A study of phylogenetic analysis algorithms and methods using MEGA X

phylogenetic analysis:

It is a study of evolutionary relationships between organisms of different species that are originated or evolved from a common root species. Such relationships are commonly represented as a ‘phylogenetic tree’, a tree diagram whose root represents the ancestral parent of the organisms, nodes of the tree represent the time/point in evolutionary history where a species is branched/separated into two different species.

In this document it is attempted to give a clear overview of the steps involved in building a phylogenetic tree with molecular data as input. The input section of this document will talk in detail about the various types of data that can be used as an input for performing a phylogenetic analysis, including molecular data and morphological data.

Softwares of phylogenetic analysis:

A huge number of free softwares can be used to perform various [1] steps of the phylogenetic analysis. Each software is designed, tailored and used for different specific purposes, but the algorithms used repeat in almost all of the softwares. The variation is seen in the prepossessing of the input data, choice of algorithm for analysis, specific operations like visualization, aligning input data, choice of models etc.

Some of the most notable open source softwares are:

- Mr Bayes – Primarily based on Bayesian inference
- MEGA X – Good for maximum parsimony, Distance based, maximum likelihood methods.
- SeaView – Multiple sequence alignment editor for molecular data based phylogenetic analysis
- PHYLIP - Good for maximum parsimony, Distance based, maximum likelihood methods.
- PAUP - Good for primarily parsimony based algorithms and also includes Distance based, maximum likelihood methods.
-

The above mentioned softwares are just a small sample of many softwares that are used for generating a phylogenetic tree. In this document, the software MEGA-X and the pipeline of generating a phylogenetic tree is explored.

With different combinations of various prepossessing steps, various analysis algorithms, types of trees etc, there are hundreds of variations of pipelines that can be performed in this one software alone. It is impossible to go through all the possible ways this software can be used in this document, hence it is confined to a most commonly performed pipeline for phylogenetic analysis.

Fig 1 shows the simplistic version of the phylogenetic analysis pipeline which addresses all the steps involved in a simple example task. Fig 2 is a more detailed version of the same pipeline, showing all the possible options that can be chosen at each step. Each choice will lead to a different/unique pipeline, so, the choices that led to the pipeline shown in the fig 2 are highlighted in green.

