

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.

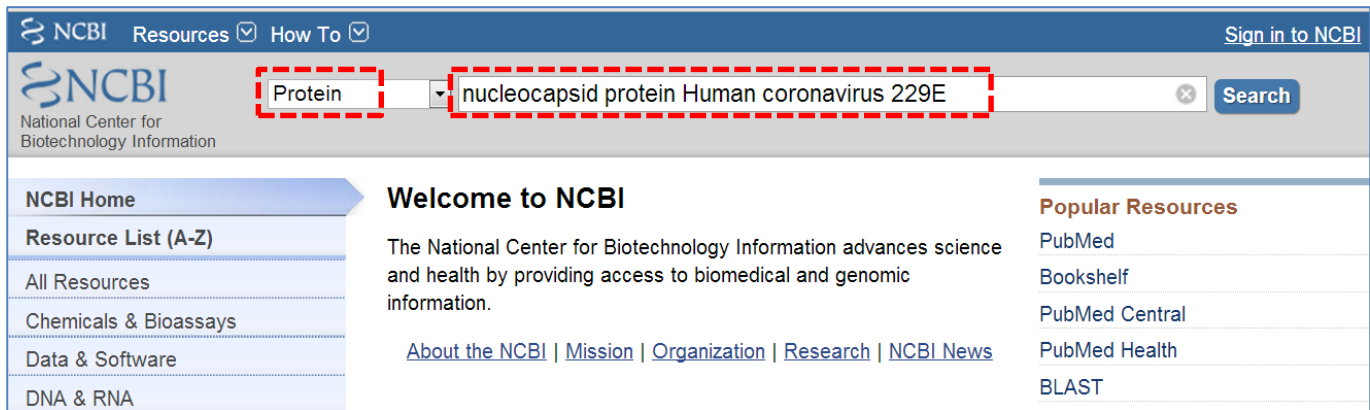
In many cases, protein sequences are preferable to nucleotide sequences because protein sequences are relatively more conserved as a result of the degeneracy of the genetic code in which sixty-one codons encode for twenty amino acids, meaning thereby a change in a codon may not result in a change in 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:

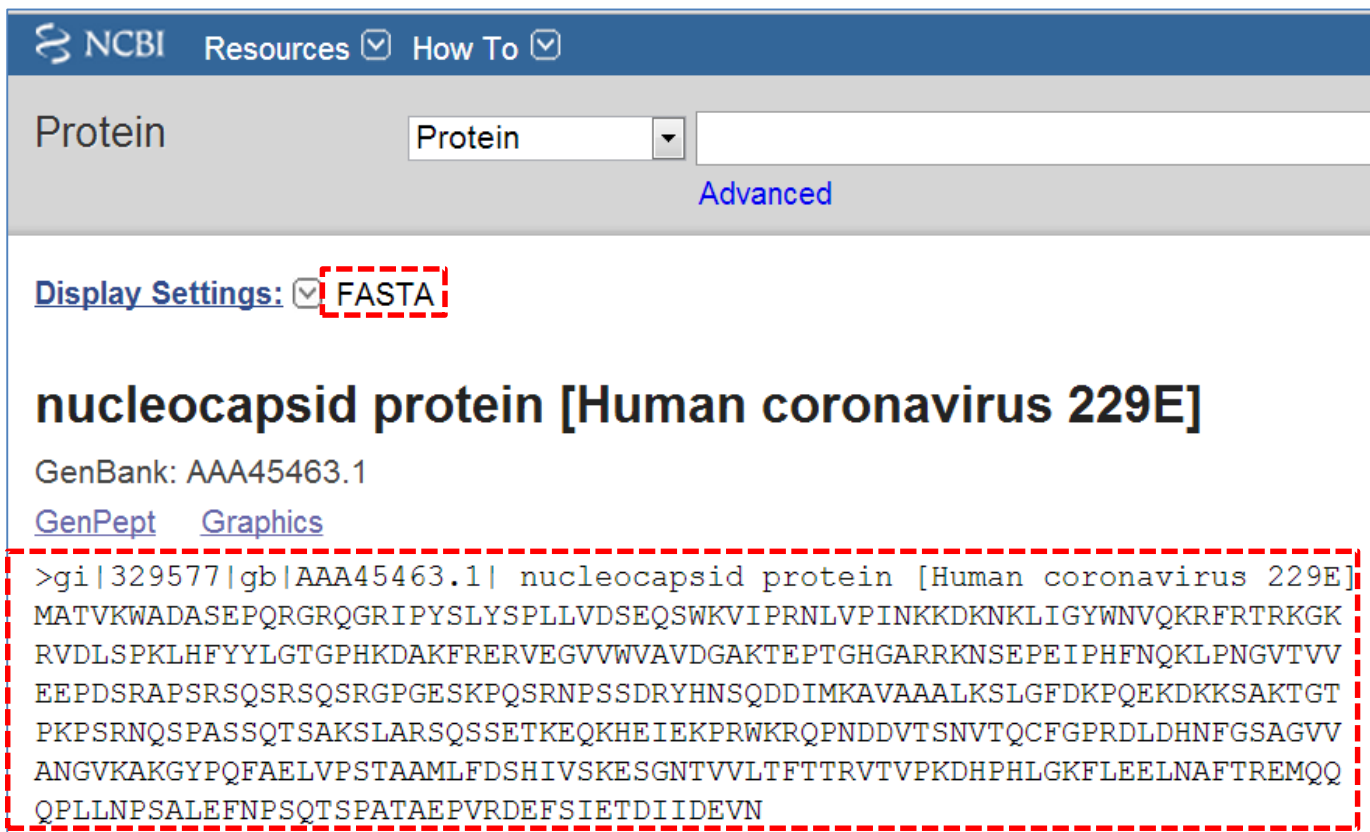
- A) Nucleocapsid (N) protein: (Phosphoprotein)**
- B) Spike (S) protein: (Glycoprotein)**
- C) Haemagglutinin-esterase (HE) protein**
- D) Integral membrane (M) protein**
- E) Small membrane (E) protein**

