Contents

General	3
Aim of the Project	3
Procedure	4
Step-1: Choice of Molecular Markers	4
Step-2: Multiple Sequence Alignment	6
Step-3: Determining a tree building method	10
Step-4: Evaluation of phylogenetic tree	15
Result Interpretation	17

General

Biological sequence analysis is founded on solid evolutionary principles. Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic trees. Thus, molecular phylogenetics is a fundamental aspect of bioinformatics.

Phylogenetics is the study of the evolutionary history of living organisms using treelike diagrams to represent pedigrees of these organisms. The tree branching patterns representing the evolutionary divergence are referred to as phylogeny.

Phylogenetics can be studied in various ways. Molecular data that are in the form of DNA or protein sequences can provide very useful evolutionary perspectives of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes. Because genes are the medium for recording the accumulated mutations, they can serve as molecular fossils. Through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

The field of molecular phylogenetics can be defined as the study of evolutionary relationships of genes, proteins and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules. Based on the sequence similarity of the molecules, evolutionary relationships between the organisms can often be inferred.

Aim of the Project

The aim of the project is to delineate the phylogenetic relationship among different species of Coronaviridae. Coronaviridae is a family of enveloped, positive-stranded RNA viruses that cause common cold and respiratory disease in humans. To construct phylogenetic relationship 10 species of the family Coronaviridae were selected.

Genus: Alphacoronavirus	Human coronavirus 229E, Human coronavirus NL63, Porcine epidemic diarrhea virus, Transmissible gastroenteritis virus
Genus: Betacoronavirus	Human coronavirus OC43, Human coronavirus HKU1, Pipistrellus bat coronavirus HKU5
Genus: Deltacoronavirus	Bulbul coronavirus HKU11, Munia coronavirus HKU13
Genus: Gammacoronavirus	Beluga whale coronavirus SW1

Procedure

Molecular phylogenetic tree construction can be divided into four steps:

- (1) Choosing molecular markers;
- (2) Performing multiple sequence alignment;
- (3) Determining a tree building method; and
- (4) Assessing tree reliability.

Step-1: Choice of Molecular Markers

For constructing molecular phylogenetic trees, one can use either nucleotide or protein sequence data. The choice of molecular markers is an important matter because it can make a major difference in obtaining a correct tree. The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study.

For studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins, can be used. For example, for evolutionary analysis of different individuals within a population, noncoding regions of mitochondrial DNA are often used.

For studying the evolution of more widely divergent groups of organisms, one may choose either slowly evolving nucleotide sequences, such as ribosomal RNA or protein sequences.

In many cases, protein sequences are preferable to nucleotide sequences because protein sequences are relatively more conserved as a result of the degeneracy of the genetic code in which sixty-one codons encode for twenty amino acids, meaning thereby a change in a codon may not result in a change in amino acid. Thus, protein sequences can remain the same while the corresponding DNA sequences have more room for variation, especially at the third codon position. The significant difference in evolutionary rates among the three nucleotide positions also violates one of the assumptions of tree-building. In contrast, the protein sequences do not suffer from this problem, even for divergent sequences.

Therefore in the present study, protein sequences were aimed to use for phylogenetic tree construction. The major viral proteins of coronaviruses include:

- A) Nucleocapsid (N) protein: (Phosphoprotein)
- B) Spike (S) protein: (Glycoprotein)
- C) Haemagglutinin-esterase (HE) protein
- D) Integral membrane (M) protein
- E) Small membrane (E) protein

Two of them (Nucleocapsid protein and Small membrane protein) were chosen for study. The protein sequences for each of the selected 10 species of Coronaviridae were collected from NCBI in FASTA format. The procedure is as follows:

a) Go to NCBI home page, select 'Protein' from the options in top left box and write the name of desired protein for search.

SNCBI Resources 🗵 How To 🖯	2	<u>Sign in to NCBI</u>
SNCBI National Center for Biotechnology Information	 nucleocapsid protein Human coronavirus 229E 	Search Search
NCBI Home Resource List (A-Z)	Welcome to NCBI	Popular Resources PubMed
All Resources	and health by providing access to biomedical and genomic	Bookshelf
Chemicals & Bioassays	information.	PubMed Central
Data & Software	About the NCBI Mission Organization Research NCBI News	PubMed Health
DNA & RNA		BLAST

b) Copy the FASTA format of protein sequence.