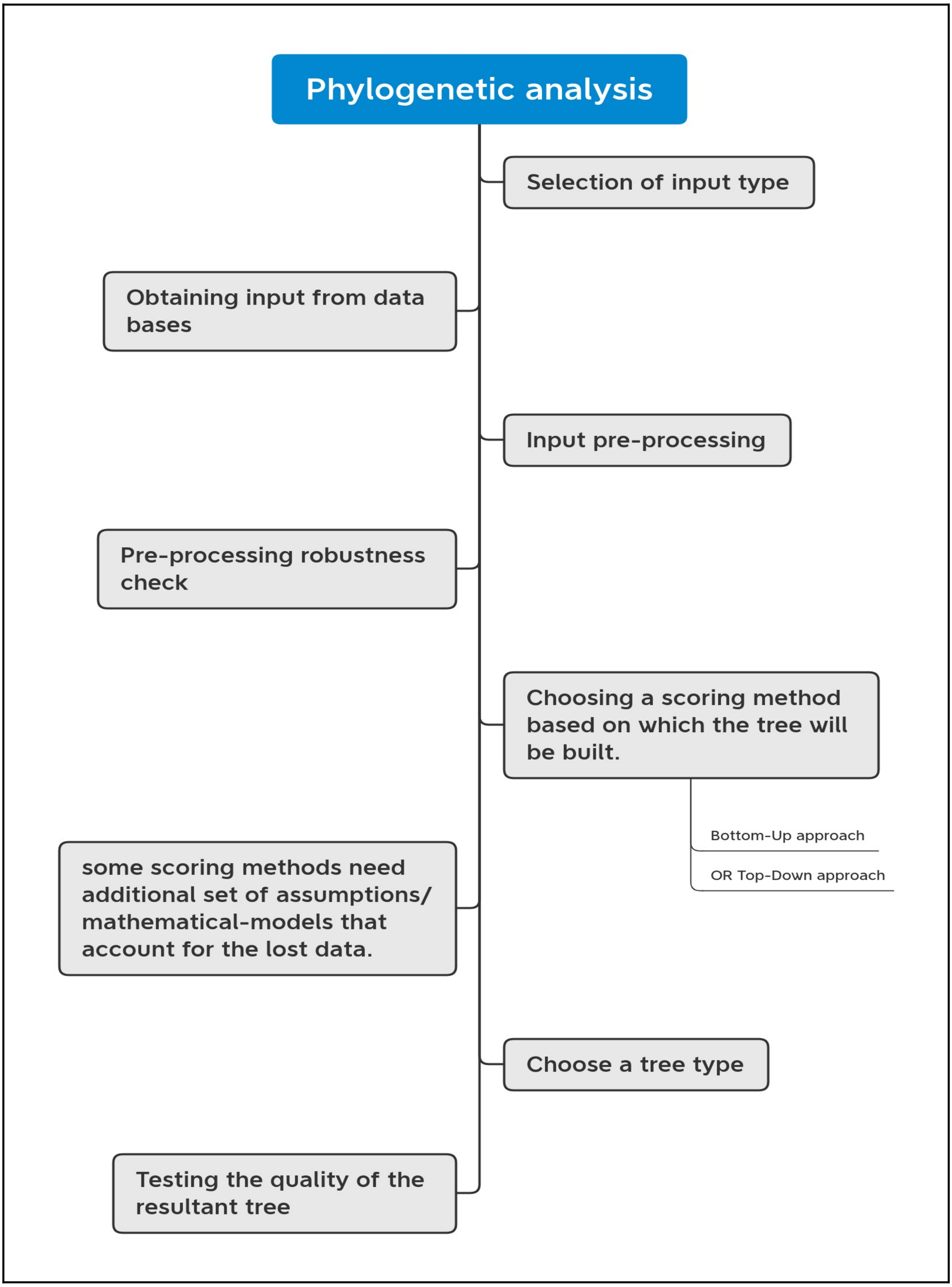

Phylogenetic analysis
Selection of input type
Obtaining input from data bases
Input pre-processing
Pre-processing robustness check
Choosing a scoring method based on which the tree will be built.
Bottom-Up approach
OR Top-Down approach
some scoring methods need additional set of assumptions/ mathematical-models that account for the lost data.
Choose a tree type
Testing the quality of the resultant tree