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.



- b) Copy the FASTA format of protein sequence.



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.

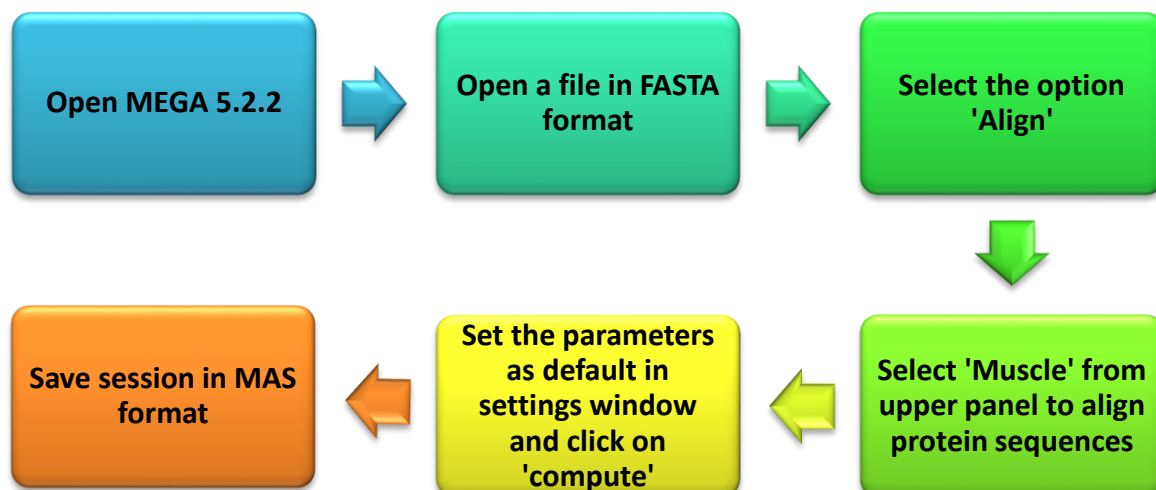
For Multiple Sequence Alignment the program “MEGA” was used. The software MEGA (MOLECULAR EVOLUTIONARY GENETICS ANALYSIS- Windows v5.2.2) was downloaded from <http://www.megasoftware.net/> and installed in the Windows 7 environment. For aligning the sequences, two alignment methods are provided by MEGA:

