Date:31/12/20

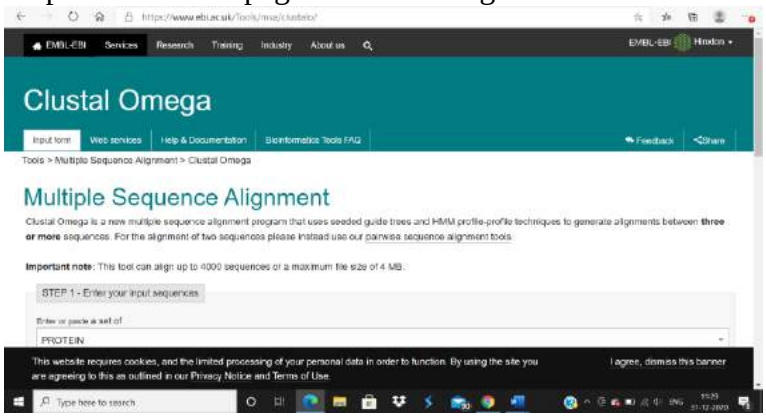
Phyloinformatics

Assignment-1

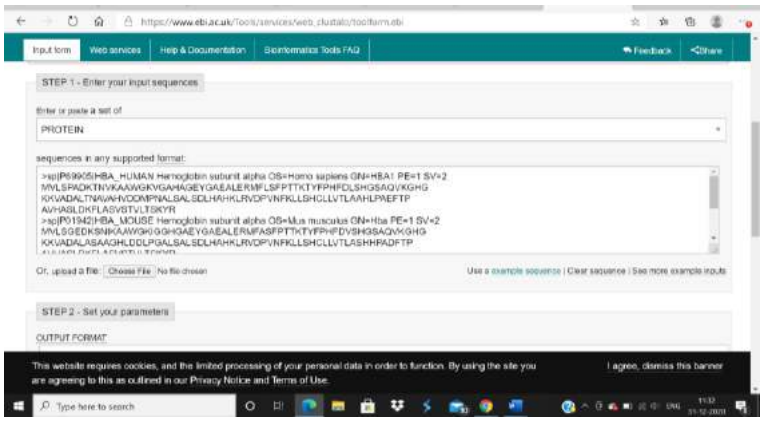
Q1-Prepare a report on "multiple sequence alignment using Clustal Omega" include the following sessions.

Answer: MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large Protein, DNA and RNA multiple sequence alignments. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a linkage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences. **Multiple Sequence Alignment (MSA)** is generally the alignment of **three or more** biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

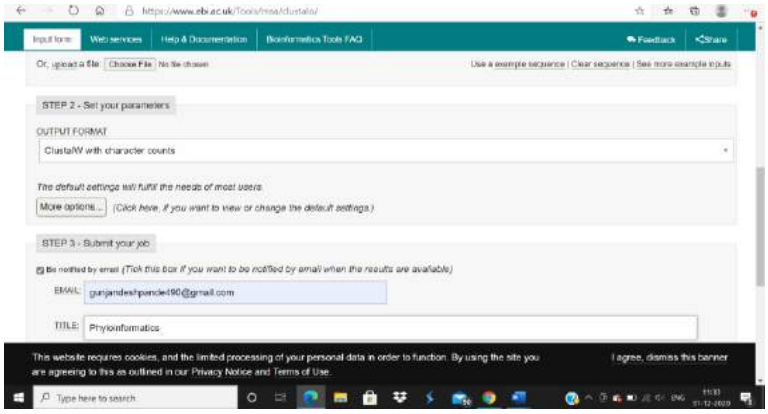
Step1: Go to homepage of Clustal omega in web browser.