S NCBI	Resources 🖂	How To 🗹
Protein		Protein Advanced
<u>Display Se</u>	ettings: 🕑 FAS	TA
nucleo	ocapsid p	orotein [Human coronavirus 229E]
GenBank:	AAA45463.1	
<u>GenPept</u>	<u>Graphics</u>	
>gi 3295 MATVKWAD RVDLSPKL EEPDSRAP PKPSRNQS ANGVKAKG QPLLNPSA	77 gb AAA45 ASEPQRGRQGR HFYYLGTGPHK SRSQSRSQSRG PASSQTSAKSL YPQFAELVPST LEFNPSQTSPA	463.1 nucleocapsid protein [Human coronavirus 229E] IPYSLYSPLLVDSEQSWKVIPRNLVPINKKDKNKLIGYWNVQKRFRTRKGK DAKFRERVEGVVWVAVDGAKTEPTGHGARRKNSEPEIPHFNQKLPNGVTVV PGESKPQSRNPSSDRYHNSQDDIMKAVAAALKSLGFDKPQEKDKKSAKTGT ARSQSSETKEQKHEIEKPRWKRQPNDDVTSNVTQCFGPRDLDHNFGSAGVV AAMLFDSHIVSKESGNTVVLTFTTRVTVPKDHPHLGKFLEELNAFTREMQQ IAEPVRDEFSIETDIIDEVN

Step-2: Multiple Sequence Alignment

The second step in phylogenetic analysis is to construct sequence alignment. This is probably the most critical step in the procedure because it establishes positional correspondence in evolution. Only the correct alignment produces correct phylogenetic inference because aligned positions are assumed to be genealogically related. Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree. For that reason, it is essential that the sequences are correctly aligned.

For Multiple Sequence Alignment the program "MEGA" was used. The software MEGA (MOLECULAR EVOLUTIONARY GENETICS ANALYSIS- Windows v5.2.2) was downloaded from <u>http://www.megasoftware.net/</u> and installed in the Windows 7 environment. For aligning the sequences, two alignment methods are provided by MEGA:

✓ ClustalW

✓ MUSCLE

MUSCLE was chosen instead of ClustalW for current study, because ClustalW has some major pitfalls-

- ClustalW is not suitable for comparing sequences of different lengths because it is a global alignment-based method.
- > The final alignment result is influenced by the order of sequence addition.
- Any errors made in the early steps of alignment cannot be corrected. This problem of "once an error, always an error" can propagate throughout the entire alignment.

A flowchart of the procedure is given below:



Open A File/Session Open a Recently Used File	e 🕨	L Distance	 • Diversity	• 또 Phylogeny •	User Tree	প Ancestors *	Selection *	₩ Rates •	() Clocks	- Of Links
≝ Edit a Text File ம் Convert File Format to M	EGA						·			
Printer Setup										
Exit MEGA	Alt+X									

M Open a File	and the second				X	
Look	in: 🕕 New folder		▼ ⇐ €) 💣 🔳 🕇		
œ	Name	•		Date mo	odified	
	🖪 1.NC protei	n multi.fasta		26-Dec-	13 12:42 PM	
Recent Places	3. SM prote	in multi.fasta		26-Dec-	13 3:58 PM	
Desktop						
Libraries						
Computer						
Network	•	111			•	
	File name:	1.NC protein multi.fasta		-	Open	
	Files of type:			-	Cancel	
		,				
					Help	
IEGA 5.2.2 (5130611)	-					
Analysis Help						
≣ - ∦ 10 ign - ∛ Data - M	odels Distance	<u>⊼i</u> ▼	(福) 양 User Tree Č Ancestors	- Selection	Rates Clocks	• • Cinks •
		Analyze or Align File?		X		
		How would you li	ke to open this MAS file?			
		Analy	ze Align 🗲			

💪 Examples

Tutorial

First time User?

MEGA release #5130611

 Citation E Report a Bug Updates?