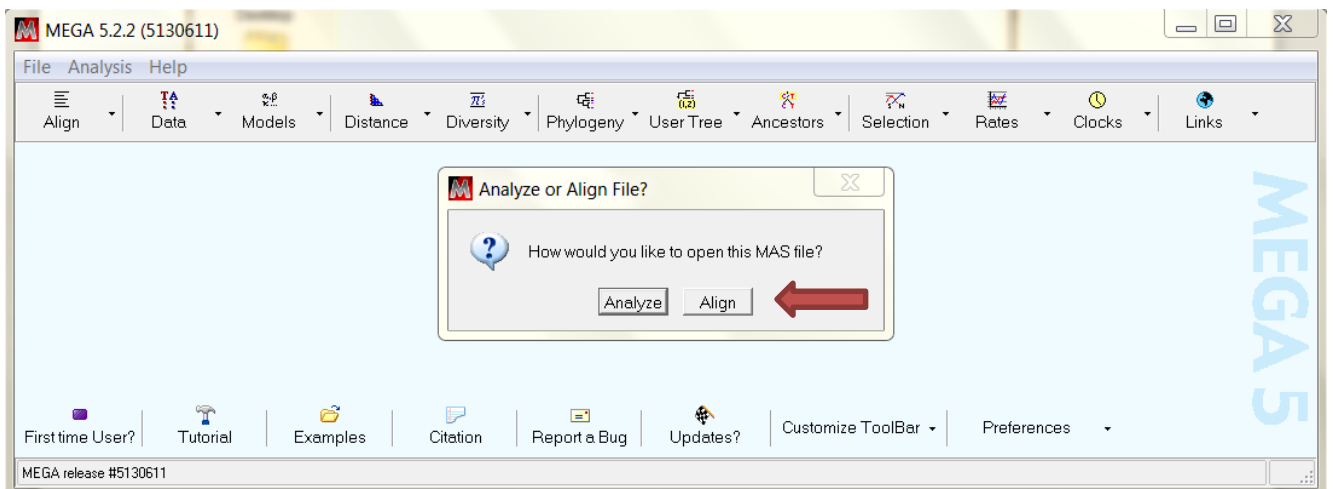
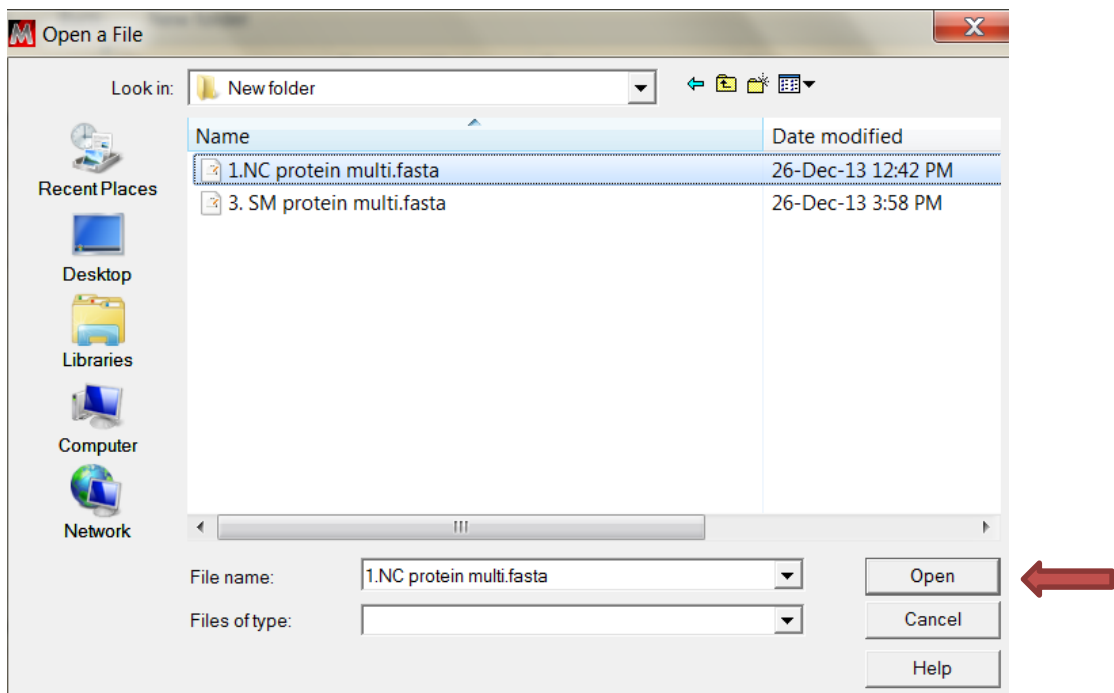
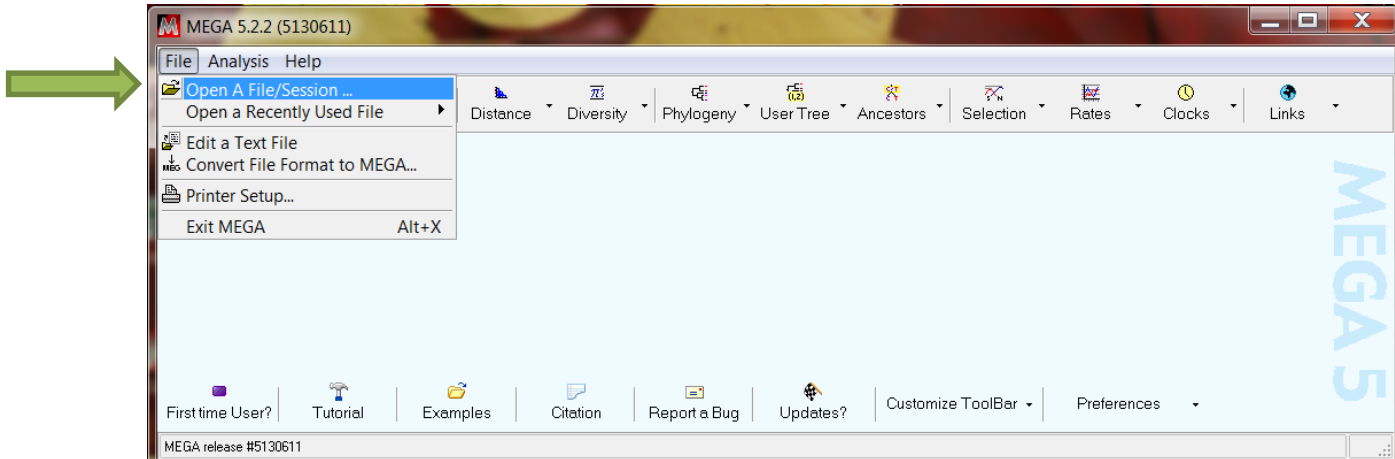
- ✓ **ClustalW**
- ✓ **MUSCLE**