Step2:We can use the sequences already given in the website below the sequence drop box as: "Use a example sequence | Clear sequence | See more example inputs"

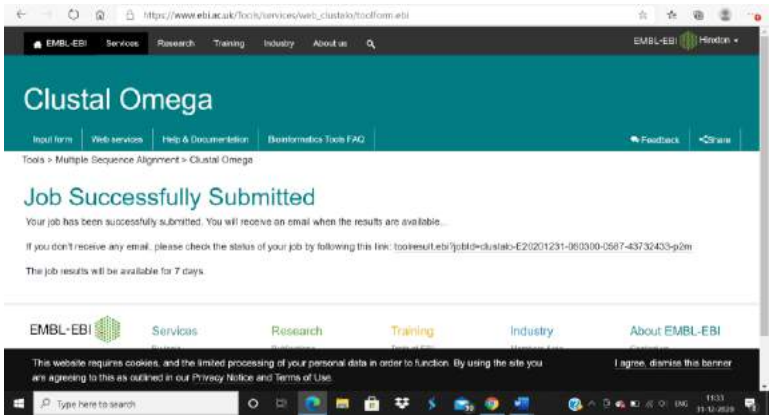


Step3: Now we have to paste sequences in the dialog box given. Just in case the sequences can also be submitted through file by clicking on the option "choose file" such that all the sequences should be in similar format. The other two steps the user can select on his/her own to set the parameters for pair wise alignment options and multiple sequence alignment options, to select the scoring matrices and scoring values.

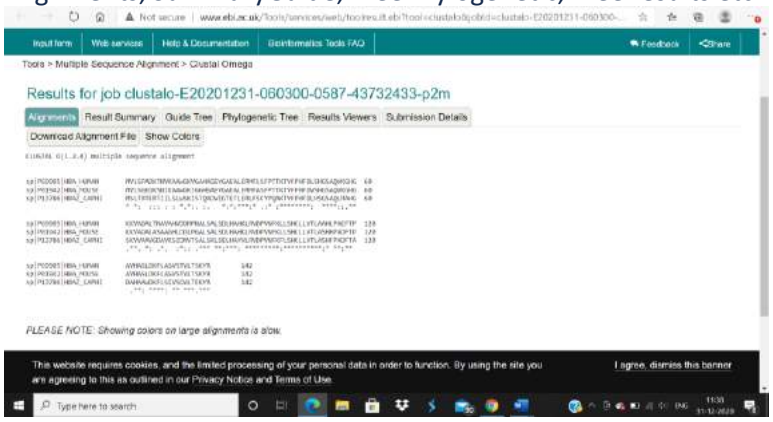


Step4: We will paste our sequences and wait for the results. Results can be notified by email when the user checks the button email notification. After the submission of the job the results can be downloaded into a file by clicking on the option Download alignment file. The result summary tab gives the links to different outputs summary and link to each output. The result files are with different formats of input and output files of the alignment.The user can enable the java plug-in in the browser, if it is disabled and thus the user can use Jalview to see the alignment with the colours. The user can view the output file and can save by clicking on the button “View output file”. The output file represents the length of each sequence , and the score of each alignment individually.

Step5:You have submitted the sequences.



Step6:Results will be mailed on your email id .It gives us information about all the aspects about Alignments, SummaryGuide, TreePhylogenetic, TreeResults etc.

