**Assignment 3 – Phylogenetic tree’s, Artemis and ACT (worth 10% of MB6301)**

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

**Demonstrator: Jamie FitzGerald (**[**jamie.fitzgerald@ucc.ie**](mailto:jamie.fitzgerald@ucc.ie)**)**

**Student: James Henry Hehir (**[**120224791@umail.ucc.ie**](mailto:120224791@umail.ucc.ie)**)**

**Deadline: By the end of Sunday the 11th of April, 2021.**

**Writing this assignment:**

**For each question, p**lease record your answer in a Word document, including what you typed on the command-line to get this answer. Graphical output can be captured in a screenshot if necessary.

**When no specific question is asked, perform the tasks and paste in the prompt (e.g. “bio@linux:/home/bio/dir$“ ) and full command input, 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.**

**You may use any up-to-date version of the programs mentioned (e.g. in biolinux, in windows via java, etc..) as long as it is recorded and provided in the assignment.**

**To figure out how a command should be used, try e.g. ‘man command’, ‘info command’, ‘command –help’ or ‘command -h’.**

**Submitting this assignment:**

When finished, upload the document (with your name in the title) with your answers to Canvas under Assignments and MB6301 Assignment 1.

**Ensure you submit your entry before the deadline, or the system may reject your assignment!**

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

**1) Download the following genomes:**

**AE005176.1 Lactococcus lactis subsp. lactis Il1403**

**CP003132.1 Lactococcus lactis subsp. cremoris A76**

**CP004884.1 Lactococcus lactis subsp. cremoris KW2**

**CP000425.1 Lactococcus lactis subsp. cremoris SK11**

**CP003157.1 Lactococcus lactis subsp. cremoris UC509.9**

**CP001834.1 Lactococcus lactis subsp. lactis KF147**

**CP000024.1 Streptococcus thermophilus CNRZ1066**

**Using artemis, extract both amino acid and nucleotide sequences of the following genes from each genome and save in FASTA format: *dnaA; groEL; grpE***

**jimmy@jimmy-VirtualBox[assignment3] ls [ 3:59PM]**

**AE005176.1 CP000024.1.tar CP001834.1 CP003132.1.tar CP004884.1**

**AE005176.1.tar CP000425.1 CP001834.1.tar CP003157.1 CP004884.1.tar**

**CP000024.1 CP000425.1.tar CP003132.1 CP003157.1.tar**

**jimmy@jimmy-VirtualBox[jimmy] sudo art**

**2)** Concatenate the amino acid sequences of *dnaA*, *groEL* and *grpE* from each genome into respective fasta files. These will be used to construct a *groEL*, *dnaK*, *grpE* multilocus tree.

*Note: See this paper for additional information http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107232#pone-0107232-g002*

I moved all fasta files for each gene into a separate folder for each genome and I used the following command to concatenate the files in each folder

**jimmy@jimmy-VirtualBox[AE005176.1] cd amino\_acids [ 5:08PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > ae005176.1.fasta [ 5:08PM]**

**jimmy@jimmy-VirtualBox[CP000425.1] cd amino\_acids [ 5:10PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > cp000425.fasta [ 5:10PM]**

**jimmy@jimmy-VirtualBox[CP003132.1] cd amino\_acids [ 5:11PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > cp003132.1.fasta**

**jimmy@jimmy-VirtualBox[CP004884.1] cd amino\_acids [ 5:12PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > cp004884.1.fasta [ 5:12PM]**

**jimmy@jimmy-VirtualBox[CP000024.1] cd amino\_acids [ 5:13PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > cp000024.1.fasta**

**jimmy@jimmy-VirtualBox[CP001834.1] cd amino\_acids [ 5:14PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > cp001834.1.fasta [ 5:14PM]**

**jimmy@jimmy-VirtualBox[CP003157.1] cd amino\_acids [ 5:16PM]**

**jimmy@jimmy-VirtualBox[amino\_acids] cat \* > cp003157.1.fasta**

**3)** Concatenate the seven resulting FASTA files and build a MUSCLE alignment.

Moved all the files made in question 2 into a separate folder and concatenated the seven fasta files into 1.