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:





M5: Alignment Explorer (1.NC protein p...i.fasta)

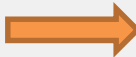
Data Edit Search Alignment Web Sequencer Display Help

Protein Sequences Align Protein

Species/Abbrv	
1. gi 329577 gb AAA45463.1 _nucleocapsid_protein_Human_coronavirus_229E	MAV...ADAA...SGR...G...I...Y...L...S...L...V...D...E...L...V...I...
2. gi 112785273 gb ABI20791.1 _nucleocapsid_protein_Human_coronavirus_NL63	MAV...ADD...AAK...F...S...F...M...L...L...V...D...A...Y...V...I...R...
3. gi 85816020 gb ABC84375.1 _nucleocapsid_protein_Porcine_epidemic_diarrh	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
4. gi 77021885 gb ABA60797.1 _nucleocapsid_protein_Human_coronavirus_HKU1	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
5. gi 344332714 gb AEN19370.1 _nucleocapsid_protein_Human_coronavirus_OC43	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
6. gi 520688469 gb AGP04915.1 _nucleocapsid_protein_Pipistrellus_bat_coron	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
7. gi 211907046 gb ACJ12048.1 _nucleocapsid_phosphoprotein_Bulbul_coronavi	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
8. gi 212681394 ref YP_002308510.1 _nucleocapsid_phosphoprotein_Munia_coro	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
9. gi 187251967 ref YP_001876448.1 _nucleocapsid_protein_Beluga_Whale_coro	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...
10. gi 335211 gb AAA47915.1 _nucleocapsid_protein_Transmissible_gastroente	MAV...F...D...G...K...V...L...L...A...L...V...L...L...V...L...A...V...A...V...

M5: MUSCLE - AppLink

Option	Selection
<input type="checkbox"/> 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
<input type="checkbox"/> 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.



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:

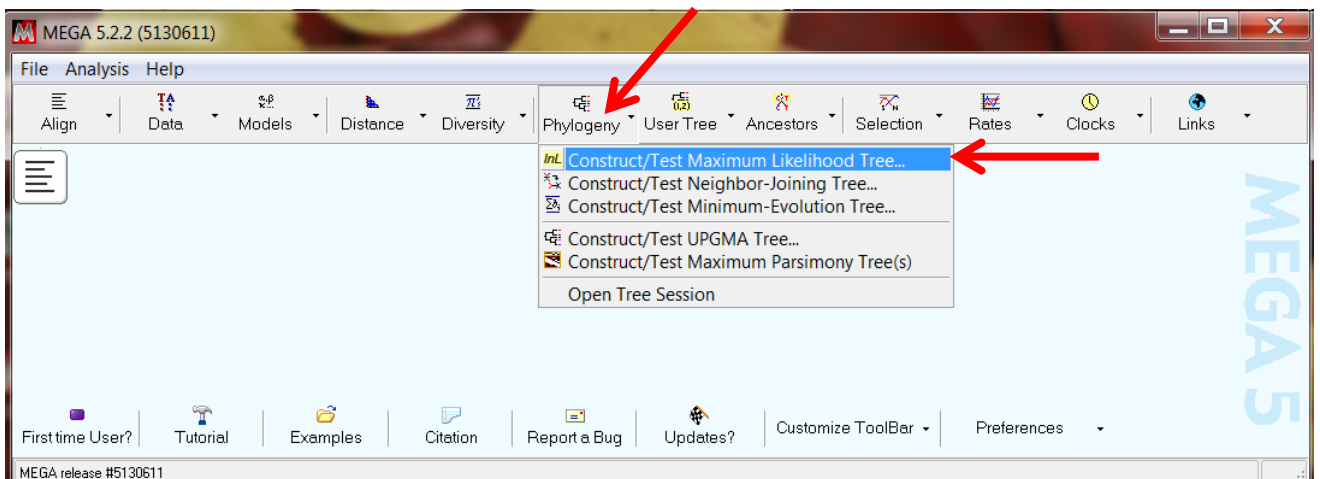
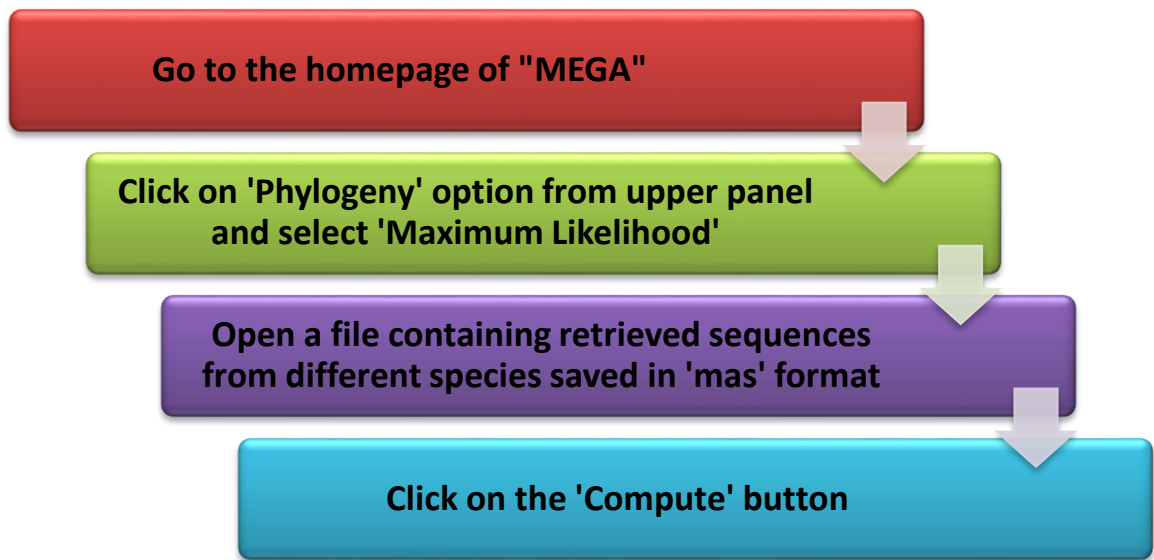
- The drawback of **all distance-based methods** is that the actual sequence information is lost when all the sequence variation is reduced to a single value.
- The basic assumption of the **UPGMA** method is that all taxa evolve at a constant rate and that they are equally distant from the root, implying 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.

