**Bioinformatics and Genomics Practicals-Practical 1**

We will doing some introductory DNA sequence manipulations and database searching/investigation in the three practicals. You WILL make mistakes in ‘button clicking/parameter setting’ (I still make them all the time….) but it is important that you try to understand when the results are not as expected and thus you have made an error somewhere. This takes time and practice. Spotting errors is just as important as knowing how to do the analyses correctly. ‘Rubbish in, rubbish out’ can be especially apparent in bioinformatics! I will be providing lots of examples of where I deliberately make errors to demonstrate this. I will also be discussing the answers to the questions during the practicals.

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

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!).

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

Question 1: Which frame contains the longest ORF?

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

**Translating the Sequence**

The next aspect of simple *in-silico* sequence manipulation I want you to try is DNA sequence translation (note: ORFfinder also does this but the graphics on the next site are clearer).

Go to: https://web.expasy.org/translate/  
Paste in the DNA sequence again into the box and translate in all six frames using the standard genetic code.

Note: Have a look at the translated sequences-what do *in-silico* protein sequences look like? i.e. what features do they have that non-coding regions of DNA lack. I will discuss what a correct protein sequence translation looks like (it may be more obvious by using the ‘verbose’output format).

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

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 (other translation programmes use \* instead)?

Not all codes introduce or lose stop-codons though (see the first ten amino acids of the yeast mitochondria verses the standard code)

This site has about 15 codes but there are many others that are not shown here (see link for a few more): <https://www.researchgate.net/figure/Standard-genetic-code-and-naturally-occurring-variants-The-standard-genetic-code-and-the_fig2_44671813>

Click on the first M (initiator) residue in your suspected protein sequence to recover the protein sequence.

**Searching Sequence Databanks**

Having identified the protein sequence, you can now try to identify it. Search programs and databases were discussed in the tutorial/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 ‘fasta’ protein sequence you have just created into the search box and run a blastp search against the swissprot database (scroll on the ‘database’ box).

ALSO click on the algorithm parameters box and change ‘expect’ from 0.05 to 20

Click on the ‘show results in a new window box’

Normally you don’t need to this precise-the nr database under default parameters would be fine, but I am trying to illustrate a few important concepts in this section.

A new screen will appear – after a little while the ‘hits’ will appear. 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)

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 (think of the accession number as the ‘home page’ of the sequence). 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

**Reverse complementing sequences**

Genomic DNA sequences can often encode proteins on the minus strand and, additionally, some viruses are negative sense (e.g. morbilliviruses). Bioinformatics analysis and alignments must always be performed with sequences in the plus sense. It is important to check that the sequences are in the correct sense, especially after downloading them. Otherwise, you might end up with a set of sequences (in a multiple alignment for example) with a mix of polarities.

Sequences can be reverse complemented very easily.

Go to:

<https://reverse-complement.com/>

Reverse complement sequence 2 at the end of the practical.

Question 9. How can you check and identify that the sequence has been reverse complemented successfully?

**Pairwise alignments**

It is sometimes important to determine the exact percentage identify between two sequences by running a specialised pair wise alignment. For example, you might want to know a pairwise distance between viruses in different hosts to support a hypothesis of a recent horizontal transmission event.

Go to:

<https://www.ebi.ac.uk/jdispatcher/psa>

Launch GGEARCH2SEQ. Click on DNA as the sequence type.

Now paste sequence 2 into the upper box and sequence 3 into the lower box.

Question 10. How similar are the sequences to each other? Is this a corrected or uncorrected distance?

To obtain a corrected distance we need to apply Jukes-Cantor or similar model.

Go to:

<http://www.insilicase.com/Web/JukesCantor.aspx>

Calculate the number of changes from the similarities and length given in the previous analysis and calculate.

Question 11. What is the corrected distance? Note that if two sequences are closely related the observed and expected/corrected distances will be very similar.