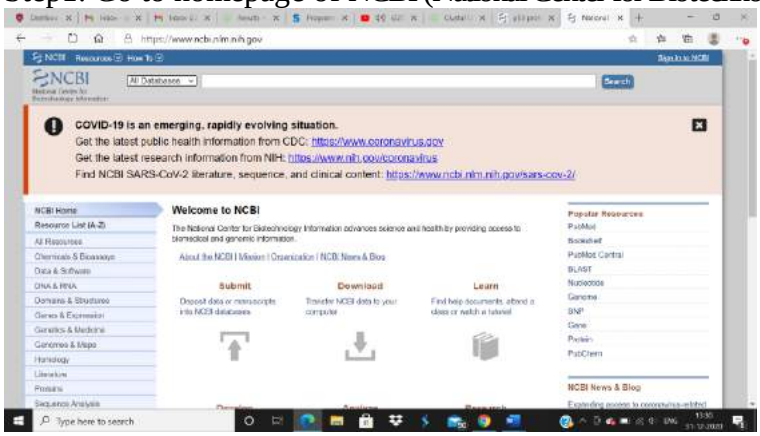


1-Introduction about MSA, Clustal Omega

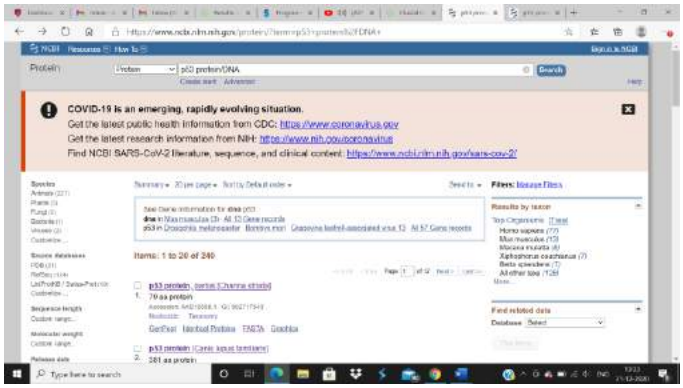
Answer: Alignment of three or more biological nucleotides or protein sequences, simply defines multiple sequence alignment. The genes which are similar may be conserved among different species. Multiple sequence alignment (MSA) is defined as to be the process or the result of sequence alignment of three or more biological sequences, basically protein, DNA, or RNA.

Clustal is a series of widely used computer programs used in Bioinformatics for multiple sequence alignment. There have been many versions of Clustal over the development of the algorithm that are listed below. The analysis of each tool and its algorithm are also detailed in their respective categories. Available operating systems listed in the sidebar are a combination of the software availability and may not be supported for every current version of the Clustal tools. Clustal Omega has the widest variety of operating systems out of all the Clustal tools. Clustal Omega has the widest variety of operating systems out of all the Clustal tools. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. Clustal Omega is nothing but a multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. It produces biologically meaningful multiple sequence alignments of divergent sequences. To be honest Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few hours. It will also make use of multiple processors, where present. In addition, the quality of alignments is superior to previous versions, as measured by a range of popular benchmarks. Clustal Omega is currently a command line-only tool.

2-Download any five p53 protein/DNA sequences (OTUs) from different organism
Answer:
Step1: Go to homepage of NCBI (National Center for Biotechnology Information) in web browser.

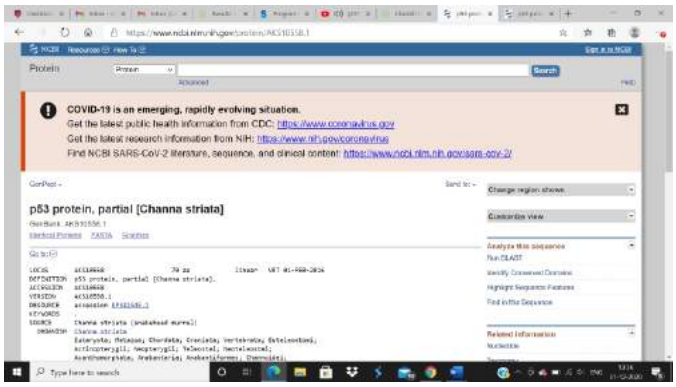


Step2: In the All Database drop box select ‘Protein’ and type p53 protein and click on search.



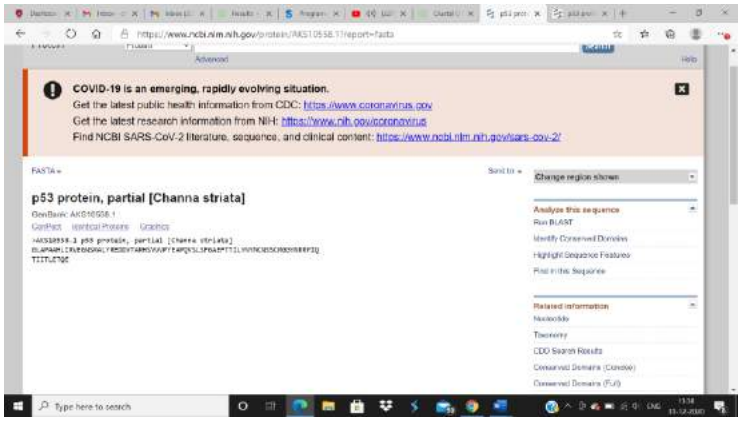
Step3: You will see 1 to 20 of 240 searches . You can choose any 5 amoung them.