**jimmy@jimmy-VirtualBox[all\_genomes\_fasta] cat \* > muscle\_file.fasta [ 5:20PM]**

**jimmy@jimmy-VirtualBox[all\_genomes\_fasta] muscle -in muscle\_file.fasta -out output.fasta**

**MUSCLE v3.8.31 by Robert C. Edgar**

**http://www.drive5.com/muscle**

**This software is donated to the public domain.**

**Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.**

**muscle\_file 21 seqs, max length 542, avg length 392**

**00:00:00 22 MB(2%) Iter 1 100.00% K-mer dist pass 1**

**00:00:00 22 MB(2%) Iter 1 100.00% K-mer dist pass 2**

**00:00:00 29 MB(2%) Iter 1 100.00% Align node**

**00:00:00 29 MB(2%) Iter 1 100.00% Root alignment**

**00:00:00 29 MB(2%) Iter 2 100.00% Root alignment**

**00:00:00 30 MB(3%) Iter 3 100.00% Refine biparts**

**00:00:01 30 MB(3%) Iter 4 100.00% Refine biparts**

**00:00:01 30 MB(3%) Iter 5 100.00% Refine biparts**

**00:00:01 30 MB(3%) Iter 5 100.00% Refine biparts**

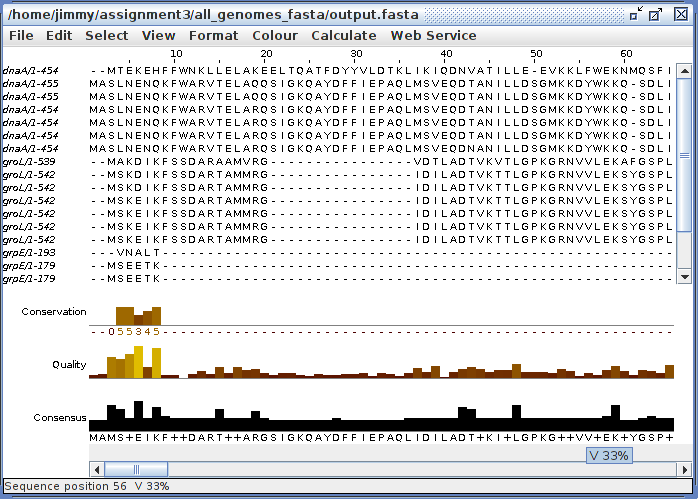
**00:00:01 30 MB(3%) Iter 6 100.00% Refine biparts**

**00:00:01 30 MB(3%) Iter 7 100.00% Refine biparts**

**00:00:01 30 MB(3%) Iter 8 100.00% Refine biparts**

**00:00:01 30 MB(3%) Iter 8 100.00% Refine biparts**

**jimmy@jimmy-VirtualBox[all\_genomes\_fasta] [ 5:25PM]**

**4)** With the output from **Q3),** use Jalview to create an ungapped alignment. Select an appropriate percentage of gaps to remove and explain your choice.

Graphical user interface, text, application

Description automatically generatedI removed over 50% of the gaps to make a better alignment overall.

**5)** Use the EMBOSS command ‘fprotdist’ to build a distance matrix. Correct for multiple amino acid substitutions using the Daythoff PAM model.

**jimmy@jimmy-VirtualBox[all\_genomes\_fasta] fprotdist -sequence gaps\_removed.fasta -method d -outfile dayhoff\_method.fasta**