There are also some additional sites that may be of help in investigating your DNA or protein sequences. Spend a little time in investigating these sites and their outputs

Paste in DNA for:

<https://www.sciencebuddies.org/science-fair-projects/references/genomics-g-c-content-calculator>

<https://www.bioinformatics.org/sms2/codon_usage.html>

(for this one the first base must be the A of ATG start codon, so use sequence 3)

Paste in a protein sequence for:

https://web.expasy.org/protscale/

**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 or similar 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 2 (measles virus-partial)**

TTTTCTTTAATATATTGCAAATAATGCCTAACTACCTAGGGCAGGATTAGGGTTCCGGAGTTCAACCAATTAATCCTTAATCAGAGCGCTGTATCCGACTAACTTGTACCATTCTTTGGTCTCCTTGACTGTTACCTTAAAAACCCACTCACGTTTTAAACCCCCCGTCATAATAATCTGTTTCTCTGACTTGGATAGATTCTTAACGAAGATATTCTGGTGTAAGTCTAGTACTAGATAACCGGACTTGAGATTCTGGATAAACCGATTTATCAACTTTCTGTTCCCGGAGTAAAGAAGAATATGCCCCCAAAATTTGCGAGTGATCCTAGATACAAGTTCTCGTTGCCTACTACTTACCAATACGGGGTAAGCGTGGAACATCCCTTGTTGACTTCTTTGGTTGTCTTTGAATCTTGCCAACTCCCTGTAGAGGATGAGTATAGAGTTAAGCAATCCATCTTGCCCTGAGGCAACATCATGGTGGATTAATTCTTTGCACAGTTTAGGTCCGTTAATTGCCAACCCGCAACTGATCAGCACCTGCTCTATAGGTGTAAGTTTTTTCAGAATAGGGTTGATATCACCTCTACTAACTGCGCCTCCCACAATTGCTTGTATGCAGCTTAGTTGCTTAATGGATAGGATGTGACCTATAAGTCCAGGTGAAGTCCGCACAGATGATTCAATTATCTGCTGCTTGATCTTTTCAGGATTCATTAGCCGGTTAGCTTTGAGATCTGTCATAACTAAATAAGATTCAGTAGATATGAAGTTGCTGTACCTAGGGTAGACAAGGTTCACTTCTCTATAATGAGACCCTACATAGCTTATAAATCCCTGAACAAAATCCCCGCTGAAAGGCATAAGCTTAATCACCAGTATTGATCCTATTTTGCCAAGGAGTAGAGCCATCGATAAGATGGCTGCCAATTCCTCTAGCTTCTCTATAGTATCTTTGTTAGGTAAGGTCTCTATATCTGAATGGATAAACCCCACACTAGAGGTAGGGATATTACTGACTATGAAATTGAAGCAATCTATACTGCCTACCCACGTGACTTCGGGCCTCCCGTTAAAGAGCACCTTGACAATATTACCTACTCCCATTCTGTGTTCGACAAGGCCAACTTCGGAGGGATAGGGTGCTAATTCCCTTTGACCAGATCTAGAATTGGCGGAAACCCCACTATTATAGAAGCACTTGTTTAGTTTTAGTATCTCCTTATAAGTGATCAACATAGAACCCGACCCCTCACCCAAGAACAAGCCGTCTTCCCCTGGCTCAAGGCATCTCCTAATTAATGTTGATATCTCAACAGCTTTGTAGCAAGCAGATGAGTTTAACCCGATTCTGCGGAAAGCATGGATTTCATAATTGGCGAGACTACCCCCTGAAATGGGAAGATTGTGCTTGCTTGTGTTGATATCTTTGAGCAATTTTGCAACATCATCGTGTGGAGGTCTGAAATCCTTGATGCTCATATTTGAGATGTTGTTGCTGCCGACCTTTGGCTGACTGACATTTACCTCAGCGAGGGCGTCAAAAATGAATCCTGGATCAACTCTCAATCTTATCTGTTTGATAGATCCTCGACGGAGATAAGTCAGAGAGCATGAGTAATGGTCTACAATAATTGGATTTATGTTCCACGAAGATCCTGCTGGAGATAACCTAGCCTCTGCCTTGATATGATCGGTTAGAACTGCACATTTCTCTACCGGCCTTAGACCTCGAATCGGTGGGCAGGTCCCTGGTTGACAGTACAAATCTGCCAGAACACACAAGTGTTTTGCCTGGATGTTGTCGAATCTGTCCGGTACTACATCCTCATCGCTTTCACACAAAAGAAATGTGAACTCTTCTAACTCTTCATTCAACAACAGGTCGAGGTAGGTCATATAGCATGTGTAAACCATGTTGCACACAGTTGTGTGCAAGTTTTGAGCATCAAGTGAAGGACCATGGATAGGCTCTATAATACCACAATGCCAGAATTTCTTGTAGATCTTTGGGTGGCTTAGAGCATTGACAAGCACCTTAAACACTCCTTTGCTCATTCTAGAAAGGAACGAAGACAACAGCTCACCCATCTGATATTTCCCTGATGGTCTATGATAATGTACATCAAATGCCCAATTGATGGCTGCACACTGGCCCAAGTAGATGGTGAATAATCTTGGCTCTATAAGCAGAAACTCAGTTATGAAACTATTGATATCGTCATCCCCTATGAGAGCTGAAATTTCATTCATATGGTCCTTCTCAAATTTTGTTACCAGGTCAATCATAGATAGTGCTGTGGACTTAGCTAGAATGTGATATAGTTGGGGTGTGGACCATGTAACAAATTCCACAAGGTGCCTCCTATGGCTCTGGGTGTATAGCCTTGTTGCATCTCTGTCAATTAAAGGTGCATTATCATATATCAATGGGTTGGTACATAGCTCTGCCCTCAGCTCTAGCTTGCGGGAGCTGGGTATCCTGGGATGATCTATCATCGGGATCACGCAACAATCTGTTTCGACGTGAAGATGTAATACCGTGTTAGATGATCCAGTATCTTTCTCGAGTCGAAACAATGTTTCTAAAACACCCAACCCTAGAAGCATTCCTTGTTGGTATATAAAGTTAGTATCAACCTTCTTATCTGATATGACAAATGAGAGATTGTCGTTGGAGATTGTGGTATACCTTGCCACTCGGACAAGGGATGTACCTGAGTATTTCACTTGAGTGCTACGATCCCTCAACCTATGCGCTAAATTAGTCGAAGTCGAGATGGGAGTGATCACCCTTAGCTCCTCCAGGCTCACATTGGCCCTTTGCCTTGCCAACAACCAGGCTTCGTTCCAAGAGCTATCATCATCACCGTAAGCCCATGAGTACACTGTTGCTATTCTAACGGCAGATCGCAAGGATCTACTTGGGGCTCTTACGAAGGCGAGCTTCATGTCTGTTCTCTCATCAGTGGTAGAACCAATATATGGGACTCTCAAGGATGATGTTTCCTTGTCAATATCATCCAGTTGGCAACCCGAGGGGACAAAAAACCATCCGTAGTTGACTGATCCACACTCGCAGATGACACATGTCTCATGACGCCGAATAAGGTGGCCTCGCATAGATTCTAGTACATCAGGGACCTCAAGGCCGTAAATAGGCCGTCCTCGAGCTAGTCTTGCCCACATATGGCTTCTTAGGGCTCTAGCCAGCTGCACTGAACATGACTCTTTGTCAATGAGGACATTTCTCTTTCTTCCTGTCAATAGCACCATCCCTGCTCTAAATTGTTCATAGTCATAATTGGACAATCTGGTTATCACTCGAGAGGTTAACCCCCCCTTCCTCATGCTGGCTCGAATCAGGCCTTTTGTGGTATCTAGCATGCCTGCAATAGACTCTCTTGCCCCTGTGACACTATGATCCAGGATTTCATGAGCTGCCCTAGGTACTATAATATGCCTGTCCATGAGGAATGCCGCCAGTCTCTCGTCCTCTTCTTTACTGTCATCATGGAATAACCCTTTTAACATTGGGTTTGGACTATGGATTAGGACAAACCTTGCAGTTATGTTCTTGAGGAGTCTAGTGATGCTCTGGACGCATACAAGATTTGCTGAGTAAGGGTCGCTAGCCCAGTCTAGGAATGAAGAGTCCCCCGGTTGTTGTGTCATTACTTGATGGAGGGTCTCTTCAGGCATTAGTGATGCGAGAATCATTCTCTTGAGATCAGCAATTGATGATGTTACTGGATCACCGATGTTTCTGACAAACAGCCTGCTCATATTCAGATAATTCATCCCCCCAATAGGAGCGGGCAACAGTGCCATCCTTATTAAGAGATCGTTGTTTGTGAGGAGGGGTATGACTACATCTCGGGTCATGGTTGAATTGATTGTGAAGCCAAGAGAGATCAAAATTTGCTGTATCACTTTTAGGACGTTCAGGGAATATGCAAGATAACGGTCATAACCTCTCTCGATGCTTTTAGCCATTGTTGTAGCAATATTACTGCATGCTGCCCTTGTTTCATCAACTATAGTCTCTGACCAGAATACACATCTTGCGATGCTCTTGAGTGATTGGGACACAAGTAGCCCATCATAATATATTCCTTTTGAATAGACAAAAAAATGTGATGAAACAATTGTCTCATTTGCCTTGAGGTGATGGCCAATGTCATGTAGCCTTTGCCTAAGAATTACAAAGTAATCTCTAGTTACTCTAGCAGCTTCCCGTTTCTTAAGGTTGTAAGGCCATGTGCTGGGTACCCTTTTTGTTACGGCTATGGTCTGATTGTCCCCTTGCACTAACGAAGCAATCCTTACCCCGCTCTCATAAGCAGCCAGGTATAAGTAGGGAATGGTGCTGATGGTCCACAGCTTCTGACAATACCCTTCTATACCTCCCATAGGGTACTTGATGAAGATTTGGTCATTGGGGACTTTGCATAACGGGACATGGGCGTCAAGGTCGGGGGGGCAATGAGGGTCACTTACATAGAGGACAGAGGTTTCAAGCCTCTTATGCAGCCACTGAAAAAATGAGGGTAATCCGTAAATCTCATTTAGCCTCTGTGCAAATAAGCTGATGGTCTCATATCTCCAATTAAGGCAGTACTTCTTGAGATCAGTCGTGATAAATGCGCTGACTGTCTCGTAGGTTTCTATATTCTCCGGATGATCAGTGTCTTGATTCTGCCGAATTACATGAGGGAATCCTACAAACCCTTTTTCTGCTTTAACGTTCCTGGTACTTGTGTGGACTGGGCTTCGGGAGTAGGTTTTTAAGACTGGCCCCCCCCTGTGACTTTCTTTGAGATCTTTGGGGACTCCTGAGACAGCCAGAGTGTGGAGTGCCTTAGTCAAATCGTGCTCATCCTTGGCCATCCCATTGTCCTTAAAATACTTGCCAATCCCGTTTGAGATTAGATTTTCAGCGATCACTTGGCATGCCCTCATTTTGTAAGTCATTTTAGCGAAAAGTCTACCTGTTTCCTTGATCTCCTTTTCTTTCAGGCTGTAAGACAGGTTGAACTCAGGGTCATGGAGGTAGGCTCCACTTACGACATACATTATCATATCATATGGGTCAAAGCTCGAATCATTAAGGAAAACATCTACAAGCCTCCGTGACCCGGTTCCCTTGGGAGGATCGTAACGCAGGAACTCTTTCGGGTAAACTGAATCCCATTCCCTTTGGAGAGCAGCAAGTGCCTTGTCCTTTAGGTACATTGTCAGATCACTGTCCAGGCTAAGAGGCATAAAACAGCCAAATCTCACTCCAGCAAATGATTTCCAGTTATCAACGCACTGCTCATGTGTTAACCCTTCACCTGAAGCTTGAGCATTCCGGATTGTGTCTGCAGCATGCAGGGGGAGGGTCAGGGGTGGCCAACTGCCTCCGTGCCTGTCACGATAGCCGTTGATTATGATTCCACAAAATATGGCATGACCTTTCATCAGAGTCTCATACACAATGACTTTAGGCTGATTCATGTATTTCCTGACATTTTCAGCAGCCGTTACTGCTTCAAGTCTGGGGTGGCCGAAACTTCTGAAAAATGAGAAAATCTCCCCTGTCAGATGTATGTCATCAGTTATGAAAATGTAATCTAGGGCTTCAATTAACTCATGATAAGTACCTTCATCAGAAAACCCGTTTTGGTCAAGAACATCATGTATTTCAGTAAAGCAGTGGTTAAGGAAAGCACCTCTGAGTTCTACTGTTATATCCCTCAGTTGCAGGTAAGCAAGTGAAAGTGGCTCCAGCATGGCTACAATTTGATAAGTTGGATTCCCGAGTGCAGGGAAGAAACCATCTATCAGTTTCCACATGTATCTGACTCTTCCTAGAAGTTCTGCATACCTAGCATCAATGGTCATAGCGGTCTCTGTCATTAACCTCCCCTCTATGACATCACAATACATCAAAACCAGTTCAAACGTCAGGTAATATACATGTTGAGACTCCTTACTGATTATAGCAACAAGGTCACGAGAGATTAACAGCTCAACTGAACTACCAGTGAAGAATACAGGTGTGTGTCTCCTCCTATGGCAAGTATGGGTTTGTGATTTAATCACTGACCTCATCTCAGTCTTGACTGTAAACCAAAACAGAAAGGGCTCAAACCATTGGGAGCTGTGCATGTAAACTCCCAAGTTAATAATTTTCTCCTTGATGTCCTCCCTCAATTCGGAGCCTAGGCCAAGCCGTGAGTTAGTGTCCCTCAGGCATTGGAAAACCTTATCACTGACTTTGGAGTACAGCGAATTTCCCTTTTTTAGGAGCTCACGGATCTTCCTTGTTGACTCTTTGTCTTCTATGTTAAATAAATCCTGATTACAATTTGGATATGGAATATGAGAGTGGGCCGGATAACTCCTAAGCTTGGACTTGATGACATTCCCAACTTCCACATTGTTTATAATCATTTGGTTGGAGAATCCGTTTTTTAGGCGGTGCTTGATGTTCTGACACAGTGTAGGGTCCTCCAGGCTGTAAGCGTGAGGGACTCGAGCATACTCCAGGATAGCTACTATCTTATTGGTAACTATCGGGCTATCTAGGTGAACTTCAGGGTATAAGATCTGGTTGACAGATAGCGAGTCCAT