Step4: Click on the first one protein to get the fasta sequence that is : p53 protein, partial [Channa striata].

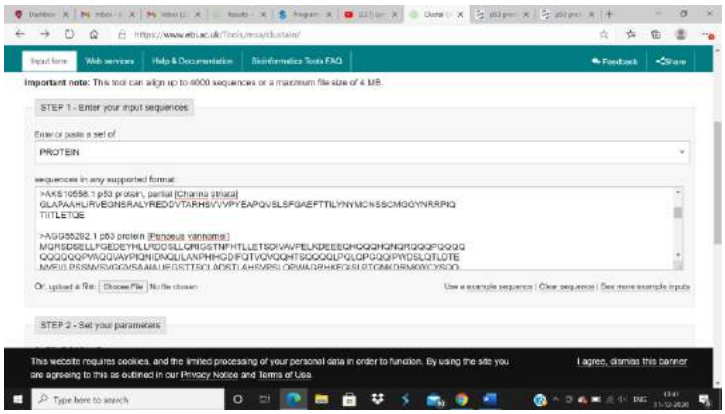


Step5: You will get the fasta format sequence of p53 protein, partial [Channa striata].

>AKS10558.1 p53 protein, partial [Channa striata]
GLAPAAHLIRVEGNSRALYREDDVTARHSVVVPYEAPQVSLSFGAEFTTILYNYMCN
SSCMGGYNRRPIQ
TIITLETQE

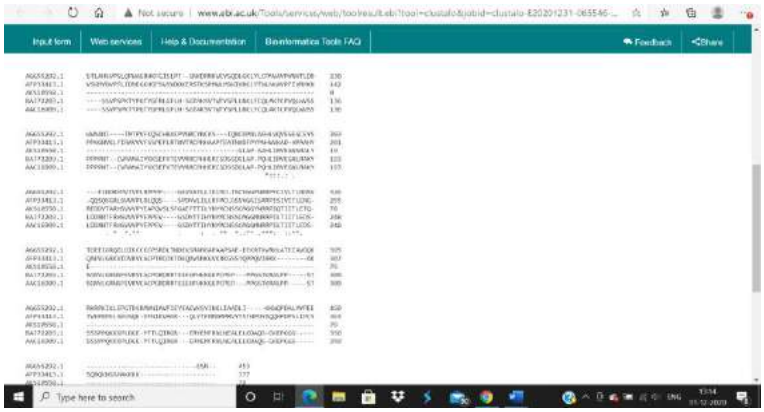


Step6: Collect fasta format sequences for all 5 from different organism and paste in Clustal omega. Go to homepage of Clustal omega in web browser. Now we have to paste sequences in the dialog box given. Just in case the sequences can also be submitted through file by clicking on the option "choose file" such that all the sequences should be in similar format. The other two steps the user can select on his/her own to set the parameters for pair wise alignment options and multiple sequence alignment options, to select the scoring matrices and scoring values. We will paste our sequences and wait for the results.



Step7: Results can be notified by email when the user checks the button email notification. After the submission of the job the results can be downloaded into a file by clicking on the option Download alignment file. The result summary tab gives the links to different outputs summary and link to each output. The result files are with different formats of input and output files of the alignment. The user can enable the java plug-in in the browser, if it is disabled and thus the user can use Jalview to see the alignment with the colours. The user can view the output file and can save by clicking on the button "View output file". The output file represents the length of each sequence , and the score of each alignment individually.