Customize ToolBar 👻

Preferences

-

M M5: Alignment Evplorer (1 NC protein p (1 firsta)	~
M M3. Alghment explorer (Live plotenty diasta)	X
Data Edit Search Alignment We Sequencer Display Help	
D 😅 🖬 Ψ : 🧮 🌚 🗰 W 👽 💥 🐩	🌼 🐴
Protein Sequences Align Protein	
Species/Abbrv	
1. gi 329577 gb AAA45463.1 _nucleocapsid_protein_Human_coronavirus_229E A VK ADASERORGROGRISYSLYSELLYSE LLYSE	SW <mark>kvi</mark> e
2. gi 112785273 gb ABI20791.1 _nucleocapsid_protein_Human_coronavirus_NL63MANVNWADDRAARKKFPPPSFYMPLLVS5DKAPY	V I P R N
3. gi 85816020 gb ABC84375.1 _nucleocapsid_protein_Porcine_epidemic_diarrh <mark>MASVSFQDRGRKRVPLSLYAPL8LYA</mark> VTNDKFLSKVL	A N N <mark>A V E</mark>
4. gi 77021885 gb ABA60797.1 _nucleocapsid_protein_Human_coronavirus_HKU1 RGRKTOPKFTVSTOPOGNTIPHYSTF5GITOFOK	<mark>g r d</mark> f <mark>k</mark> f
5. gi 344332714 gb AEN19370.1 _nucleocapsid_protein_Human_coronavirus_OC43 MSFTPGKQSSSRASSGNRSGNGILKNAADQSDQFR	V Q T <mark>R</mark>
6. gi 520688469 gb AGP04915.1 _nucleocapsid_protein_Pipistrellus_bat_coronMATPAPPRAVVFANDNETPTNNORSGRERTKFRP	A P N T T V
7. gi 211907046 gb ACJ12048.1 _nucleocapsid_phosphoprotein_Bulbul_coronaviMSAPVVPTTDASWFQVLKAQNKKATHEKFKGDGV	INSAI
8. gi 212681394 ref YP_002308510.1 _nucleocapsid_phosphoprotein_Munia_coromATPAVPTTDASWFQVL4AQWKKAIHFQF%GWGW	VNSAI
9. gi 187251967 ref YP_001876448.1 _nucleocapsid_protein_Beluga_Whale_coro <mark>MASISGKGKNPADKSVKFQAPKVGRQAPALPIS</mark> G	¶ <mark>V</mark> SW <mark>F</mark> Ç
10. gi 335211 gb AAA47915.1 _nucleocapsid_protein_Transmissible_gastroenteMANQGQRVSWGDESIKIRGRSWSBGEKWWWIFLS	FFN <mark>P</mark> I1

M5: MUSCLE - AppLink	
Option	Selection
Presets	None
Gap Penalties	
Gap Open	-2.9
Gap Extend	0
Hydrophobicity Multiplier	1.2
Memory/Iterations	
Max Memory in MB	959
Max Iterations	8
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMA
Clustering Method (Other Iteration	UPGMA
Min Diag Length (lambda)	24 📫
Genetic Code (when using cDNA)	Standard
Alignment Info	MUSCLE Citation: Edgar, Robert C. (2004), MUSCLE:
	multiple sequence alignment with high accuracy and
	high throughput, Nucleic Acids Research 32(5), 1792-1797.
	Compute X Cancel ? Help Restore Defaults