**Sequence 3 (canine distemper virus-partial)**

ATGGACTCTGTATCAGTGAACCAGATTCTATACCCTGAGGTCCATCTAGATAGCCCAATTGTAACCAATA

AGCTAGTATCTATTTTAGAATACGCACGAATTAGACATAACTATCAGCTCCTTGATACAACATTAGTGCG

TAATATCAAAGAGAGAATTTCAGAAGGGTTCTCAAACCAGATGATCATTAACTGCATCGAAATTGGGAGC

ATTATTAATCAGACCTTGTTATCTTATCCCAAACACAACCATGTGATATACCCAAATTGCAACAAACTTC

TATTTCATGCACAGGATCGAGTCATCTCTCTGAGGCTGAGAAATATATTCAAAAGAGGAAATAGCATCTA

TAGCAAAATAACAGACGGGGTCAAAAAATGCTTAAACGATATTAATCTTAATATTGGTTTAGGGGGTGCA

CTGGACAAGACTATTGGGACCAAAATTGATGAAGCAGGCATAATTATGCAAAGCTCACAGTGGTTCGAAC

CTTTCCTTCTATGGTTTACAATTAAAACAGAAATGAGATCAGTGATTAAATCCTCTACTCACAACTGTCG

CAAGCGGAGGCAGAACCCTGTCTTTGTAAAAGGTGAATCATTGAATGTGTTAGTCTCTAGGGACCTTGTA

TGTATTATTGACCTCACCAGTCACATTGTTTATTACCTAACATTTGAAATGGTCCTGATGTACTGTGATG

TAATAGAAGGGAGGCTAATGACTGATACTGCTATGGCAATTGATCAACGTTACTCAACTTTGCATGTCAG

GATCAGGTATCTCTGGGATCTAATTGACGGATTCTTCCCGGATCTGGGAAATTCAACCTATCAATTGGTG

GCTCTACTGGAGCCTCTCTCATTGGCTTACTTGCAATTAAAAGACATCACCTTCTCTCTCAGGGGTGCTT

TTCTGAGTCACTGCTTTGCTGAAATTCAGGAGATTTTACAGGACAATGGCTTCTATACTGAAGAGACGTT

CCAAACTTTAACCCAAGCTCTAGACTTCGTTTTCATCACAGAGGATATACATATAACAGGAGAAATCTTT

TCCTTCTTTAGAAGTTTCGGTCACCCAAGGTTAGAAGCAATAACAGCAGCAGAGAACGTACGGAAACACA

TGAATCAACCCAAAGTTGTCTCCTATGAGACTATGATGAAGGGACACGCTATATTCTGTGGGATAATCAT

TAACGGTTATCGGGATAGACATGGGGGGACTTGGCCTCCGATGGATCTTCCTGTTCATGCATCTCCTATC

ATCAGAAATGCTCATGCCTCAGGGGAGGGAATCACCTATAGTCAATGTATAGAAAATTGGAAATCCTTCG

CAGGAATTCGATTTAAATGCTTTATGCCTCTTAGCCTAGACAGTGATTTGACCATGTACCTGAAAGATAA

GGCTTTGGCAGCCCTAAGAAAAGAGTGGGACTCAGTGTACCCAAAAGAATTCCTCAGGTACAATCCGCCT

CGCTCCACTGAGTCTCGGAGACTTGTTAATGTGTTTCTAGAGGACTCTCAGTTTGACCCTTATAACATGA

TTATGTACGTTATCTCAGGTCAATATCTAGAAGATCCTGATTTCAACCTATCATACAGTCTCAAAGAGAA

AGAGATTAAAGAGGTAGGGAGATTATTCGCTAAAATGACCTACAAAATGCGAGCCTGTCAGGTCATAGCA

GAAAACTTGATATCTAATGGAATTGGGAAGTACTTCAAGGACAATGGGATGGCAAAGGATGAACACGATC

TCACTAAATCATTGCACACTCTGGCTGTGTCCGGGGTTCCGAAAGACAAGAAAGATTCCCATCGTGGCCT

CACTAACCAGCGTAAATCCCTGAAGCCTGCACCTTATCGAGGAACCCGTCACTCCGTCTCTTCCCCAAGT

AGTAGATATATAGACCCAAACCCAAATTTTTGCACCAGTAGAAGAGAAGACAATGACATAGAGATCTATG

AAACTGTAAGTGCATTTATAACTACAGATCTCAAAAAGTACTGTCTGAATTGGCGTTATGAGACCATCAG

TATTTTTGCTCAGAGATTAAATGAAATCTATGGTCTCCCCTCATTTTTCCAATGGTTGCACAGAAGATTG

GAACAGTCGATCTTATACGTAAGTGACCCCCACTGCCCTCCAGATCTCGATCGTCATGTGGACTTGAATA