Clustal Omega
Results for job clustalo-E20201231-065546-0037-5603331-p2m




PLEASE NOTE: Showing colors on large alignments is slow.
EMBL-EBI
Services
Research
Training
Industry
About EMBL-EBI

i) Eriocheir sinensis
>AFP33413.1 p53 protein, partial [Eriocheir sinensis]
MDDETEVAKAKLRHQELMGHIKVEMDEGSEESADDPTDDHTMVIATQAVTLPQENGYYASPVPLTAGQSQ
LIVSQSSPLQGGVSGMVDVPTLIDMEGKHGFSVSVDDKERSTKSPMWLMSNIVNKLYTNLNKAVPFEVRM
KNPPKGNVKLFIRAVVVFSSPEFLRTNVTRCPNHAAPTEATNHDFPYPNHVVKADHPAAHYQQSQSGRLS
VVVPLDLQQSSPDYVLILLRFMCLGSSVGGISRRPISIVITLENGQAEVLGRKVIDVRVCACPTRDIKTD
EQAVSNKGVKRKGSSTQPPQVIRKKAKTVEPRPELSEGSQEVFNIKVHGRQLYTFMMDMMRVYYSTHPDY
AQQHPDPSLIPCSSQRQKNSSHAKKKK

ii) Channa striata
>AKS10558.1 p53 protein, partial [Channa striata]
GLAPAAHLIRVEGNSRALYREDDVTARHSVVVPYEAPQVSLSFGAEFTTILYNYMCNSSCMGGYNRRPIQ
TIITLETQE

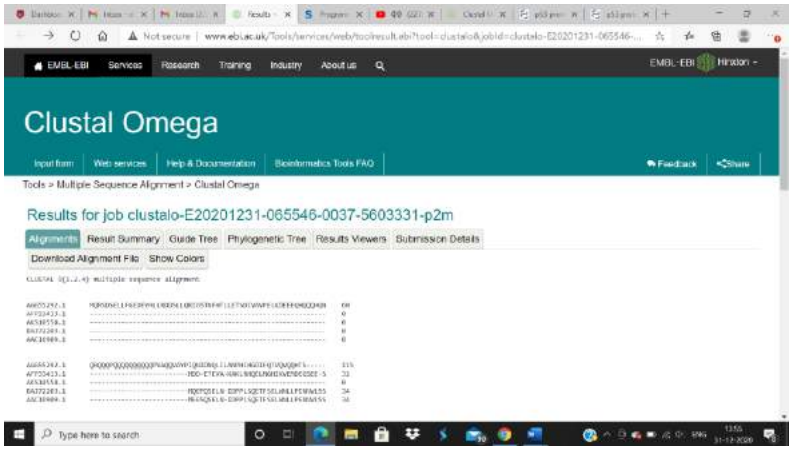
iiiPenaeus vanname
)>AGG55292.1 p53 protein [Penaeus vannamei]
MQRSDSELLFGEDEYHLLRDDSLLQRIGSTNFHTLLETSDIVAVPELKDEEEQHQQQHQNQRQQQPQQQQ
QQQQQQPVAQQVAYPIQNIDNQLILANPHIHGDIFQTVQVQQHTSQQQQLPQLQPGQQIPWDSLQTLDTE
NVEVLPSSNVSVGGVSAAIALIEGSTTSCLADSTLAHSVPSLQPWAGRHKFGISLPTGNKDRNKWCYSQD
LGKLYLCPNVAVPVNVTLDDWVNANITMTPVFKQSCHRAEPVNRCYNCKSIQNCDPNLAEHLVQVEGEGC
EYSFINDRYMVTVPLRPPPPGEVSSTLLIKIMCLTSCVGGPNRRPFCIVLTLRNSVTGEEIGRQILDIKC
CKCPSRDLTNDEKSRARGAPAAPSAEEEKRTKVRKLATEIAVGQKRKRPKIKLEPGTDSRMVNIAVPIEY
EAEVKSYINKLIAADLIKKWQPDALMYPEEESN