Data Edit Search Alignment	Web Sequencer Display Help		
Create New Copen Peopen] 🕫 🚡 💼 🗙 💥 🍓	28 4 D A # # #
Close		*	* * *
Phylogenetic Analysis	D K A BYR VI PRNLVPICKC - NK D E D ICYNNVCK D K A BYR VI PRNLVPICKC - NK D E D ICYNNVCK	R FRIRKGKRVDLSPKLHFYYLG R WRMRRGORVDLPPKVHFYYLG	I G P H K D A K F R E I G P H K D L K F R O
Save Session Ctrl+S	DKPLSKVLANNAVPINKG – NKDQQIGYWNEQI	R WRMRRGERIEQPSNWHFYYLG	I G P H A D L R Y R I
🖺 Export Alignment	KGROFKFSOGOGVPIAFGVPPSEAKGYWYRHS	R R S F K I A D G Q Q K Q L L P R W Y F Y Y L G	T <mark>GP</mark> YANAS <mark>YGE</mark>
DNA Sequences Protein Sequences 		RKOFKIA GOVELLEKWYFYYLG RK - INTGNGTK - PLAPRWYFYYLG R OXBGGT OTBDSYAFYYTG	TGPEANLPFR5 TGPEANLPFR5
땨 Translate/Untranslate 瓋 Select Genetic Code Table	KAIHPOFKONOVPVNSAIKPALNHOVHLKFA Ageogi Fkorovevnevvsknonhovnig	R Q K P G G A Q I P P S Y A F Y Y F G R I I P G G K K L A P L Y Y F Y Y I G	T G P R G N L K Y G E Y P P N D S T G P R D Q A K Y G Q
I Reverse Complement Reverse Complement	GSKFNNLCORDEV DRGIG- NRODOIGYNNROD	R YRMVKGQRKCLFRWFFYYLG	geh a dak fko
Exit Alignment Explorer			

<u>Result</u>

a) Multiple Sequence Alignment for Nucleocapsid protein

M5: Alignment Explorer (1.NC protein multi.fasta)	
Data Edit Search Alignment Web Sequencer Display Help	
] D 😅 🖬 📽 🗮 🌚 🎆 W 😏 💥 🏗	- 🗠 🖻 🐰 🛍 🗙 💥 糩 🚵 💶 🕨] 🗛 🦊 🗛 🦂
Protein Sequences	
Species/Abbrv *	
1. gi 329577 gb <mark>3LYSPLLVDSEQS</mark> -WKVI PRHLVPINKK-DKNKLIGYWWVQKR	FRIRKGKRVDLSPKLHFYYLGIGPHKDAKFRDRV
2. gi 112785273 <mark>5fymp</mark> llv s5dkapyr vI prnlvpigrg-nkdeqigynnvqer	NRMRRGQR <mark>VDLPPK</mark> VHFYYLG T GPHKDLKFRQRS
3. gi 85816020 gBLYAPLRVINDKPLSKVLANNAVPINKG-NKDQQIGYNNZQIR	WRMRRGERIEORSNWHFYYLGEGPHADLRYRTRT
6. gi 152068846915WFTGLT-OHGKOPLAFPPGGGVPLNANSTPAONAGYWRRODRK-	INTENETR - PLAPENY FYYTE TEPEANLPFESVK
7. gi 211907046 SWFQVLKAQNKKATHPKFKGDGVPINSAIKPAENHGYWLXYWR	- QK PGGT QIPPSYAFYYTGTGPRGNLKYGEYPPNON I ET
8. gi 212681394 <mark>50 FQVLKAQNKK</mark> AI <mark>HFQ</mark> FKGNGVPVNSAIKBAENHGYULRFAR	- QKPGGA QIPPSYAFYYTGTGPRGNLKYGEYPPNDSKET
9. gi 187251967 <mark>5NFQ</mark> AI <mark>KGRKAGEEGIIFKGPGVPVNENVSKNQNHGYNIQRSR</mark>	- IIPGGKKLAPLYYFYYIGIGPRDQAXYGQBI
10. gi 335211 gb <mark>sffnp</mark> IILQQGsKfWnLCPRDFVPKGIG-NRDQQIGYNNRQIR	YRMVKGORKELPERNFFYYLGIGPHADAKFKDKL
	٩
Site # 536 🔹 🔍 🗰 with 🔿 w/o Gap:	



b) Multiple Sequence Alignment for Small Membrane protein

Step-3: Determining a tree building method

There are currently two main categories of tree-building methods, each having own advantages and limitations.

The first category of phylogenetic methods is based on distance, which is the amount of dissimilarity between pairs of sequences, computed on the basis of sequence alignment. The distance-based methods assume that all sequences involved are homologous and that tree branches are additive, meaning that the distance between two taxa equals the sum of all branch lengths connecting them.

The algorithms based on distance matrix include-

• Unweighted pair group method using arithmetic average (UPGMA) and

• Neighbor joining (NJ)

The second category is based on discrete characters, which are molecular sequences from individual taxa. The basic assumption is that characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved. Therefore, the character states of the common ancestor can be traced from this dataset. Another assumption is that each character evolves independently and is therefore treated as an individual evolutionary unit.