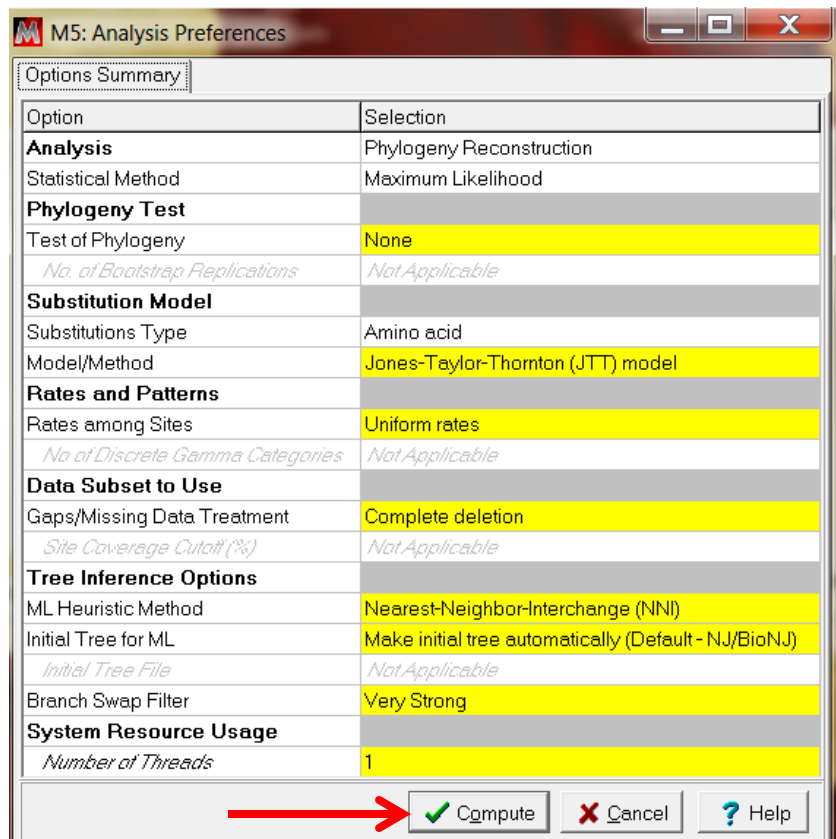
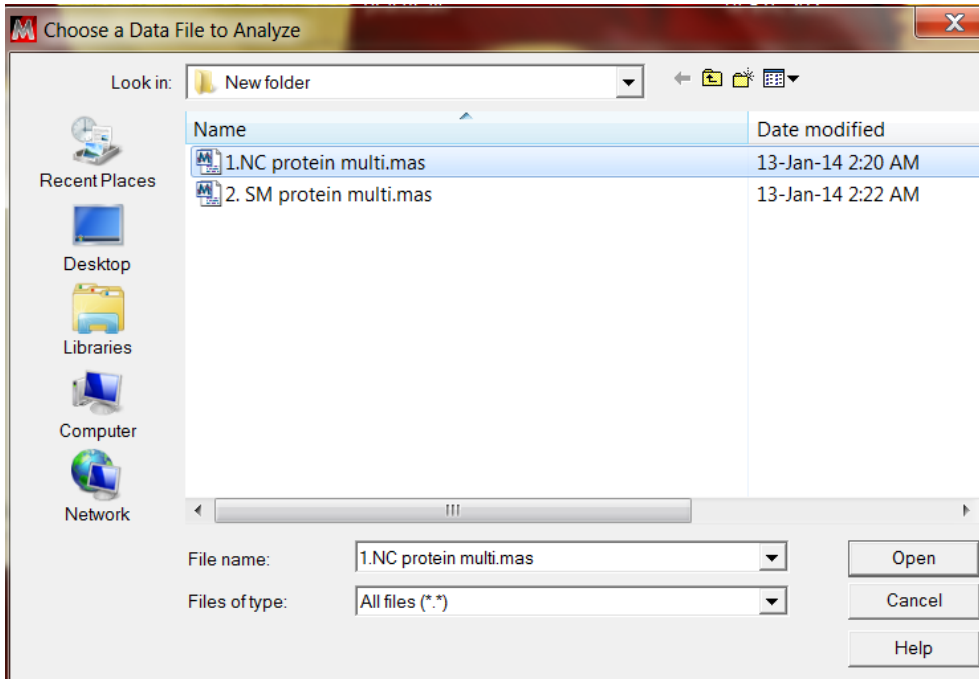
Advantages of Maximum Likelihood

