**Assignment 2 – EMBOSS, BLAST and Alignment (worth 10% of MB6301)**

**Course co-ordinator: Dr Marcus Claesson (**[**m.claesson@ucc.ie**](mailto:m.claesson@ucc.ie)**)**

**Demonstrator: Brian Forde**

**Student: James Henry Hehir (120224791)**

**For each question, please record your answer in a Word document, and also what you typed on the command-line to get this answer. When finished, upload the document (with your name in the title) with your answers to Canvas under Assignments.**

**Deadline: End of Sunday 21st of March**

**If you cannot meet this deadline you will have to fill in and submit the Late Submission Form and provide any supporting documentation (e.g. medical cert)**

**When no specific question is asked, perform the tasks and paste in the syntax including the full prompt, as well as the output that is not directed into a file, so that it is clear where you are located when doing the tasks.**

**To figure out how the command should be used, use ‘man’ or ‘info’.**

**Remember, an important part of Bioinformatics is the ability to find information on how to solve a problem online.**

**\*\*Include code that gives the required output. Comment code where necessary\*\***

**Q1. From the commandline download the genome, “Mycoplasma Genitalium” extract the reverse nucleotides between the coordinates 150 and 250 with the output in genbank entry format. Paste in the genbank output**

**Link to genome:**

<ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Mycoplasma_genitalium_G37_uid57707/NC_000908.fna>

jimmy@jimmy-VirtualBox[jimmy] wget "ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old\_refseq/Bacteria/Mycoplasma\_genitalium\_G37\_uid57707/NC\_000908.fna"

jimmy@jimmy-VirtualBox[jimmy] seqret -sreverse -sbegin 150 -send 250 "NC\_000908.fna" -osformat genbank

jimmy@jimmy-VirtualBox[jimmy] cat nc\_000908.genbank [ 7:37PM]

LOCUS NC\_000908.2 101 bp DNA linear UNC 19-MAR-2021

DEFINITION Mycoplasma genitalium G37 chromosome, complete genome.

ACCESSION NC\_000908

VERSION NC\_000908.2 GI:108885074

ORIGIN

1 TTGTTATTAG TATTACAGTA ATATTTACTT AATACATAAG TAGTATTTAT TAAGTATTAA

61 AGCAATATAA AATTAGTATT TGTCTAATAT CTAAATAATA C

//

**Q2. Download and un-compress (use “unzip”) the 10 Nucleotide sequences from Canvas (seq1-10). Concatenate all the nucleotide sequences together into a new assignment directory called assignment2. Find the length, description, percentage GC content and name of each sequence. Which sequence has the highest %GC content?**

jimmy@jimmy-VirtualBox[jimmy] unzip Assignment\ 2\ sequences.zip [ 8:08PM]

jimmy@jimmy-VirtualBox[jimmy] mkdir assignment2 [ 8:08PM]

**jimmy@jimmy-VirtualBox[jimmy] mv \*.fasta ~/assignment2 [ 8:08PM]**

**jimmy@jimmy-VirtualBox[jimmy] cd assignment2 [ 8:09PM]**

**jimmy@jimmy-VirtualBox[assignment2] cat \*.fasta > assignment2 [ 8:13PM]**

**jimmy@jimmy-VirtualBox[assignment2] infoseq assignment2 -only -length -desc -name -pgc**

**Display basic information about sequences**

**Name Length %GC Description**

**Lactobacillus\_acidophilus\_ATCC\_4796\_uid55489 1845 38.81**

**Lactobacillus\_johnsonii\_ATCC\_33200\_uid55505 1875 37.60**

**Lactobacillus\_helveticus\_DPC\_4571\_uid58761 1827 40.07**

**Lactobacillus\_gasseri\_ATCC\_33323\_uid57687 1872 38.41**

**Lactobacillus\_delbrueckii\_bulgaricus\_ATCC\_11842\_uid58647 1845 50.57**

**Lactobacillus\_crispatus\_ST1\_uid48359 1854 39.37**

**Lactobacillus\_casei\_ATCC\_334\_uid57985 1875 47.47**

**Lactobacillus\_buchneri\_ATCC\_11577\_uid55495 1887 39.37**

**Lactobacillus\_brevis\_ATCC\_367\_uid57989 1860 45.81**

**Lactobacillus\_amylovorus\_GRL\_1112\_uid61179 1863 42.57**

**The sequence with the highest %GV content is Lactobacillus\_delbrueckii\_bulgaricus\_ATCC\_11842\_uid58647 with 50.57**