Fig 1: Simplified version of phylogenetic analysis pipeline

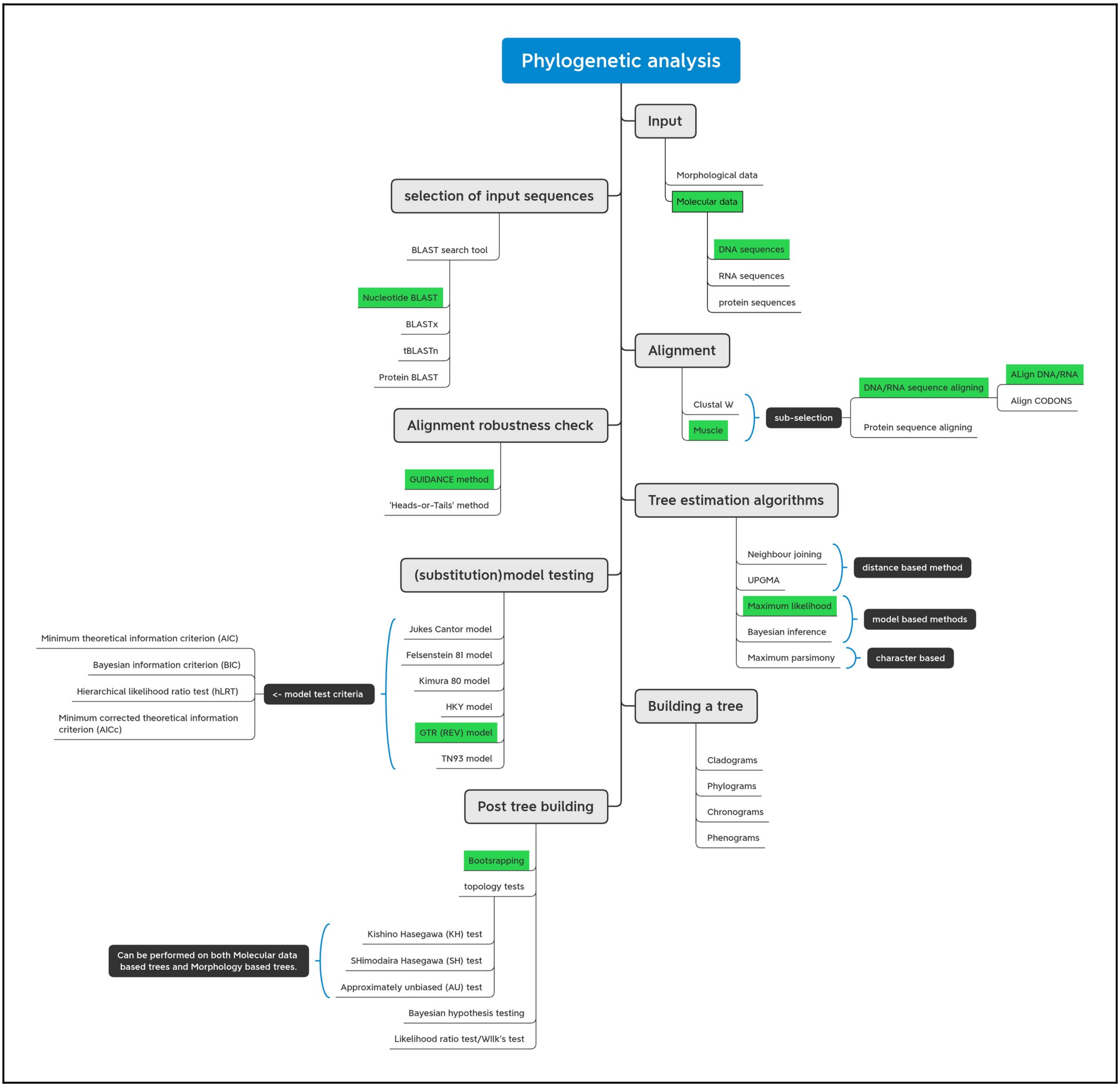


Fig 2: A detailed map/pipeline of the phylogenetic analysis with choices in each step highlighted in green.

The following sections will address each of the steps in more detail.

Input:

Information extracted from various organisms will be used to understand the relationship between those species. Information that is used for phylogeny used to be morphological data, i.e, physiological and behavioral observations of organisms. Lately, Molecular data like DNA, RNA and Protein sequences are being used to infer phylogenetic relationships.

Types of input:

- Morphological data:
  - behavioral properties
  - physiological and structural properties
  - developmental processes etc.,
- Molecular data
  - DNA sequences
  - RNA sequences
  - Protein sequences etc.,

Morphological data:

Morphological data implies physiological and behavioral feature that are observed with a lot of care and precision. It is hard to collect such data when fossils are involved. Because extremely small percentage of remains of ancient species, in the form of fossils, will make it to the researcher. Majority of the physical evidence is lost in time and the study of available small remains is extremely difficult by using just morphological data. Morphological data also suffers from consistency issues. Unlike molecular data, which is discrete in nature, morphological features have no universally accepted standards. The features observed vary from organism to organism that are being analyzed. For example, features like limb lengths, bone structure, placement of eyes etc are only suitable when the study is performed on creatures that are of at-least certain size. Such features may not be suitable when the study is performed on microscopic organisms like bacteria, fungi, viruses etc., If a generalized feature matrix is to be built for a large scale comparison, then for large number of organisms the feature matrices tend to be extremely sparse.

Molecular data:

The building blocks of all life form are proteins, amino acids and other complicated formulations of carbohydrates. DNA, RNA sequences have been part of every life form since the beginning of life itself and they contain the instructions needed for life to happen. The discrete nature of such molecular sequences comes from this fact that they all are made of same ingredients. But the amount of information in such sequences are extremely long.