CAGCCCCTAACTCTCAAATATTCATCAAATACCCAATGGGAGGGGTGGAGGGGTATTGTCAGAAGTTATG

GACTATTAGCACCATACCTTATCTGTACTTGGCGGCACATGAGAGTGGTGTCAGAATTGCATCACTTGTC

CAAGGTGATAACCAAACTATTGCTGTCACTAAAAGAGTACCAAGCACCTGGTCCTATGCCTTGAAGAAAT

CTGAAGCCAGTCGAGTGACCACAGAATACTTTATAGCCTTGAGACAGAGGTTACATGATGTCGGACATCA

TTTGAAAGCAAATGAAACAATAATCTCTTCCCACTTTTTTGTATACTCAAAAGGAATCTATTATGATGGA

ATGTTAATTTCACAATCCCTGAAGAGTATAGCTAGGTGTGTATTTTGGTCAGAAACAATAGTGGATGAGA

CCCGAGCCGCGTGCAGCAACATTTCAACAACATTAGCGAAAGCCATTGAGAAAGGGTTTGACCGATATTT

AGCCTACACGCTGAACATTTTAAAAATCATCCAACAAGTATTAATTTCATTAGGATTCACTATCAATTCA

GCGATGACACGGGATGTGATAGAACCCCTCTTACAAGATCACTGTCTCTTGACCAAGATGGCAATTCTCC

CCGCACCCATTGGCGGTTTTAATTACCTCAATATGAGTAGGCTCTTTGTCAGGAATATCGGGGATCCCGT

GACATCTTCTATTGCTGACCTCAAACGAATGATCCGATCAGGCCTTCTTGGAGTGGAGATTCTACATCAG

GTCATGACTCAATACCCAGGTGACTCTTCTTATTTAGATTGGGCAAGTGACCCTTATTCTGCCAATCTGC

CCTGTGTCCAGAGCATAACCCGACTCCTTAAAAATATCACAGCTAGGCATGTCCTTATCAACAGTCCAAA

TCCGATGCTGAGAGGATTGTTCCATGATGAAAGTCAGGATGAGGATGAAGCTTTAGCAGCTTTCTTGATG

GATAGGAAAATTATTATCCCAAGGGCTGCACATGAAATTCTGGATAACACAATCACGGGTGCAAGAGAGG

CAATTGCCGGAATGCTAGACACCACAAAGGGGTTGATAAGAGCAAGCATGAAAAGAGGAGGTCTAACCCC

TAGAATAATAACCCGTTTGTCAACTTACGATTATGAGCAATTTAGGGCAGGTATCAGACTGTTCTCAGGG

AAGGGGCATGATCAGCTCATCGATCAAGACTCATGTTCCGTCCAGTTAGCGAGAGCATTAAGGAACCACA

TGTGGGCCAAGCTGGCGAAGGGTCGTCCTATTTATGGTCTAGAAGTCCCGGATATCCTTGAATCAATGAA

GGGTTATATGATTAGAAGACATGAATCCTGTTTGCTTTGCGCATCAGGCTCTCATAACTATGGTTGGTTT

TTTATACCAGCGAATTGCCAATTGGATAGTATTACAGAGGGAACATCTGCACTGAGGGTGCCATACATAG

GGTCCACAACAGAAGAAAGAACAGACATGAAACTAGCATTCGTCAAATCTCCTAGTAGGTCTCTAAAATC

AGCAGTGAGAATAGCAACTGTGTACTCATGGGCCTATGGTGATGATGACGAATCTTGGCAAGAGGCTTGG

ACCTTGGCAAAACAGAGAGCGGACATCTCACTTGAGGAATTACGAATGATTACCCCAATTTCCACTTCTA

CTAATCTAGCTCACCGACTAAGAGACAAGAGTACTCAAGTCAAATACTCAGGGACCTCTCTCATCAGAGT

AGCACGTTATGCAACTATCTCGAATGATAATCTTTCTTTTATTATAGATGACAAGAAAGTGGACACAAAT

TTTATTTATCAACAAGGTATGCTCCTGGGCCTGGGGATCCTTGAGCACTTATTTAGATTGTCTTCAACCA

CCGGCGACTCTAACACCGTGTTACATTTACATGTTGAAACAGATTGTTGCGTAATACCCATGAGTGACCA

TCCAAGAGTCCCAGGGCTCAGAAAGGTCGTCATACCAAGAAATATTTGTACAAATCCTTTGATTTATGAC

AGTAACCCTATTATTGAGAAAGATGCAGTCAGACTTTATAACCAGAGTCACAGAAAGCACATTGTAGAGT

TTGTCACATGGACAACAGGGCAGCTTTATCATGTGCTAGCTAAATCTACTGCTATGTCTATGGTTGAGAT