**Q3. Perform a Needle-Wunsch global alignment on the two sequences with the highest %GC content. (Chose a suitable substitution matrix).**

**jimmy@jimmy-VirtualBox[assignment2] needle -asequence seq4.fasta -bsequence seq6.fasta -outfile ednafull**

**Q4. Translate the Nucleotide sequence with the highest %GC into all 6 open reading frames.**

**jimmy@jimmy-VirtualBox[assignment2] sixpack [ 8:55PM]**

**Display a DNA sequence with 6-frame translation and ORFs**

**Input nucleotide sequence: seq4.fasta**

**Output file [lactobacillus\_delbrueckii\_bulgaricus\_atcc\_11842\_uid58647.sixpack]: question4**

**protein output sequence(s) [lactobacillus\_delbrueckii\_bulgaricus\_atcc\_11842\_uid58647.fasta]: question4.fasta**

**Give an explanation as to how you found the command and why you selected this command.**

**This is a popular command among users of BioLinux and a quick search and you will find the command ‘sixpack’ which allows a user to display a DNA sequence with 6-frame translation and ORFS.**

**Paste in the first 15 lines.**

**Lactobacillus\_delbrueckii\_bulgaricus\_ATCC\_11842\_uid58647**

**M S K V I G I D L G T T N S A V A V L E F1**

**C L K L L G L T \* V P P T Q Q L R F W K F2**

**V \* S Y W D \* L R Y H Q L S S C G F G R F3**

**1 ATGTCTAAAGTTATTGGGATTGACTTAGGTACCACCAACTCAGCAGTTGCGGTTTTGGAA 60**

**----:----|----:----|----:----|----:----|----:----|----:----|**

**1 TACAGATTTCAATAACCCTAACTGAATCCATGGTGGTTGAGTCGTCAACGCCAAAACCTT 60**

**X D L T I P I S K P V V L E A T A T K S F6**

**X T \* L \* Q S Q S L Y W W S L L Q P K P F5**

**H R F N N P N V \* T G G V \* C N R N Q F F4**

**How many ORF’s are in frame 4?**

**Total ORFs in frame 4 : 20**

**Q5. Download and unzip the swissprot database from the local server at:** <http://bioinfo.ucc.ie/temp/swissprot.gz> **and prepare for BLAST searches. Concatenate the 5 files with the highest %GC content from question 2, and translate (one frame is sufficient) the sequences putting the output into a new file.**

jimmy@jimmy-VirtualBox[assignment2] wget "http://bioinfo.ucc.ie/temp/swissprot.gz"

jimmy@jimmy-VirtualBox[assignment2] gunzip -d swissprot.gz [ 9:17PM]

jimmy@jimmy-VirtualBox[assignment2] formatdb -i swissprot -p T -o T [ 9:18PM]

jimmy@jimmy-VirtualBox[assignment2] cat seq4.fasta seq6.fasta seq8.fasta seq9.fasta seq2.fasta > question5.fasta

**Q6.** **Perform a BLAST search of the concatenated file created in question 5 with the results being displayed in tabular, including comment lines (header), only allowing E-values less than 1e-50 and returning only the top hit per sequence**

jimmy@jimmy-VirtualBox[assignment2] blastp -query question5.fasta -out question5.blast.txt -db swissprot -outfmt 7 -max\_target\_seqs 1 -evalue 1e-50

OUTPUT:

# BLASTP 2.2.28+

# Query: Lactobacillus\_delbrueckii\_bulgaricus\_ATCC\_11842\_uid58647

# Database: swissprot

# 0 hits found

# BLASTP 2.2.28+

# Query: Lactobacillus\_casei\_ATCC\_334\_uid57985

# Database: swissprot

# 0 hits found

# BLASTP 2.2.28+

# Query: Lactobacillus\_brevis\_ATCC\_367\_uid57989

# Database: swissprot

# 0 hits found

# BLASTP 2.2.28+

# Query: Lactobacillus\_amylovorus\_GRL\_1112\_uid61179

# Database: swissprot

# 0 hits found

# BLASTP 2.2.28+

# Query: Lactobacillus\_helveticus\_DPC\_4571\_uid58761

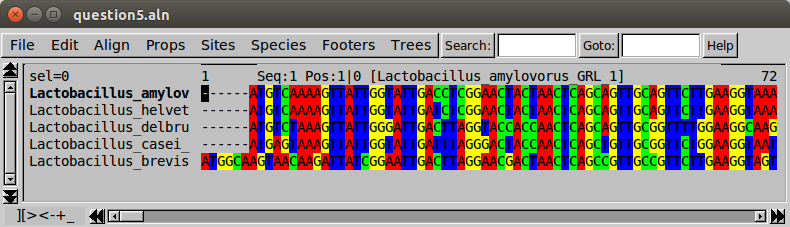
# Database: swissprot

# 0 hits found

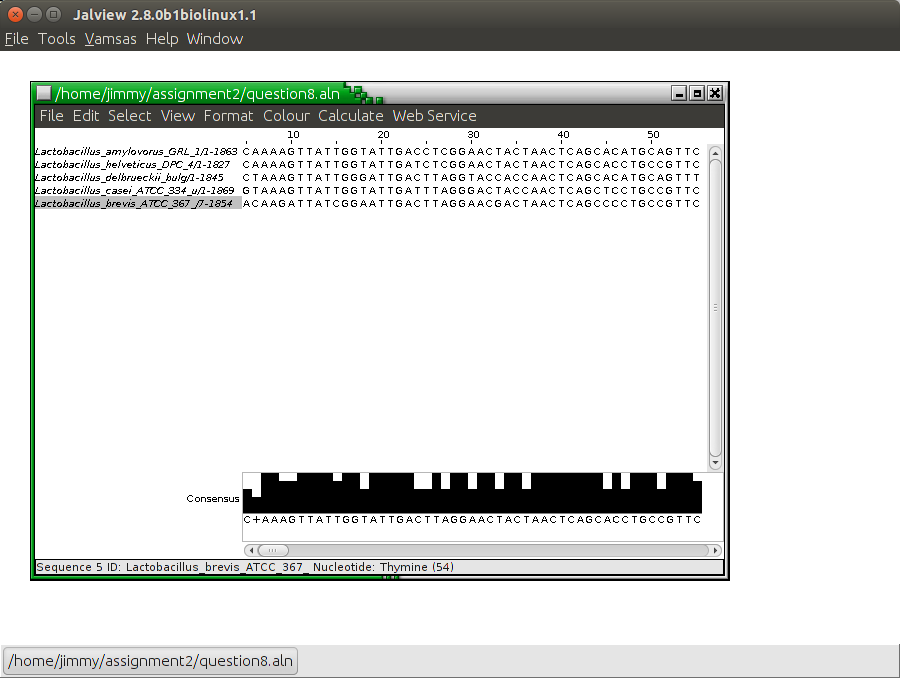
# BLAST processed 5 queries

**Q7. Perform a clustalw to align the protein sequences created in Q5.**

**jimmy@jimmy-VirtualBox[assignment2] clustalw question5.fasta [ 9:33PM]**



**Q8. Use Jalview to remove alignment columns with ≥50% gaps. Save the ungapped alignment in a new clustal file.**



**Q9. Create a distance matrix from your original alignment (with gaps), with the gap penalty being set to 5. (Paste the results using suitable font and size)**

jimmy@jimmy-VirtualBox[assignment2] distmat -gapweight 5 [10:22PM]

Create a distance matrix from a multiple sequence alignment

Input (aligned) sequence set: question5.aln

Multiple substitution correction methods for nucleotides

0 : Uncorrected

1 : Jukes-Cantor

2 : Kimura

3 : Tamura

4 : Tajima-Nei

5 : Jin-Nei Gamma

Method to use [0]: 0

Phylip distance matrix output file [question5.distmat]:

Distance Matrix

---------------

Uncorrected for Multiple Substitutions

Using base positions 123 in the codon

Gap weighting is 5.000000

1 2 3 4 5

0.00 20.31 28.99 31.44 35.25 Lactobacillus\_amylovorus\_GRL\_1 1

0.00 31.90 36.77 39.52 Lactobacillus\_helveticus\_DPC\_4 2

0.00 32.05 36.70 Lactobacillus\_delbrueckii\_bulg 3

0.00 30.22 Lactobacillus\_casei\_ATCC\_334\_u 4

0.00 Lactobacillus\_brevis\_ATCC\_367\_ 5

**Q 10. Compare the statistics on the clustal alignment with gaps, the alignment without gaps and muscle alignment. In the table include the following; number of different positions, align length, number of gaps and the headings for each.**

jimmy@jimmy-VirtualBox[assignment2] infoalign question7 -only -diffcount -alignlength -gaps -heading -outfile 10A

jimmy@jimmy-VirtualBox[assignment2] infoalign question8 -only -diffcount -alignlength -gaps -heading -outfile 10B

jimmy@jimmy-VirtualBox[assignment2] infoalign qmuscle -only -diffcount -alignlength -gaps -heading -outfile 10C