iv) Canis lupus familiaris
>BAJ72203.1 p53 protein [Canis lupus familiaris]
MQEPQSELNIDPPLSQETFSELWNLLPENNVLSSELCPAVDELLLPESVVNWLDEDSDDAPRMPATSAPT
APGPAPSWPLSSSVPSPKTYPGTYGFRLGFLHSGTAKSVTWTYSPLLNKLFCQLAKTCPVQLWVSSPPPP
NTCVRAMAIYKKSEFVTEVVRRCPHHERCSDSSDGLAPPQHLIRVEGNLRAKYLDDRNTFRHSVVVPYEP
PEVGSDYTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNVLGRNSFEVRVCACPGRDRRTEEENFHKK
GEPCPEPPPGSTKRALPPSTSSSPPQKKKPLDGEYFTLQIRGRERYEMFRNLNEALELKDAQSGKEPGGS
RAHSSHLKAKKGQSTSRHKKLMFKREGPDSD

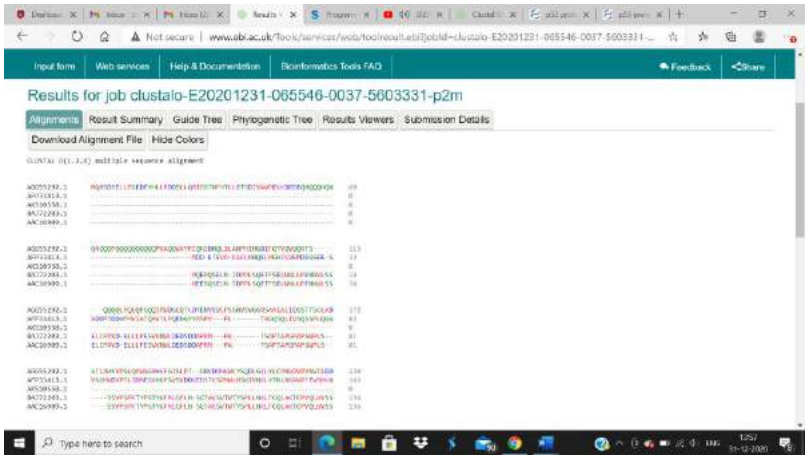
v) Canis lupus familiaris
>AAC16909.1 p53 protein [Canis lupus familiaris]
MEESQSELNIDPPLSQETFSELWNLLPENNVLSSELCPAVDELLLPESVVNWLDEDSDDAPRMPATSAPT
APGPAPSWPLSSSVPSPKTYPGTYGFRLGFLHSGTAKSVTWTYSPLLNKLFCQLAKTCPVQLWVSSPPPP
NTCVRAMAIYKKSEFVTEVVRRCPHHERCSDSSDGLAPPQHLIRVEGNLRAKYLDDRNTFRHSVVVPYEP
PEVGSDYTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNVLGRNSFEVRVCACPGRDRRTEEENFHKK
GEPCPEPPPGSTKRALPPSTSSSPPQKKKPLDGEYFTLQIRGRERYEMFRNLNEALELKDAQSGKEPGGS
RAHSSHLKAKKGQSTSRHKKLMFKREGLDSD

3-Show the alignment with color codes (give general interpretation)

Step1: When you receive results on your email id ,click on show colours.



Step2: You will be able to see the alignment with color codes.