Finding useful information from such long strings of discrete sequences is computationally heavy task. This is exactly the reason why molecular phylogenetic analysis had to wait till the advent of affordable computational power.

Selection of input sequences:

Choice of which type of sequence also depends on the type of analysis that is being performed.

- Phylogenetic analysis within a population:
  - Non-coding regions of Mitochondrial DNA are often used. These are the regions of DNA that don’t have much biological role.
- Phylogenetic analysis between widely divergent groups of organisms:
  - Slowly evolving nucleotide sequences like RNA and Protein sequences
- Phylogenetic analysis between deepest levels of organisms like bacteria and eukaryotes (cells with nucleus enclosed in them)
  - Conserved protein sequences (sequences that are continued by natural selection) perform better than nucleotide sequences.

The first step of selecting a sequence is the most intellectually demanding step. The basic assumption is that , all the sequences on a tree are homologous in nature. Homologous sequences are the gene sequences that are inherited in 2 species by a common ancestor. The most reliable way to identify sequences that are homologous to sequences_of_interest is, to do a BLAST search, using the sequence_of_interest as a query. [2]

What is Blast search?

BLAST means ‘Basic Local Adjustment Search Tool’. It finds the local similarities in Nucleotide or Protein sequences of different organisms in the database. It finds a statistically significant match for the query sequence from the database.

- Types of algorithms in BLAST
  - Blastn – For Nucleotide sequences
    - Mega_blast – highly similar sequences
    - Discontigous mega_blast – dissimilar sequences
    - blastn – somewhat similar sequences
  - Blastp – For Protein sequences
  - tblastn – a protein query sequence translated and compared to nucleotide sequence in database
  - Blastx – a nucleotide sequence translated and compared to protein sequence in database
  - tblastx – Both the query and batabase nucleotide sequence are translated and compared using Blastp algorithm.

Simply put, Blast search tool is used to statistically find the sequences from the database. Once the sequences of different taxa or organisms are obtained, the next step is to find the alignment between all the sequences.

What is sequence aligning?

It is a way arranging the nucleotide or protein sequences to identify similar regions. Some portions of the sequences remain unchanged/un-mutated across several species. Identifying such regions of such extremely long strings of discrete sequences, will make identification of variation among those species, easier. Finding such alignments is very labor intensive task for a human to perform, since the sequences that are under comparison are too long for humans. Several algorithms and methods are developed to perform this operation automatically. Fig 3 shows the unaligned sequences of an example phylogeny analysis.