The two most popular character-based approaches are-

- Maximum parsimony (MP) and
- Maximum likelihood (ML)

In the present study the **"Maximum likelihood (ML)"** method was chosen to construct the phylogenetic tree. This is because ML has some advantages over other tree-building methods.

Drawbacks of various tree-building methods:

- The drawback of all distance-based methods is that the actual sequence information is lost when all the sequence variation is reduced to a single value.
- The basic assumption of the UPGMA method is that all taxa evolve at a constant rate and that they are equally distant from the root, implying that a molecular clock is in effect. However, real data rarely meet this assumption. Thus, UPGMA often produces erroneous tree topologies.
- One of the disadvantages of the NJ method is that it generates only one tree and does not test other possible tree topologies. This can be problematic because, in many cases, in the initial step of NJ, there may be more than one equally close pair of neighbors to join, leading to multiple trees. Ignoring these multiple options may yield a suboptimal tree.
- The Maximum Parsimony method chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths. It tends to produce more accurate trees than the distance-based methods when sequence divergence is low. However, when sequence divergence is high, or the amount of homoplasies is large, tree estimation by MP can be less effective, because the original parsimony assumption no longer holds.
- Estimation of branch lengths may also be erroneous because MP does not employ substitution models to correct for multiple substitutions. This drawback can become prominent when dealing with divergent sequences.
- In addition, MP only considers informative sites, and ignores other sites. Consequently, certain phylogenetic signals may be lost.
- MP is also slow compared to the distance methods, and is very sensitive to the "long-branch attraction" (LBA) artifacts.

Advantages of Maximum Likelihood

- ✓ Maximum Likelihood (ML) uses probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data. It finds a tree that most likely reflects the actual evolutionary process.
- ✓ ML is based on well-founded statistics instead of a medieval philosophy. It is thus considered mathematically more rigorous than MP. In fact, it is the most rigorous among all approaches.

- ✓ ML is an exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites and therefore may be more robust.
- ✓ ML employs substitution models and is not sensitive to LBA.

A flowchart of the procedure is given below:



File Analysis Help
E TAign Data Models Distance Diversity Phylogeny User Tree Ancestors Selection Rates Clocks Links
ImL Construct/Test Maximum Likelihood Tree Image: Construct/Test Neighbor-Joining Tree Image: Construct/Test Neighbor-Joining Tree Image: Construct/Test Minimum-Evolution Tree Image: Construct/Test UPGMA Tree Image: Construct/Test Maximum Parsimony Tree(s) Image: Open Tree Session
First time User? Tutorial Examples Citation Report a Bug Updates? Customize ToolBar - Preferences -

🕅 Choose a Data F	ile to Analyze				X
Look in:	👢 New folder		• E (*	
œ.	Name	A		Date mod	dified
	🐘 1.NC protein r	nulti.mas		13-Jan-14	4 2:20 AM
	🖳 2. SM protein	multi.mas		13-Jan-14	4 2:22 AM
Desktop					
Libraries					
Computer					
Network	•	111			4
	File name:	1.NC protein multi.mas		•	Open
	Files of type:	All files (*.*)		•	Cancel
					Help

Options Summary		
Option	Selection	
Analysis	Phylogeny Reconstruction	
Statistical Method	Maximum Likelihood	
Phylogeny Test		
Test of Phylogeny	None	
No. of Bootstrap Replications	Not Applicable	
Substitution Model		
Substitutions Type	Amino acid	
Model/Method	Jones-Taylor-Thornton (JTT) model	
Rates and Patterns		
Rates among Sites	Uniform rates	
No of Discrete Gamma Categories	Not Applicable	
Data Subset to Use		
Gaps/Missing Data Treatment	Complete deletion	
Site Coverage Outoff (%)	Not Applicable	
Tree Inference Options		
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)	
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)	
Initial Tree File	NatApplicable	
Branch Swap Filter	Very Strong	
System Resource Usage		
Number of Threads	1	

<u>Result</u>

a) Phylogenetic tree based on Nucleocapsid protein sequences



b) Phylogenetic tree based on Small membrane protein sequences



<u>Step-4: Evaluation of phylogenetic tree</u>

After phylogenetic tree construction, the next step is to statistically evaluate the reliability of the inferred phylogeny. There is a question that needs to be addressed i.e., how reliable the tree or a portion of the tree is. To answer this question, we need to use analytical resampling strategies such as bootstrapping or jackknifing, which repeatedly resample data from the original dataset.

Bootstrapping is a statistical technique that tests the sampling errors of a phylogenetic tree by repeatedly sampling trees through slightly perturbed datasets. The rationale for bootstrapping is that a newly constructed tree is possibly biased owing to incorrect alignment or chance fluctuations of distance measurements. To determine the robustness or reproducibility of the current tree, trees are repeatedly constructed with slightly perturbed alignments that have some random fluctuations introduced. A truly robust phylogenetic relationship should have enough characters to support the relationship even if the dataset is perturbed in such a way. Otherwise, the noise introduced in the resampling process is sufficient to generate different trees, indicating that the original topology may be derived from weak phylogenetic signals. By doing so, the robustness of the original tree can be assessed.

For statistical testing 'Bootstrap method' was selected from 'Phylogeny Test' option and clicked on 'Compute'.

Options Summary	
Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Jones-Taylor-Thornton (JTT) model
Rates and Patterns	
Rates among Sites	Uniform rates
No of Discrete Gamma Categories	Not Applicable
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
nitial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	Not Applicable
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

15: Progress					
Progress	25%				
Details -					
Status / Options					
Run Status					
Start time	13-Jan-142:55:03 AM				
Status	Bootstrapping ML tree				
Log Likelihood	-1590.1123				
Analysis Optior	8				
Analysis			*		
Analysis Statistic Phylogeny 3	cal Method	Phylogeny Reconstruction Maximum Likelihood	=		
Test of Phylogeny No. of Bootstrap Replications		Bootstrap method 500			
Substitutio	on Model	Amino acid			
Model/Met model	chod	Jones-Taylor-Thornton (JTT)	Ŧ		

<u>Result</u>

Result obtained from bootstrapping for both Nucleocapsid protein and Small membrane protein was the same as the original result, which supports the reliability and robustness of the constructed phylogenetic trees.





Result Interpretation

For interpretation one phylogenetic tree is taken for consideration (as both trees exhibit almost similar kind of phylogenetic relation among different speceis of Coronaviridae).



The phylogenetic tree is broken down into nodes and branches.

- The branches are horizonal lines that represent evolutionary lineages changing over time. The longer the branch in the horizonal dimension, the larger the amount of change. The bar at the bottom of the figure provides a scale for this.
- The nodes are represented as circles. There are two types of nodes;
 - external nodes, also called 'leaves' and
 - ➢ internal nodes

The leaves are shown here with green circles and these represent the virus protein sequences taken as samples.

The internal nodes are represented by blue circles and these represent putative ancestors for the sampled viruses.

The numbers next to each node represent a measure of support for the node computed by bootstrapping. A high value (e.g. 100) means that there is strong evidence that the sequences to the right of the node cluster together to the exclusion of any other.

For the convenience of discussion the nodes are termed by alphabets like A,B,C,D,E etc.

The tree is rooted which suggests an ultimate common ancestor (A) of all the sampled viruses (represented as red circle).

According to the inferred phylogenetic tree based on protein sequences from different species of Coronaviridae -

- ✓ Two broad subgroups (B and C) have descended from a common **ancestor A**.
- ✓ In the subgroup B, Bulbul coronavirus HKU11 and Munia coronavirus HKU13 are the closely related groups which are related to Beluga whale coronavirus SW1 and they are descended from the ancestor F. The group F is related to another group E that includes two closely related virus species, Human coronavirus OC43 and Human coronavirus HKU1. The groups F and E are descendants of D which is descended from B. the group B gives rise to an outgroup, Pipistrellus bat coronavirus HKU5, which is more close to group E than F.
- ✓ In the subgroup C, Human coronavirus 229E and Human coronavirus NL63 are the closely related groups which are related to Porcine epidemic diarrhea virus and they are descended from the ancestor H. The group H is descended from the ancestor C which gives rise to an outgroup, Transmissible gastroenteritis virus.