- ✓ 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.

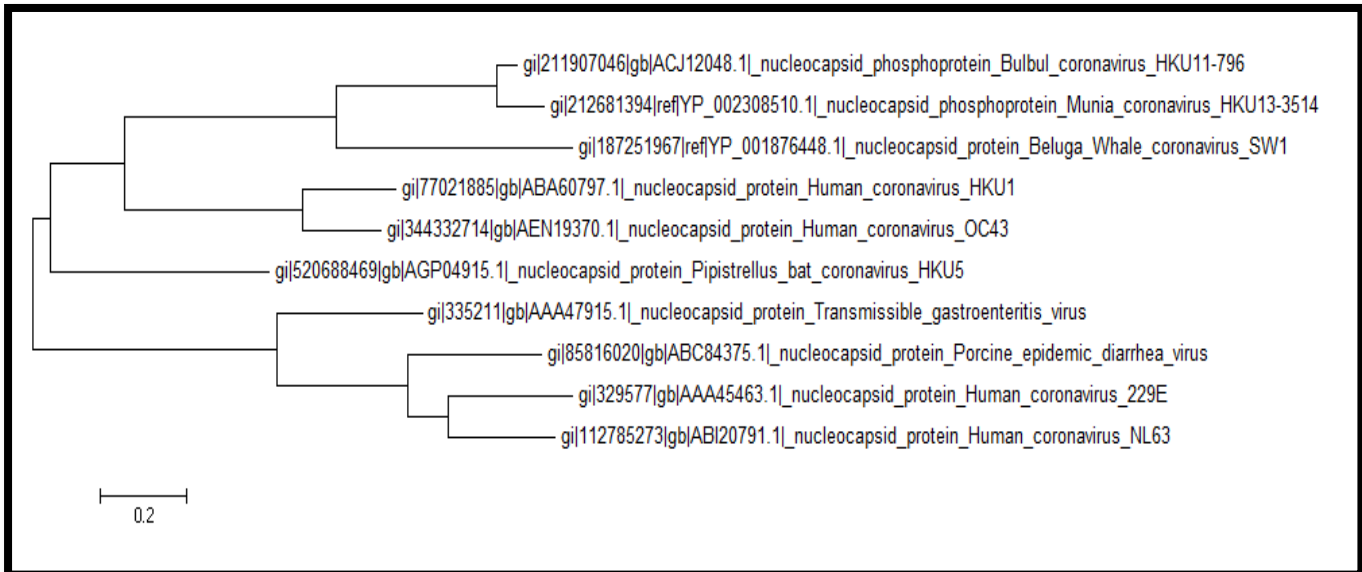
A flowchart of the procedure is given below:



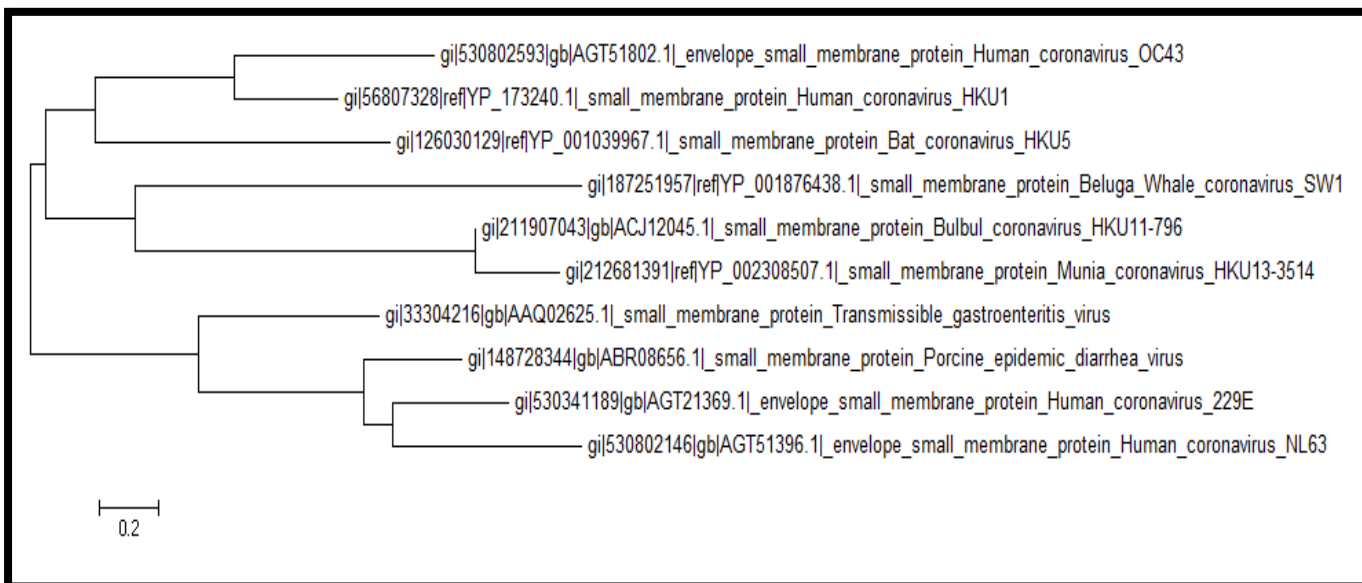


Result

a) Phylogenetic tree based on Nucleocapsid protein sequences



b) Phylogenetic tree based on Small membrane protein sequences

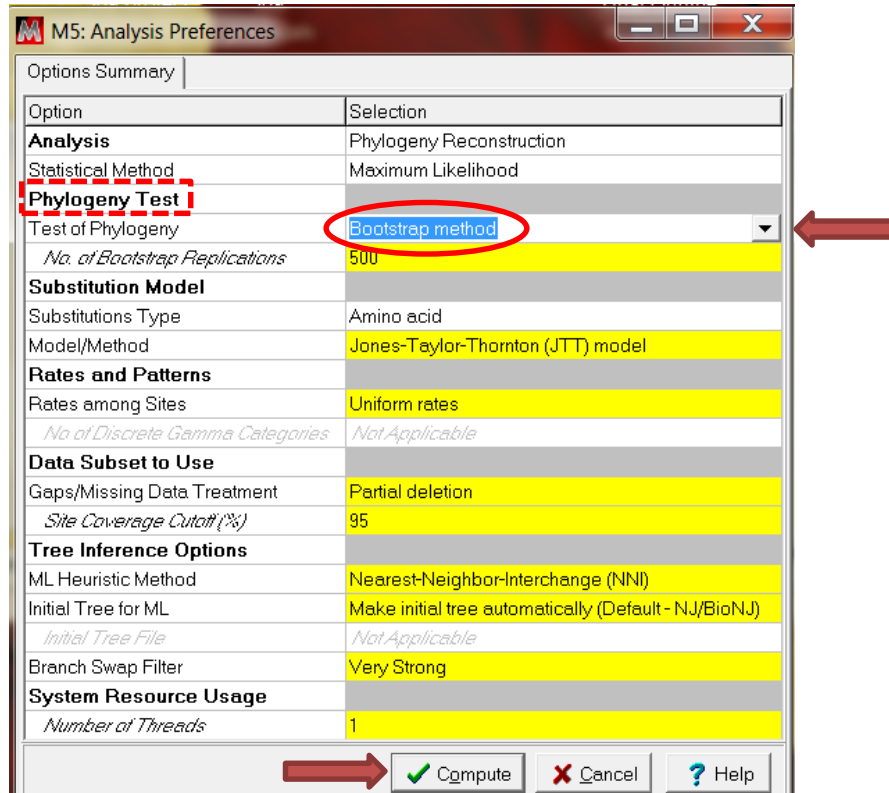


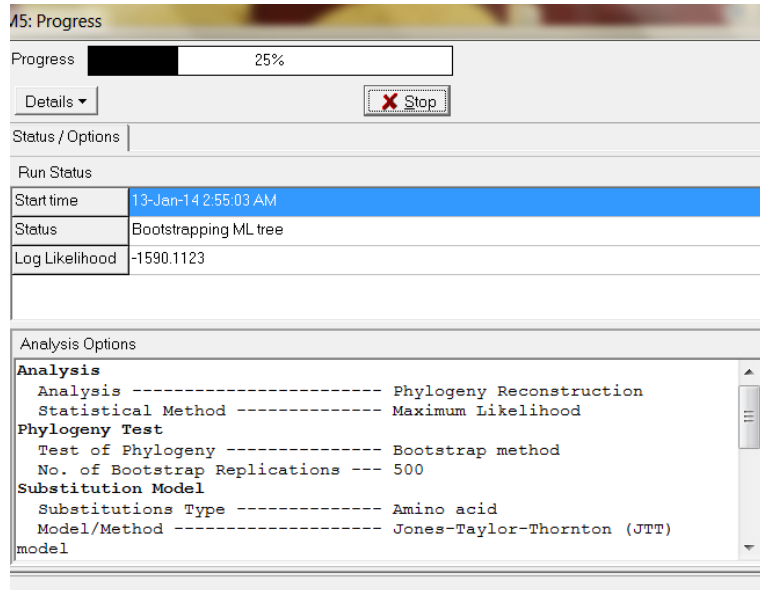
Step-4: Evaluation of phylogenetic tree

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.