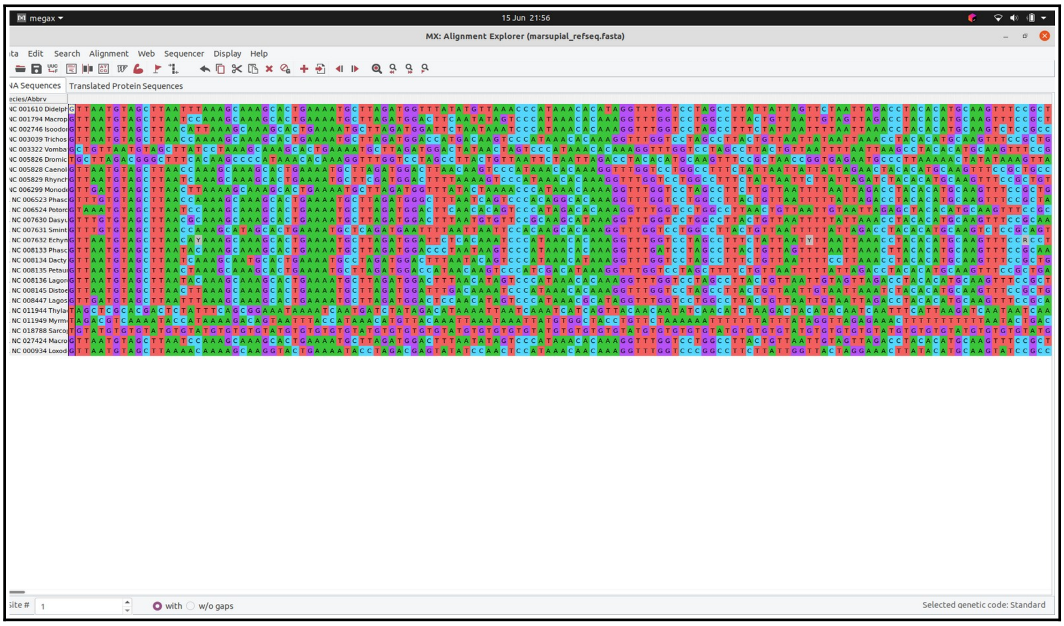
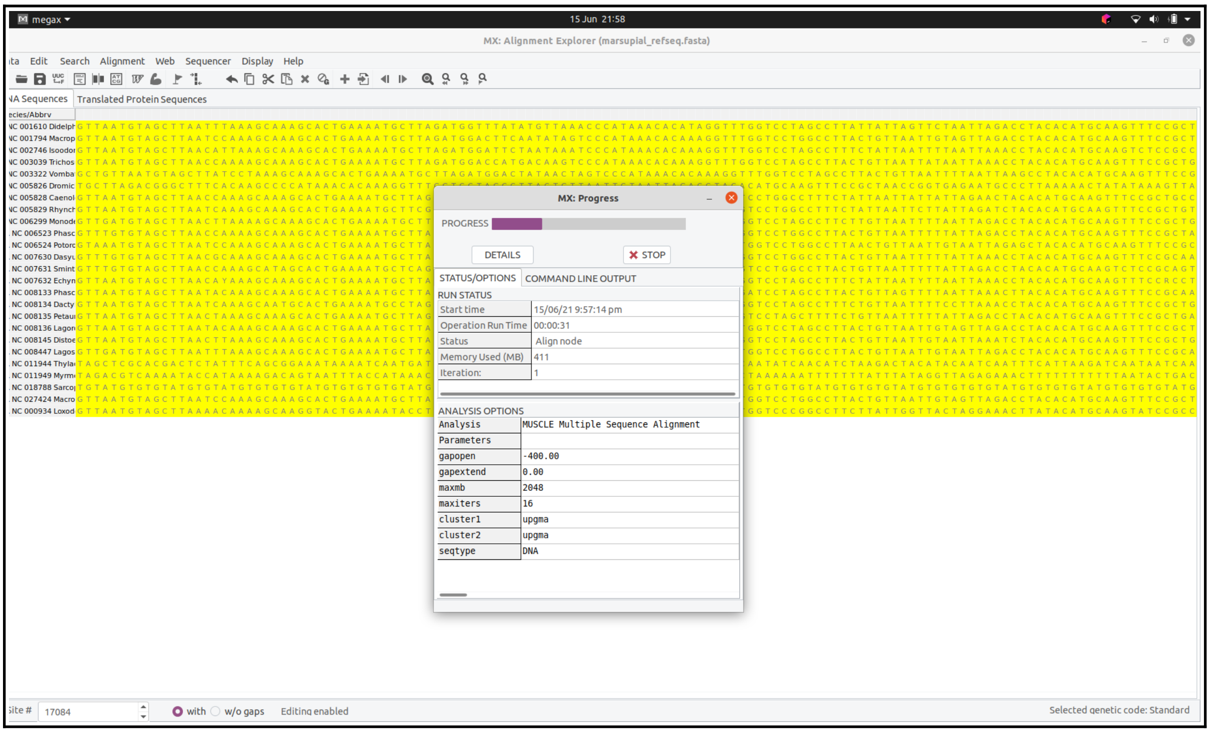


Fig 3. unaligned nucleotide sequences

MEGA X software offers two of such most popular algorithms are CLUSTALW and MUSCLE. Of the two, MUSCLE alignment algorithm is most commonly used because it is faster and achieves better alignments compared to CLUSTALW. Fig 4 shows the sub selections involved in aligning sequences of an example phylogeny analysis.