GATTACAAAGTTTGAAAAGGACCACCTAAATGAAGTCACTGCGTTAATTGGCGATGATGATATCAATAGT

TTTATCACTGAGTTTCTTCTAGTTGAGCCTAGATTATTTACTGTATATCTAGGTCAATGTGCTGCAATCA

ACTGGGGCTTTGAAATTCATTATCACCGACCTTCTGGAAAGTACCAAATGGGTGAGTTGTTGTTCTCTTT

CCTGAGTAGAATGAGTAAAGGAGTCTTCAAAATTTTAGCCAATGCATTGAGCCATCCTAAAGTATATAGA

CGGTTTTGGGACAGTGGGATGATTGAACCTGTTCATGGACCCTCTCTTGACTCCCAAAACCTACATATAA

CTGTATGCAACCTGATCTATAACTGTTACATGATTTACCTAGACCTTCTGTTAAATGATGAATTAGATGA

TTTCTCATTCATTTTATGCGAAAGTGACGAGGATGTCATACCTGAAAGATTTGACAACATACAAGCCAGG

CACCTATGCATCTTATCTGACCTTTATTGTAACCCTCGTGATTGTCCCCAGATTCGTGGGTTGACACCAA

CACAGAAATGTGCTGTGTTGTCGGGGTACTTAAAATCAAAAGCCCTAGAATCCCATGTTGGTCTGACATG

GAATGACAAACCTATCTTAATAGATCAATATTCATGTTCCCTGACATATCTTAGAAGAGGCTCAATCAAG

CAGATAAGATTGAGAGTGGATCCCGGATTCATCACTGATGCTGTTGGATGCTTAGAAAGGCGTCCTCTAA

GAAATAATTCTACCTCTAAGGCCTCAGAATTAACGTCAGGATTTGACCCACCGAAAGATGACTTGGCTAA

ACTTCTGAGTCAGCTGTCAACAAGGACACATAACTTACCTATTACAGGATTAGGAGTCCGGAACTATGAG

GTTCATTCATTCAGAAGAATTGGGATCAACTCTACTGCATGTTACAAGGCAGTTGAAATAGCTTCCGTGA

TTAAGAACGAATTTACGTCTGAAGAACACGGATTATTCCTAGGAGAAGGTTCAGGTGCAATGTTGACAGT

ATATAAAGAGCTATTAAGATTGTCAAGATGTTATTATAACAGTGGTGTGTCGGTAGAATCCAGAACTGGA

CAACGAGAGATTTCACCTTACCCTTCTGAGGTCAGTCTGGTGGAACATCAATTAGGACTCGATAAATTGG

TGACTGTGCTTTTCAATGGGAGACCAGAAGTAACTTGGGTTGGGAGTGTTGATTGTTACAAGTACATACT

GAGCCAGATCTCTGCTAGCAGTCTTGGGTTGATTCACTCGGATATAGAGTCACTACCGGATAAAGACATA

ATTGAAAAGTTGGAAGAATTGTCTGCTATATTATCAATGACTTTGATATTAGGGAAGGTAGGGTCAGTGT

TAGTAATTAAGATCATGCCAGTTAGTGGCGACTGGGTTCAAGGATTTATTTTGTATGCACTCCCACATTT

TCTTCGAAGTTTCATAGTTTACCCAAGATACAGCAATTTTGTGTCAACAGAGGCCTACCTTGTTTTTACC

GGTCTTAGAGCAGGGAGACTAATCAATCCTGAGGGGATTAAACAACAGATTTTGCGAGTCGGTATTCGAA

CTTCACCCGGGTTGGTAGGGCACATCCTTTCATCAAAGCAGACAGCATGTGTGCAGTCTTTGCATGGACC

TCCATTTCATGCTAAATCTTTCAATCCTCACCTTCAGGGTTTAACAAGTATTGAGAAGGTATTAATCAAT

TGTGGGCTTACAATTAATGGTCTTAAAGTATGTAAGAACCTGCTTCACCATGATATTTCGTCAGGCGAGG

AAGGGCTGAAAGGATCTATCACGATCCTTTACCGGGAACTCGCAAGGTTCAAGGATAACCACCAATCTTC

ACATGGAATGTTCCATGCATACCCTGTGTTAATCGCAAGTCAGGAAAGGGAGCTCGTATCTATCATTGCA

AAGAAGTACTGTGGCTATATTTTGCTGTACTCGGGCGACTTATACGAAATTACCAGGATTGTCCGAAACC

TGAAAGCCAACCACATAATTTTCGACCTGCATCGTAATTTATTCATGGATAATCTGTCCAGATCTGACAG

GTCTCTCATCCTAACGACAATCCCCAAAAAGAATTGGCTCTTTCAGCTCGAGACAAAAGAGATAAAGGAG

TGGTTCAAATTATTAGGTTATAGTGCACTGATTAGAAATCACTGA