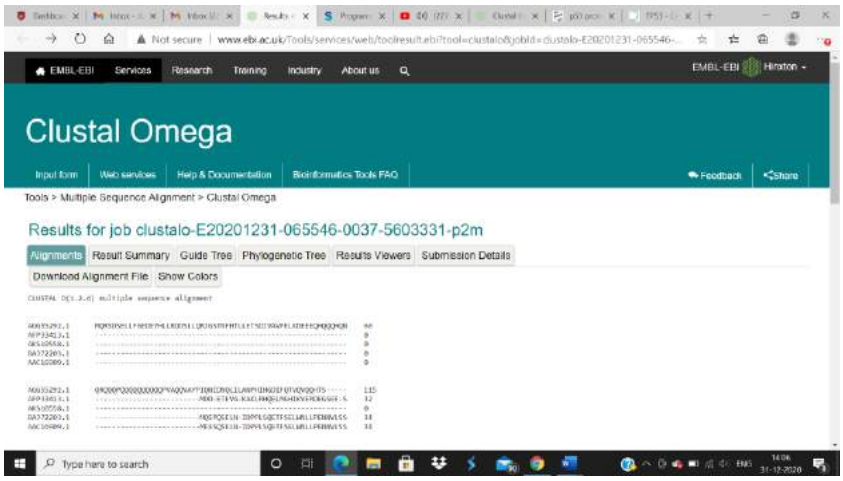


So the general color code represents as:

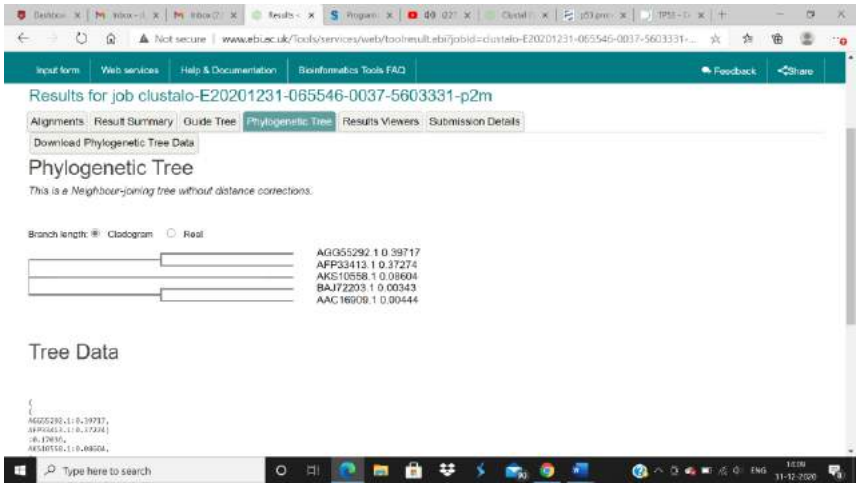
Category	Colour	Residue at position
Hydrophobic	BLUE	A,I,L,M,F,W,V
		C
Positive charge	RED	K,R
Negative charge	MAGENTA	E
		D
Polar	GREEN	N
		Q
		S,T
Cysteines	PINK	C
Glycines	ORANGE	G
Prolines	YELLOW	P
Aromatic	CYAN	H,Y
Unconserved	WHITE	any / gap

4-Give Phylogram and cladogram (give general interpretation)

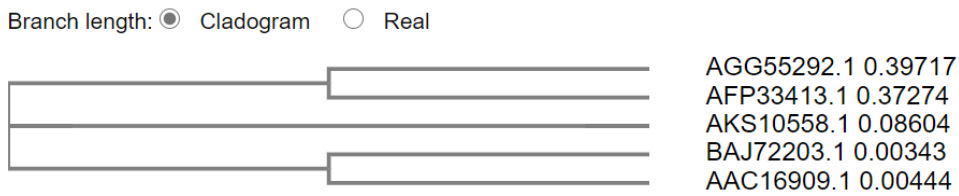
Step1: When you receive results on your email id ,click on show Phylogenetic Tree.



Step2: You will be able to see (Phylogram and cladogram) the Phylogenetic Tree and Tree Data

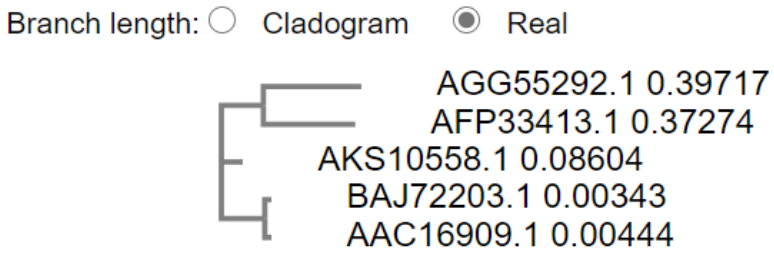


Phylogenetic Tree :
This is a Neighbour-joining tree without distance corrections.