megax
15 Jun 21:58
MX: Alignment Explorer (marsupial_refseq.fasta)
ta Edit Search Alignment Web Sequencer Display Help
NA Sequences
Translated Protein Sequences
ecies/Abbrv
DETAILS
✖ STOP
STATUS/OPTIONS
COMMAND LINE OUTPUT

MX: Progress

PROGRESS

RUN STATUS

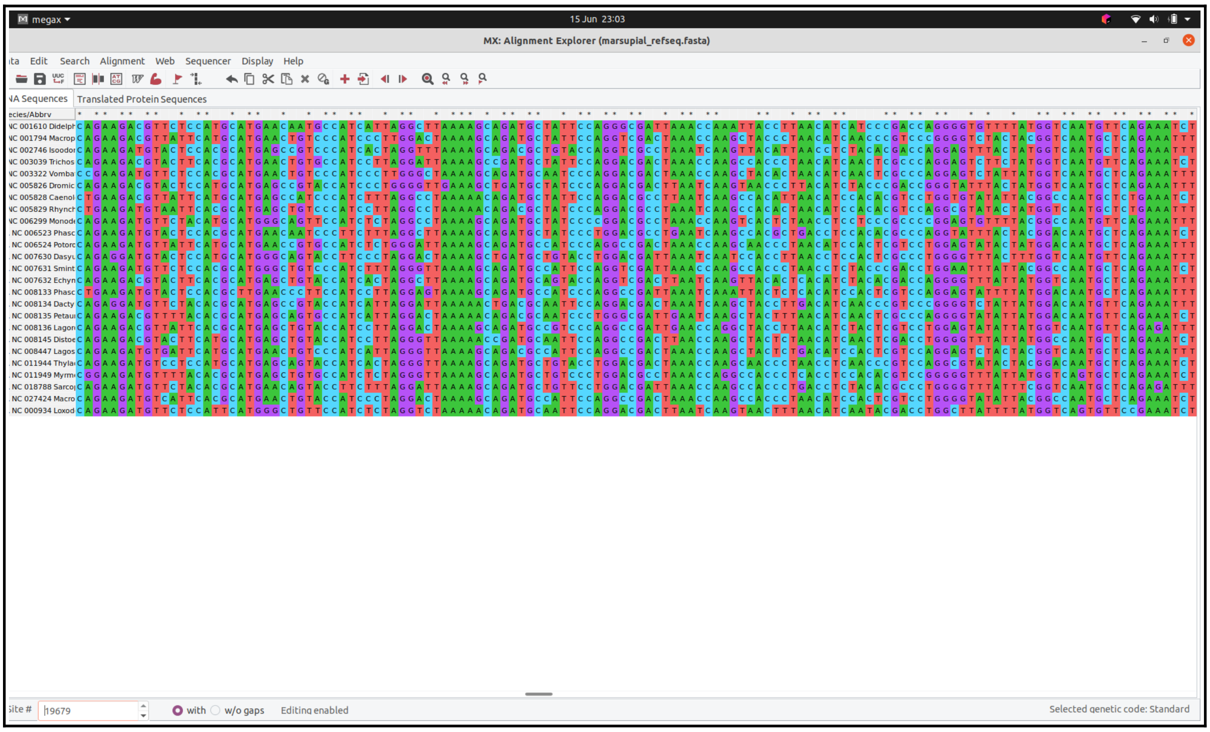
Start time	15/06/21 9:57:14 pm
Operation Run Time	00:00:31
Status	Align node
Memory Used (MB)	411
Iteration:	1

ANALYSIS OPTIONS

Analysis	MUSCLE Multiple Sequence Alignment
Parameters	
gapopen	-400.00
gapextend	0.00
maxmb	2048
maxiters	16
cluster1	upgma
cluster2	upgma
seqtype	DNA

Site # 17084 ● with ○ w/o gaps Editing enabled Selected genetic code: Standard

Fig 4. Using MUSCLE method for sequence alignment


megax
15 Jun 23:03
MX: Alignment Explorer (marsupial_refseq.fasta)
ta Edit Search Alignment Web Sequencer Display Help
NA Sequences
Translated Protein Sequences
ecies/Abbrv

Site # 19679 ● with ○ w/o gaps Editing enabled Selected genetic code: Standard

Fig 5. Nucleotide sequences after alignment

Aligning operation is time intensive, based on the number of taxa in the analysis and the lengths of the sequences. A series of sub selections will be needed before performing the MUSCLE alignment operation, like, gap-penalties, codon alignment vs DNA alignment, stop codons identification etc. Gaps represent historical insertion and deletion mutations that happen in the sequences. Purpose is to bring homologous sequences (the part of the sequences which do not change easily with generations) into alignment. Fig 5. shows the aligned nucleotide sequences.