**Protein distance algorithm**

**Computing distances:**

**dnaA\_1-454**

**dnaA\_1-455 .**

**dnaA\_1-455 ..**

**dnaA\_1-454 ...**

**dnaA\_1-454 ....**

**dnaA\_1-454 .....**

**dnaA\_1-454 ......**

**groL\_1-539 .......**

**groL\_1-542 ........**

**groL\_1-542 .........**

**groL\_1-542 ..........**

**groL\_1-542 ...........**

**groL\_1-542 ............**

**groL\_1-542 .............**

**grpE\_1-193 ..............**

**grpE\_1-179 ...............**

**grpE\_1-179 ................**

**grpE\_1-179 .................**

**grpE\_1-179 ..................**

**grpE\_1-179 ...................**

**grpE\_1-179 ....................**

**Output written to file "dayhoff\_method.fasta"**

**Done.**

**6)** Use the EMBOSS command ‘fneighbor’ to generate an un-rooted tree from the output.

**jimmy@jimmy-VirtualBox[all\_genomes\_fasta] fneigbor -datafile dayhoff\_method.fasta -outfile fneighbor**

**zsh: correct 'fneigbor' to 'fneighbor' [nyae]? y**

**Phylogenies from distance matrix by N-J or UPGMA method**

**Cycle 18: species 2 ( 0.00000) joins species 3 ( 0.00000)**

**Cycle 17: node 2 ( 4.12993) joins species 4 ( -1.33635)**

**Cycle 16: node 2 ( 0.70748) joins species 5 ( -0.70748)**

**Cycle 15: node 2 ( 0.37585) joins species 6 ( -0.37584)**

**Cycle 14: node 2 ( 0.20702) joins species 7 ( -0.20226)**

**Cycle 13: species 1 ( 3.02612) joins node 2 ( 2.94858)**

**Cycle 12: node 1 ( 14.06168) joins species 11 ( -6.21886)**

**Cycle 11: node 1 ( 3.36855) joins species 12 ( -3.36855)**

**Cycle 10: node 1 ( 1.83739) joins species 13 ( -1.83739)**

**Cycle 9: node 1 ( 1.00628) joins species 10 ( -0.97155)**

**Cycle 8: node 1 ( 0.55649) joins species 14 ( -0.56965)**

**Cycle 7: node 1 ( 0.32208) joins species 9 ( -0.31289)**

**Cycle 6: node 1 ( 0.26842) joins species 8 ( -0.05534)**

**Cycle 5: node 1 ( 5.73532) joins species 15 ( 4.35509)**

**Cycle 4: node 1 ( 1.98323) joins species 16 ( -0.02062)**

**Cycle 3: node 1 ( 0.02790) joins species 18 ( -0.00142)**

**Cycle 2: node 1 ( 0.00172) joins species 17 ( 0.00569)**

**Cycle 1: node 1 ( 0.00448) joins species 19 ( 0.00000)**

**last cycle:**

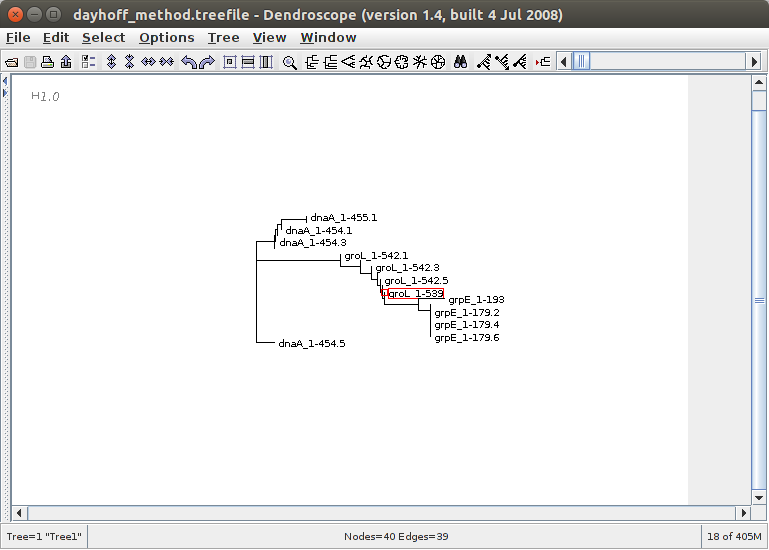
**node 1 ( 0.00000) joins species 20 ( 0.00000) joins species 21 ( 0.00000)**

**Output written on file "fneighbor"**

**Tree written on file "dayhoff\_method.treefile"**

**Done.**

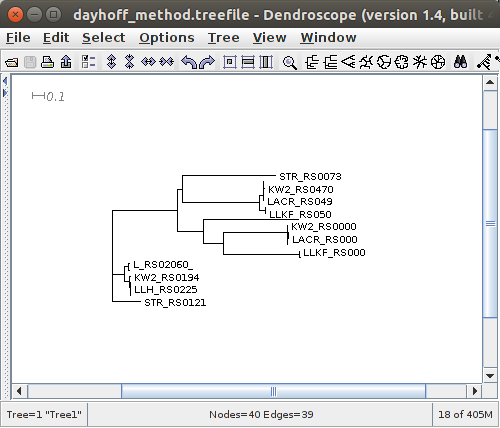
**jimmy@jimmy-VirtualBox[all\_genomes\_fasta]**

**7)** Use ‘dendroscope’ to select a suitable tree type, and paste the output to your assignment. Explain why you have chosen this tree type.

I picked the rectangular phylogram tree as it describes the genes better than the other options. It also represents the evolutionary relationships among organisms. This tree shows how the genes may have a common ancestor.

**8)** Repeat exercise using the nucleotide sequences extracted in **Q2)** of the genes in the Mega package. Provide any relevant comments on gaps, parameters selected, tree-type and any differences or similarities observed with the amino acid tree.

I repeated the exact same procedure with the bases and this is the tree that was created.



The tree is different from the amino acids tree. The tree is more readable and shows that there is greater link between the genes. The two trees have the same number of nodes and edges.

**9)** Use ACT to generate a multiple comparison at nucleotide level between the following genomes. Provide a screenshot.

[CP000425.1](http://www.ncbi.nlm.nih.gov/nuccore/CP000425.1) Lactococcus lactis subsp. cremoris SK11

[CP003157.1](http://www.ncbi.nlm.nih.gov/nuccore/CP003157.1) Lactococcus lactis subsp. cremoris UC509.9

[CP000024.1](http://www.ncbi.nlm.nih.gov/nuccore/CP000024.1) Streptococcus thermophilus CNRZ1066

**jimmy@jimmy-VirtualBox[question9] makeblastdb -in cp000435.1.fasta -dbtype nucl**

**jimmy@jimmy-VirtualBox[question9] blastn -query cp003157.1.fasta -db cp000425.1.fasta evalue 1 – task megablast -outfmt 6 > comparefile.crunch**

**jimmy@jimmy-VirtualBox[question9] makeblastdb -in cp0003157.1.fasta -dbtype nucl**

**jimmy@jimmy-VirtualBox[question9] blastn -query cp0000024.1.fasta -db cp0003157.1.fasta evalue 1 – task megablast -outfmt 6 > comparefile1.crunch**

**jimmy@jimmy-VirtualBox[question9] art**

I tried using artemis here to compare the files but every time I tried to get an output I got an error so I have no screenshot for here. However the method above shows how I got the crunch files for comparison.

**10**) Repeat using tBLASTx, how does this alignment compare to Q9? Comment. Describe any differences. Describe the levels of similarity between the genomes.