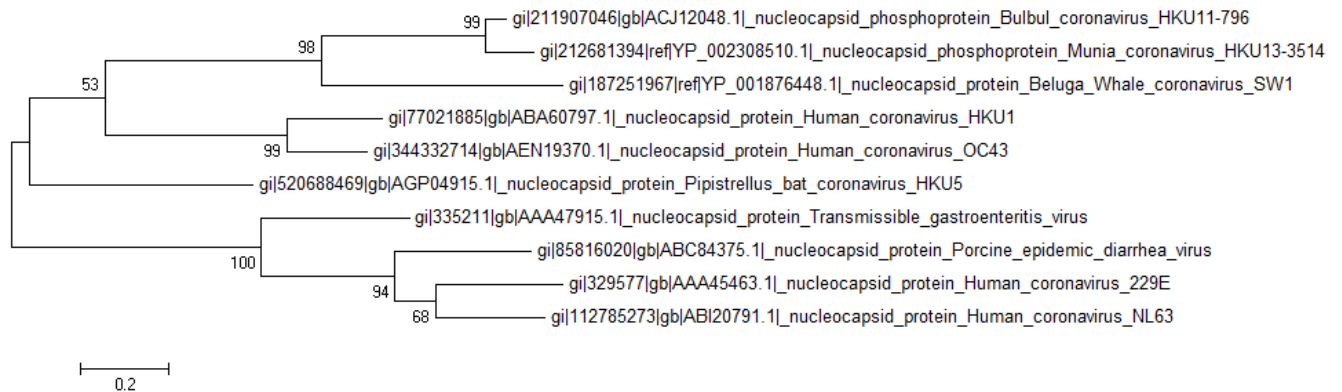
For statistical testing 'Bootstrap method' was selected from 'Phylogeny Test' option and clicked on 'Compute'.





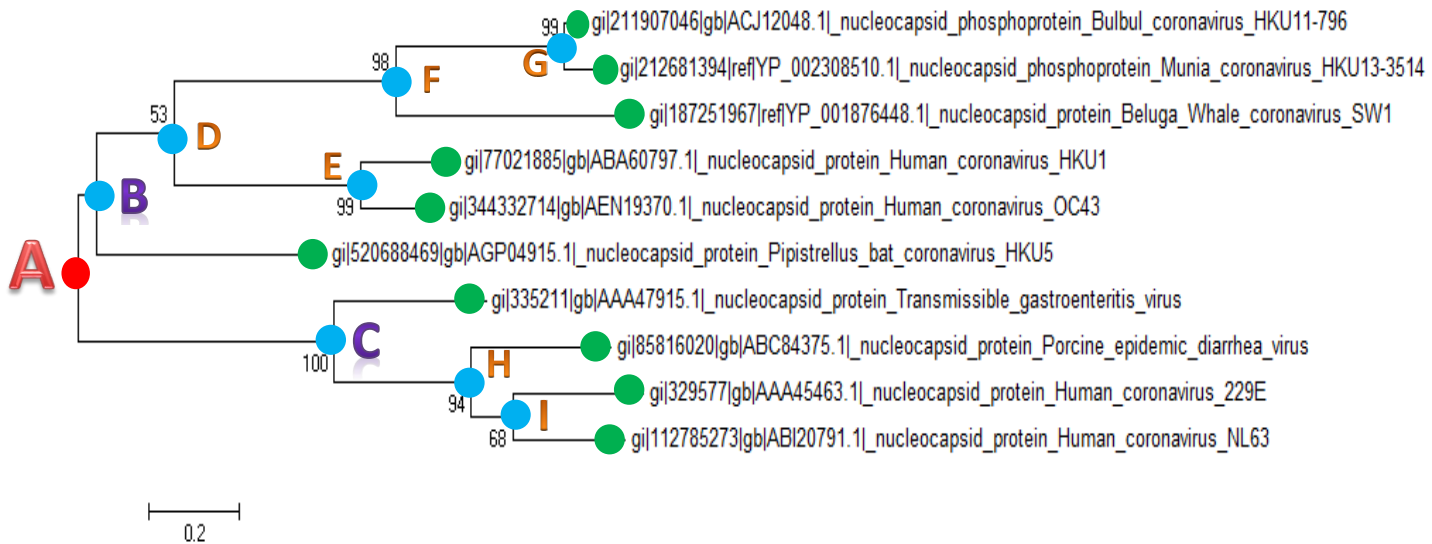
Result

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 horizontal lines that represent evolutionary lineages changing over time. The longer the branch in the horizontal 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.