Once the alignment is achieved, One can check the quality of alignment using GUIDANCE server. A web server that can be used to perform the robustness of the alignment of the sequences based on a confidence score.

- Checking alignment
  - GUIDANCE2 method
  - GUIDANCE method
  - Head or Tails method

Post alignment, The tree estimating algorithm is to be selected. These algorithms can be categorized as following,

- Distance based
  - Neighbor joining
  - UPGMA - Unweighted Pair Group Method With Arithmetic Mean
- Character based
  - Maximum parsimony
- Model based
  - Maximum likelihood
  - Bayesian inference

The above mentioned algorithms are used to calculate a distance or a score which is derectly used to build the tree. Tree building is more often a bottom-up approach, which means, the branches of the tree built first and based on the similarity score or distance score obtained from the above mentioned algorithms, the branches are combined to form nodes and root.

Model based algorithms:

The algorithm used in the pipeline shown in fig 2 is the ‘maximum likelihood’ algorithm. Model based algorithms are more reliable because here there are more parameters that can be controlled to obtain a better tree but always at a cost of computation.

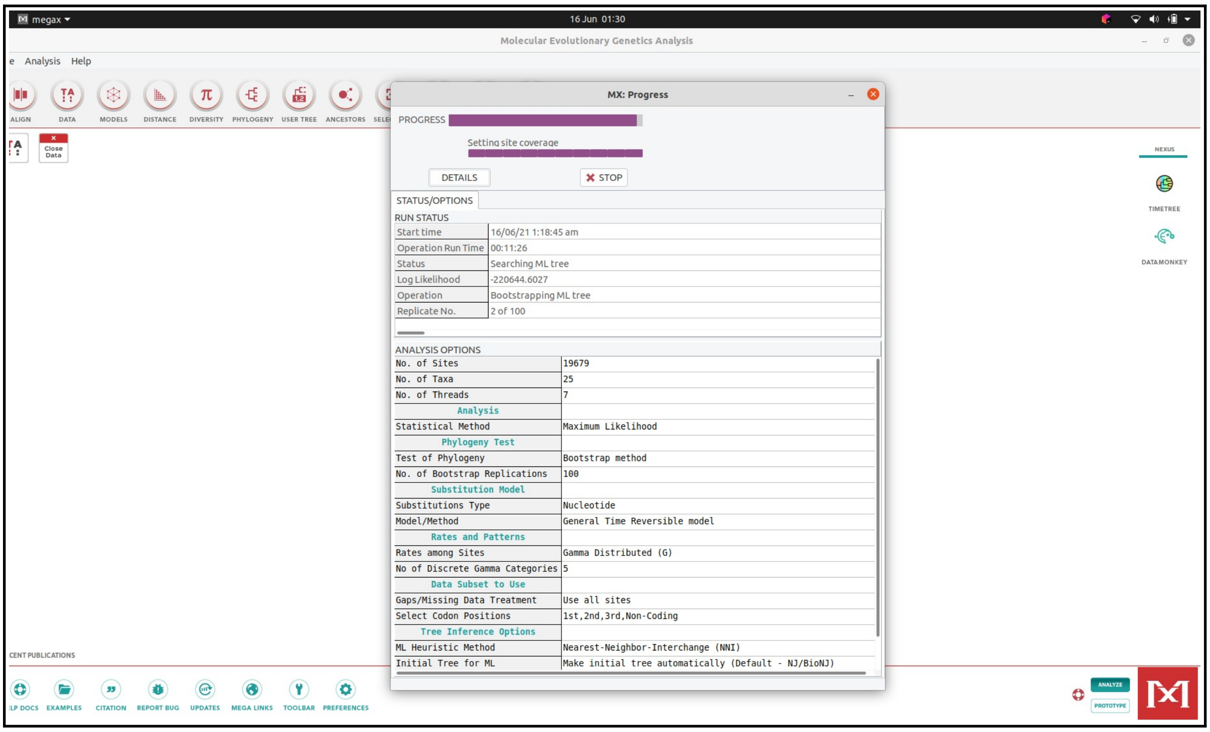
Substitution models:

The substitution models are ‘Markov models’(future state is dependent only on the current state and independent of past states) that “describes” the changes over evolutionary time. Substitution models attempt to estimate the pattern_of_mutation that occurs in a sequence during the process of evolution. This information in reality is completely lost, but the substitution models try to make a estimate/model of such mutations and the maximum likelihood algorithms checks the likelihood of a phylogeny tree based on the number of substitutions that have occurred since a pair of sequences diverged from a common ancestor [3]. Substitution models are used to correct for multiple changes at same site during the evolutionary history of the sequences.

Some of the most common substitution models are, in no particular order,

- Jukes Cantor model (JC69)
- Falsenstein81 (F81)
- Kimura 80 model (K80)
- HKY model
- GTR (REV) model
- Tamura and Niel model (TN93)

The above mentioned models vary in how the rates of substitutions and their relationships are modeled, distribution of weights etc., The more prior information is known about the sequences, the more control one can have in choosing the substitution model. Fig 6 shows the parameters that are used for an example phylogeny analysis.



RUN STATUS

Start time	16/06/21 1:18:45 am
Operation Run Time	00:11:26
Status	Searching ML tree
Log Likelihood	-220644.6027
Operation	Bootstrapping ML tree
Replicate No.	2 of 100

ANALYSIS OPTIONS

No. of Sites	19679
No. of Taxa	25
No. of Threads	7
Analysis	
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	100
Substitution Model	
Substitutions Type	Nucleotide
Model/Method	General Time Reversible model
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
No of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Select Codon Positions	1st,2nd,3rd,Non-Coding
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)

Fig 6. Maximum likelihood and respective parameters.

How to choose a model?

In MEGA X, model selection tool suggests the best suitable model for the given input sequences. The tool doesn't suggest a single model but instead offers a score under various test criteria for each substitution model. The user must choose the criteria that matters most for their experiment/phylogenetic analysis. Some of the most valued criteria, in no particular order, are

- Minimum theoretical information criteria (AIC) – Higher the better
- Bayesian information criteria (BIC) – Lower the better
- Hierarchical likelihood ratio test (hLRT) – Higher the better
- Minimum corrected theoretical information criteria (AICc)

Tree building:

Based on the model selection tool suggestions, a substitution model is selected and the software begins to build a tree based all the given parameters (phylogeny analysis algorithm, substitution model, gamma distribution etc.,).

Types of trees:

- Cladograms
- Phylograms
- Chronograms
- Phenograms

Tree quality testing:

- Bootstrapping

    - Building 100's of sets of trees by sampling from the sequences.
    - The most repeated tree is considered to be the final tree.
    - Statistical testing for phylogeny can be done for
        - study of tempo and mode in the evolution i.e., rate of evolution.
        - Variations in natural selection regimes
        - To test correlated evolution among traits.
- Likelihood ratio test or Wilk's test
- Bayesian hypothesis testing
- Topology testings
    - Kishino Hasegawa test
    - Shimodaira Hasegawa test
    - Approximately unbiased test

REFERENCES:

1. https://en.wikipedia.org/wiki/List_of_phylogenetics_software
2. Wheeler D, Bhagwat M. BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. In: Bergman NH, editor. Comparative Genomics: Volumes 1 and 2. Totowa (NJ): Humana Press; 2007. Chapter 9. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1734/
3. Wikipedia contributors. (2021, June 4). Substitution model. In *Wikipedia, The Free Encyclopedia*. Retrieved 00:05, June 30, 2021, from https://en.wikipedia.org/w/index.php?title=Substitution_model&oldid=1026805568