAAATGTAATCATATGACAGTATTAGATATTGGAGATGCATATTTCACTATACCATTAGATCCAAAGTTTAGACAATATACAGCATTTACTGTGCCATCCATTAATCATCAGGAACCAGATAAAAGATATGTATGGAATTGCTTGCCACAAGGTTTTGTGTTGAGTCCATACATATATCAAAAAACATTACAGGACATATTGCAAGCTTTTAGAGAAAGGCATCCAGATGTACAATTATATCAATATATGGATGAT

>FIV

CAATGGCCTCTATCGAGAGAAAAGATAGAAGCATTAACAGAAATAGTATATAGATTGGAAAAAGAAGGAAAAGTGAAAAGAGCAGATCCTAACAATCCTTGGAATACTCCAGTGTTTTGTATAAAGAAAAAATCAGGAAAATATAGGATGTTAATAGATTTTAGATGTTTGAATGATCTTACTGAAAAAGGAGCAGAAGTTCAATTAGGATTACCTCACCCTGCGGGATTAAAAGAAAGAAAACAAGTAGCAATACTAGACATTTCAGATGCTTATTTTACTATCCCATTAGATAAAGACTATCAGCCTTATACCGCCTTTACTTTACCAAAATTAAACAATCAAGGTCCTGGAGAAAGATTTGTTTGGTGTTCCCTTCCACAAGGATGGGTGTTGAGCCCGTTAATATATCAAAGTACCCTTAATAATATATTAAAACCTTTTAGAGAAAAACATCCTGAGATAGATCTATATCAATATATGGATGAT

>HIV1

CAATGGCCATTGACAGAAGAAAAAATAAAAGCATTAGTAGAAATTTGTACAGAGATGGAAAAGGAAGGGAAAATTTCAAAAATTGGGCCTGAAAATCCATACAATACTCCAGTATTTGCCATAAAGAAAAAAGACAGTACTAAATGGAGAAAATTAGTAGATTTCAGAGAACTTAATAAGAGAACTCAAGACTTCTGGGAAGTTCAATTAGGAATACCACATCCCGCAGGGTTAAAAAAGAAAAAATCAGTAACAGTACTGGATGTGGGTGATGCATATTTTTCAGTTCCCTTAGATGAAGACTTCAGGAAGTATACTGCATTTACCATACCTAGTATAAACAATGAGACACCAGGGATTAGATATCAGTACAATGTGCTTCCACAGGGATGGAAAGGATCACCAGCAATATTCCAAAGTAGCATGACAAAAATCTTAGAGCCTTTTAGAAAACAAAATCCAGACATAGTTATCTATCAATACATGGATG

AT

>SRV1

CAGTGGCCTTTACTCAATGATAAATTAAGTGCCGCCCAACAGTTACTGCAGGAACAACTGGAAGCAGGACATATTATAGAAAGTAATTCTCCTTGGAATACACCTATTTTTGTTATTAAAAAGAAGTCTGGTAAATGGAGACTCTTACAAGATTTAAGAGCAGTAAATATCACTATGGTCCTTATGGGTGCCTTACAACCAGGATTGCCTTCACCGGTTGCGATTCCTCAAAAATATTTTAAAATCATTATTGATCTTAAAGATTGCTTTTTTACAATTCCCCTTCACCCTGCTGACCAAAAAAGATTTGCCTTTAGTCTTCCATCTACAAATTTTAGACAACCAATGAAGCGCTATCAATGGAAAGTCTTACCTCAGGGTATGGCCAATAGTCCTACCTTGTGTCAAAAATATGTAGCTGCTGCTATAGAGCCAGTCAGAAAAACATGGACACAAATGTATATTATACATTATATGGATAAT

>MMTV

CAATGGCCCCTTAAACAAGAAAAGTTACAGGCTTTACAACAGTTAGTGACAGAACAATTACAACTGGGCCACTTAGAAGAGAGCAATAGCCCTTGGAATACGCCTGTTTTTGTCATTAAAAAGAAGTCAGGAAAATGGAGACTGTTACAAGACCTACGTGCAGTTAATGCTACAATGCACGATATGGGAGCATTACAACCCGGCTTGCCGTCCCCTGTAGCAGTCCCTAAAGGATGGGAAATAATCATAATAGATCTACAAGATTGCTTTTTTAATATAAAACTGCATCCTGAAGATTGTAAAAGATTTGCTTTTAGTGTGCCCTCCCCTAATTTTAAGAGACCTTATCAAAGATTCCAATGGAAAGTTTTGCCCCAGGGTATGAAAAATAGCCCTACTTTATGTCAAAAATTTGTGGACAAAGCTATATTGACTGTAAGGGATAAATACCAAGACTCATATATTGTGCATTACATGGATGAC

>RSV

CAGTGGCCCCTCCCTGAAGGTAAACTTGTAGCGCTAACGCAATTAGTGGAAAAAGAATTACAGTTAGGACATATAGTACCTTCACTTAGTTGTTGGAACACACCTGTCTTCGTGATCCGGAAGGCTTCCGGGTCTTACCGCTTACTGCATGATTTGCGCGCTGTTAACGCCAAGCTTGTTCCTTTTGGGGCCGTCCAACAGGGGGCGCCAGTTCTCTCCGCGCTCCCGCGTGGCTGGCCCCTGATGGTCTTAGACCTCAAGGATTGCTTCTTTTCTATCCCTCTTGCGGAACAAGATCGCGAAGCTTTTGCATTTACGCTCCCCTCTGTGAATAACCAGGCCCCCGCTCGAAGATTCCAATGGAAGGTCTTGCCCCAAGGGATGACCTGTTCTCCCACTATCTGTCAGTTGGTAGTGGGTCAGGTACTTGAGCCCTTGCGACTCAAGCACCCATCTCTGTGCATGTTGCATTATATGGATGAT