Phylogenetic analysis aims to study the evolutionary relationships among the different organisms. It is the study of evolutionary relatedness among various groups of organisms (for example, species, and populations). Owing to the technological advancement of the sequencing techniques in molecular biology and the ability to collect large amounts of data (DNA or amino acid sequences) from disparate organisms, phylogenetic analysis and evolutionary studies are still of highest interest. A phylogenetic tree is an evolutionary tree that shows the evolutionary relationships between different groups of animals.

Cladograms give a hypothetical picture of the actual evolutionary history of the organisms.
Real – this is the phylogram which shows branch length with nexus tree format

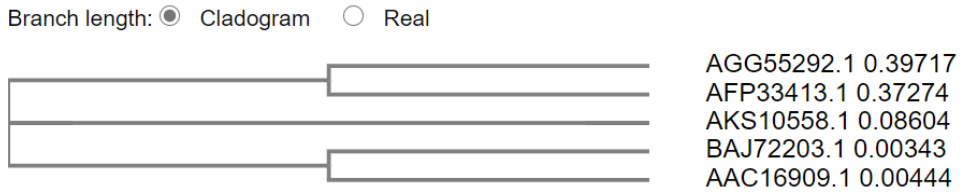


5-How many rooted and unrooted trees are possible from the above 5 OTUs (Use equation to solve and provide it in report)
Answer:
Given:

Nu = no. Of unrooted trees
NR = rooted
And n = no. Of sequence that is 5

Number of unrooted trees for n taxa Nu=(2n-5)*(2n-7)*...*3*1=(2n-5)!/[2n-3*(n-3)!]

Number of rooted trees for n taxa Nr=(2n-3)*(2n-5)*(2n-7)*...*3*1=(2n-3)!/[2n-2*(n-2)!]
So,
Number of unrooted trees=15
Number of rooted tress=105



Here we can see AGG55292.1 0.39717 and AFP33413.1 0.37274 have been grouped together with bootstrap 100%, and BAJ72203.1 0.00343 and AAC16909.1 0.00444 have been grouped together with bootstrap 100%. These four proteins have also been grouped together in a larger clade with bootstrap 100%.
As this is a rooted tree, we know the direction that evolutionary time ran. Say we call the ancestor of the four sequences (AGG55292.1 0.39717,AFP33413.1 0.37274, BAJ72203.1 0.00343 ,AAC16909.1 0.00444) ancestor1, the ancestor of the two sequences (AGG55292.1 0.39717,AFP33413.1 0.37274) ancestor2, and the ancestor of the two sequences (BAJ72203.1 0.00343 ,AAC16909.1 0.00444) ancestor3.

Because it is a rooted tree, we know that time ran from left to right along the branches of the tree, so that ancestor1 was the ancestor of ancestor2, and ancestor1 was also the ancestor of

ancestor3. In other words, ancestor1 lived before ancestor2 or ancestor3; ancestor2 and ancestor3 were descendants of ancestor1.

Another way of saying this is that AGG55292.1 0.39717 and AFP33413.1 0.37274 shared a common ancestor with each other more recently than they did with BAJ72203.1 0.00343 and AAC16909.1 0.00444

The lengths of branches in this tree are proportional to the amount of evolutionary change (estimated number of mutations) that occurred along the branches. The branches leading back from AGG55292.1 0.39717 and AFP33413.1 to their last common ancestor are slightly longer than the branches leading back from BAJ72203.1 0.00343 and AAC16909.1 0.00444 to their last common ancestor.
This indicates that there has been more evolutionary change in AGG55292.1 0.39717 and AFP33413.1 proteins since they diverged, than there has been in BAJ72203.1 0.00343 and AAC16909.1 0.00444 since they diverged. Compared to these four proteins, the AKS10558.1 0.08604 seem to be relatively distantly related.

Step1: Click on result summary.

Step2:Click on Tool Output: