A critical review on different techniques for phylogenetic analysis

[1]Shivam Bhardwaj, [2]Simran Jeet Kaur, [3]Deepti Jamwal, [4]Bhola Das
[1,2,3,4]Student, Department of Biotechnology, Manav Rachna International Institute of Research and Studies, Faridabad, India

Abstract - **Understanding the evolutionary relationship between species (called taxa that are supposed to have a common ancestor) is the basis of phylogenetic study. Since the starting of the civilization of animal kingdom, scientist's curiosity about the evolution process and origin of life has been a great advantage for the field of biology. It has helped in the classification of species and studying its relationship with the other organisms and species. This study of evolutionary study and finding relationship is phylogenetic analysis. This relationship can be inferred using construction of tree like diagram called phylogenetic tree. It is said that all organisms in the earth are part of a single phylogenetic tree, thus indicating that they have a common ancestor. there are basically 2 different types of methods of phylogenetic analysis i.e., distance-based methods and character-based methods or discrete methods. In distance-based methods trees are calculated based on similarities of sequences and are based on distances. This method compresses all the individual differences between a pair of sequences into a single number. UPGMA, neighbor joining method are some examples of this. In character-based method trees are constructed by considering the various pathways of evolution and uses each alignment position as evolutionary information. Maximum parsimony and likelihood are methods used in this type of evolutionary findings. Depending on different method for construction of phylogenetic tree different tools like dnadist, Dnamove, neighbor, etc are available. This method is widely used in different fields like forensic, comparative genomics, bioinformatics, drug discovery, and many more. It is an emerging field of biology and somehow linked to every other field. This review basically aims to analyse the methodology of different methods for phylogenetics analysis.**

Index Terms - **Bioinformatics, Phylogeny, Phylogenetic analysis, Phylogenetic tree, UPGMA.**

I.INTRODUCTION

The transfer of heritable traits in an organism or a biological population through generations is referred to as evolution. The Darwinian theory of biological evolution acknowledges that all life on Earth has a common ancestor. The broadening of this regular precursor happens through speciation occasions, for example, common choice, hereditary float, and change of hereditary material that prompts the grouping of life (Sarma et al., 2019). Many researchers have aspired since the time of Charles Darwin to recreate the evolutionary history of all creatures on Earth and present it in the form of a phylogenetic tree (Scq.ubc.ca, 2006). This increases the importance of study of evolutionary relationships, and now we will go through what actually phylogeny is? Also, their usage through bioinformatics and various publicly tools for phylogeny.

A phylogeny is a group of entities in evolutionary history. Phylogeny classifies species based on evolutionary distance or evolutionary connection (taxonomy). The main objective of phylogeny rebuilding is to describe evolutionary relationships in terms of relative recency of common ancestry. These relationships are depicted as a branching diagram, or tree, with node branches that lead to the ends of a tree (Singh, n.d.). Basically, a phylogeny is a pattern of historical evolutionary relationship among species and higher-level taxa that is often presented as a tree diagram, or phylogenetic tree. The term phylogenetics is often applied to the study of such relationships. Phylogenies are analogous to genealogies on the scale of species and higher-level taxa (e.g., genera and families). Phylogenies are typically depicted as tree-like branching diagrams, with species on the same branch assumed to be more closely related to each other than species on other branches. Interpreting such diagrams as historical patterns requires a basic understanding of hierarchy, and phylogenetic trees are often incorrectly assumed to support historical suppositions (e.g., one modern taxon is “primitive”) that are not indicated by the results. Phylogenies are typically depicted as tree-like branching diagrams, with species on the same branch assumed to be more closely related to each other than species on other branches. Additional uses of phylogenetic information include measures of phylogenetic diversity, which can be used in making conservation and habitat preservation decisions (Nixon, 2001). As a result, most phylogenies are theories based on indirect evidence. Different phylogenies often emerge using the same evidence. Nevertheless, there is universal agreement that the tree of life is the result of organic descent from earlier ancestors and that true phylogenies are discoverable, at least in principle (Phylogeny - Taxonomic systems | Britannica, 2019).

II.RELATIONSHIP BETWEEN PHYLOGENETIC ANALYSIS AND SEQUENCE ANALYSIS

When two sequences found in two organisms are very similar, we assume that they have derived from one ancestor. The sequence's alignment reveals which positions are conserved from the ancestor sequence.

The progressive multiple alignment of a group sequence first aligns the most similar pair. Then it adds the more distant pairs. The alignment is influenced by the “most similar” pairs and arranged accordingly, but it does not always correctly represent the evolutionary history of the occurred changes. But not all phylogenetic methods work in this way. Most phylogenetic approaches presume that each place in a sequence may change independently of the others. Gaps in alignments represent mutations in sequence such as: insertion, deletion, genetic rearrangements. Phylogenetic approaches deal with gaps in a variety of ways. Most of them overlook gaps.

Another approach to treat gaps is by using sequence similarity scores as the base for the phylogenetic analysis, instead of using the alignment itself, and trying to decide what happened at each position. The similarity scores based on scoring matrices (with gaps scores) are used by the DISTANCE methods (Introduction to Phylogenetic Analysis, n.d.).

1. How are sequences selected for phylogenetic analysis?

Depending on this your tree might include sequences even from different domains. On the other hand, if you are analysing micro diversity within a species population your tree might include multiple sequences from strains from the same species (>=97% sequence identity in ssu_rRNA gene).

In general, homologous sequences are the requisite for a phylogenetic tree. Exactly the same size of the sequences is not necessary because that is not not expected to occur naturally anyway. However, the larger the shared homologous region of the sequences, the better the phylogenetic analyses are expected to behave because of some statistical assumptions behind them (Mount, Harbor, 2006). Get some sequences from the respective gene from GenBank, align them and then do a quick and dirty maximum parsimony or likelihood analysis just to check if your selected region has enough information to give you a supported phylogeny (Müller, Sebastian, 2015).

2. Different Methods for Phylogenetics analysis:

Phylogenetic approaches may be utilised for a variety of reasons, including the examination of morphological and molecular data. Here are some highly recommended methods which are available in bioinformatics for analysis of phylogenetics.

2.1. Distance matrix methods:

Non-parametric distance methods were first used to phenetic data using a matrix of pairwise distances; hence distance matrices are employed in phylogeny. These distances are then added together to make a tree (a phylogram, with informative branch lengths). The distance matrix can be derived from a variety of sources, including measured distance (for example, from immunological studies) or morphometric analysis, various pairwise distance formulae (for example, Euclidean distance) applied to discrete morphological characters, or genetic distance derived from sequence, restriction fragment, or allozyme data. For phylogenetic character data, raw distance values can be calculated by simply counting the number of pairwise differences in character states (Manhattan distance) (www.bionity.com, n.d.). Approaches based on distance-matrix phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified, hence an MSA is required as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches (Mount and Harbor, 2006). Distance approaches try to generate an all-to-all matrix expressing the distance between each sequence pair from the sequence query set. A phylogenetic tree is built from this, with closely related sequences placed under the same core node and branch lengths that closely match the reported distances between sequences. Depending on the methodology used to construct them, distance-matrix approaches can create either rooted or unrooted trees. They are commonly used as the foundation for both progressive and iterative multiple sequence alignment. The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees (Felsenstein and Sinauer Associates, 2008).

A. Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

UPGMA (unweighted pair group method with arithmetic mean) is a straightforward approach to constructing a phylogenetic tree from a distance matrix. It is the only method of phylogenetic reconstruction in which the resulting trees are rooted. UPGMA implicitly implies a constant rate of substitution across time and across phylogenetic lineages (known as the molecular clock hypothesis). Since this assumption is often violated, this method is now rarely used (Sciencedirect.com, 2014).

The UPGMA algorithm

- UPGMA begins with a pairwise distance matrix D [1.n, 1.m].
- Each sample (taxon, operational taxonomic unit=OTU) is referred to as a 'cluster' in the following text.
- It starts by assigning all clusters (samples) to a star-like tree.

1. Find that pair (cluster i and j) with the smallest distance value in the distance matrix: D [i, j].

2. Define a new cluster comprising cluster i and j:

A branch connects Cluster i to the common ancestor node. The same is true for cluster j.

As a result, the distance D [i, j] is divided into two branches. As a result, each of the two branches has a length of D [i, j]/2.

3. If i and j were the last 2 clusters, the tree is finished. If this is not the case, the algorithm creates a new cluster called u.

4. Define the distance from u to each other cluster (k, with k <> i or j) to be an average of the distances dki and dkj.

For 'Weighted PGMA (WPGM)': dku = dki+dkj/2).
For 'Complete linkage': dku = max (dki, dkj).
For 'Single linkage': dku = min (dki, dkj).

5. Go back to step 1 with one less cluster. Clusters i and j are eliminated, and cluster u is added to the tree (www.sequentix.de, n.d.).

Characteristics of UPGMA:

UPGMA is the simplest method for constructing trees. It always Generates rooted trees that re-rooting is not allowed. It generates ultra-metric trees. This method is easy to compute through hand or by software tools.

Disadvantages of UPGMA:

Along with benefits, UPGMA has some disadvantages also like it assumes the same evolutionary speed on all lineages. Often, it generates wrong tree topologies. Re - rooting is not allowed in UPGMA and it relies on a randomized molecular clock (Shreya Feliz, 2017).

B. Neighbor Joining Method:

The star decomposition technique is a subset of the neighbor-joining approach. In contrast to cluster analysis, neighbor-joining tracks nodes on a tree rather than taxa or taxonomic groups. The raw data are provided as a distance matrix and the initial tree is a star tree (Treehozz.com, 2020). A neighbor-joining approach for rebuilding phylogenetic trees using evolutionary distance data is proposed. The basic idea behind this technique is to select pairings of operational taxonomic units (OTUs [=neighbours]) that have the shortest total branch length at each step of OTU clustering, beginning with a starlike tree. Using this approach, you can rapidly determine the branch lengths and topology of a parsimonious tree. The usual methodology of tree-making techniques based on this concept is to analyse all conceivable topologies (branching patterns) or a specific set of topologies that are likely to be near to the genuine tree and select the one with the least amount of total evolutionary change as the final tree (Saitou2 and Nei, n.d.).

Algorithm

Neighbor-joining is a recursive algorithm. Each recursion stage consists of the following steps:

1. Based on the current distance matrix, calculate a modified distance matrix Q (see below).

2. Find the least distant pair of nodes in Q (= the closest neighbors = the pair with the lowest distance value). Make a new node on the tree that connects the two nearest nodes: the two nodes are connected by their shared ancestor node.
3. Calculate the distance of each of the nodes in the pair to their ancestral node.
4. Calculate the distance of all nodes outside of this pair to their ancestral node.
5. Start the algorithm again, considering the pair of joined neighbors as a single taxon (the terminal nodes are replaced by their ancestral node and the ancestral node is then treated as a terminal node) and using the distances calculated in the previous step (www.sequentix.de, n.d.).

Negative branch lengths

NJ represents the data in an additive tree. Therefore, it can assign negative branch lengths. Typically, branch lengths may be taken as an estimate for the number of substitutions. However, we are having difficulty doing so in this case.

If this occurs, the negative branch length can be adjusted to zero and the difference transferred to the adjacent branch length, preserving the total distance between an adjacent pair of terminal nodes. This has no effect on the tree's overall structure (A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates., 1994) Advantages of Neighbor Joining Method:

The main advantage of the NJ method is that it is fast as compared to other processes which makes it efficient for large data sequences. The property of neighbour joining is that if the input distance matrix is right, the output tree will be accurate. Furthermore, the correctness of the output tree topology is guaranteed as long as the distance matrix is 'nearly additive. Nevertheless, neighbor joining has been largely superseded by phylogenetic methods that do not rely on distance measures and offer superior accuracy under most conditions (Wikipedia, 2021).

Disadvantages.

One of the disadvantages of the NJ method is that it only produces one tree even if there is any possibility of constructing other trees also.

C. Fitch - Margoliash Method

For grouping based on genetic distance, the Fitch–Margoliash technique employs a weighted least squares technique. To compensate for the increasing error in estimating distances between distantly related sequences, closely related sequences are given greater weight in the tree construction process. To avoid major artefacts in computing connections between closely related and distantly connected groups, the distances used as input to the method must be normalised. This approach requires linear distances to be determined. The distance linearity criterion requires that the expected values of the branch lengths for two individual branches equal the expected value of the sum of the two branch distances - a property that only applies to biological sequences when they have been corrected for the possibility of back mutations at individual sites. This correction is done using a substitution matrix such as that derived from the Jukes-Cantor model of DNA evolution (Wikipedia, 2021). The Jukes-Cantor model equation is…

$d_{ij}=-3/4 \ln (1- 4/3 * p)$

where p is the fraction of mismatches between sequences. Using the JC distances, the FM distance-based algorithm produces unrooted phylogenetic trees in which the common ancestral node is unknown (to produce rooted trees one can, use an outgroup that is known to be distantly related) (www.cs.gettysburg.edu, n.d.)

Least Square Technique in FM Algorithm:

The least squares methodology is a core process in regression analysis for approximating the solution of over determined systems (sets of equations with more equations than unknowns) by minimizing the sum of the squares of the residuals created in the outcomes of each individual equation.

The most important application is in data fitting. In the least-squares sense, the best fit minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model). When there are significant uncertainties in the independent variable (the x variable), basic regression and least-squares procedures fail; in such instances, the approach necessary for fitting errors-in-variables models should be examined instead of least squares (Theoretical foundation of the minimum-evolution method of phylogenetic inference., 1993).

To create a tree, the data can be input as a distance matrix or as a collection of DNA sequences. Entering the data as a distance matrix allows you to provide your own particular mutational distances for each of your labels in addition to those estimated using the Jukes-Cantor model. To enter data for each taxon into a properly formatted distance matrix each taxon should be given its own row starting with the taxon name and followed by the evolutionarily related distances. When a sequence is compared to itself, it has a distance of zero and must be placed into the matrix such that the zeroes are on the main. It is also important to ensure that the names do not contain spaces and that underscores are used instead. There should, however, be spaces between all other entries, and each taxon should be on its own line.

Data may also be entered as DNA sequences in order to form a tree and determine mutational distances using the Jukes-Cantor model. The sequences submitted MUST be of equal length and must include labels. The label should go first, followed by a colon and then the sequence (www.cs.gettysburg.edu, n.d.)

Advantages of FM algorithm:

Its big advantage is that it generates a tree that preserves distances between data “as much as possible” in the sense that it should minimize a selected criterion.

D. Minimum Evolution Method:

This method was proposed by Rzhetsky and Nei in 1992. It is a simple method for estimating and testing phylogenetic trees under the principle of minimum evolution (ME). The minimum-evolution (ME) approach of phylogenetic inference is based on the notion that the tree with the least sum of branch length estimates is the most likely to be correct. This assumption has previously been used without mathematical evidence.

Algorithm:

The algorithm of their method is first to construct a Neighbor - Joining Tree using Saitou and Nei’s (1987) procedure and to compute the total sum (S) of branch lengths for this tree. Following that, all tree topologies that meet specific requirements and are near to the NJ tree are investigated, and the S value for each tree is computed. The S values acquired are then compared, and the tree with the lowest S value is chosen as the final one. This last tree is often the NJ tree, though the NJ approach occasionally fails to identify the ME tree. Of course, if the number of sequences is small enough, it is feasible to evaluate all topologies, although this is typically unneeded. In addition, a statistical test for the difference in S across various topologies was created. Computer simulations have shown that the ME method is more efficient than most other distance methods of phylogenetic inference and that the statistical test proposed is conservative (Theoretical foundation of the minimum-evolution method of phylogenetic inference., 1993)(A Simple Method for Estimating and Testing Minimum-Evolution Trees, 1992).

Disadvantages:

This method is a bit slower than other clustering methods. It requires a lot of information when characters have to be converted into distance.

2.2. Character Based Method:

It is also known as the discrete method. This method is based directly on the sequence characters. The character-based technique constructs the phylogenetic tree using the aligned characters. Some of the methods which are based on the character-based phylogeny analysis are as follows:

A) Maximum Parsimony:

Maximum parsimony is a method for predicting the evolutionary tree that requires the fewest steps to produce the observed variation in sequences. It is constructed with the fewest adjustments necessary to describe (tree) the data discrepancies.

Algorithm:

Here, a phylogenetic tree T with $L(T) < 2L(T)$ is the most parsimonious tree. Consider $G(S)$ be a graph whose vertex set is the set S of taxa, and weight of every edge (ai, aj) which is known as Hamming distance between the characters of ai and aj.

It is followed by converting the T’ to a phylogenetic tree T by replacing every internal node labelled by taxa with an unlabelled node and attaching a leaf labelled by a as its child.

Advantage: Appropriate for a small number of sequences with a lot of similarities.

Disadvantages:

1. It takes a long time to test all feasible trees.
2. For divergent sequences, parsimony may fail.
3. It is susceptible to the long-branch effect (Munjal, Hanmandlu and Srivastava, 2019)

B) Maximum likelihood:

It seeks to find the model with the best chance of generating the input sequence in a particular evolutionary model.

Algorithm:

They are based on two separate computer programs. The first program is a function (called FUN) that accepts as inputs a value for the parameter vector θ and $ and the data ln [L(θ; $)] and returns the log-likelihood value as output.

Advantages:

1. Suitable for sequences that are extremely dissimilar.
2. We can build evolutionary relationships hypotheses.
3. More precisely, in a reasonable amount of time, phylogenetic trees can be generated for a limited number of taxa.

Disadvantages:

1.Slow search algorithms result in slow responses.
2.For huge datasets, it takes a long time (Munjal, Hanmandlu and Srivastava, 2019).

III.IMPORTANCE OF PHYLOGENETICS IN BIOLOGY

Phylogenetics is an emerging field in biology that is being rapidly used in different fields. One of the major importance, when we talk about phylogenetics, is that it has contributed greatly to the classification of organisms, to be more precise, speciation. It helped the scientists to trace the genetic history of species and thus classifying them. It also concluded that every species present today are the descendants of a single common ancestor (Larson, 2021).

Understanding the concept of evolution has been a widely studied topic throughout the history of biology. Evolution has been said to be the only explanation for the diversity of life. And for studying the evolutionary history and relationship amongst a group of organisms, phylogenetic analysis is a widely used method. We can find out how closely related an organism is to another one just by viewing and analysing the phylogenetic tree diagram. It gives us a major idea of how our molecular sequences (gene, genome) evolve. It also helps us to understand and predict the future evolution of species (EMBL-EBI, n.d.).

IV.APPLICFATIONS OF PHYLOGENETIC:

When we talk about applications of phylogenetics in different fields, the very first application that comes to mind is its role in the field of evolution. But the application of this emerging field is endless, from conservation biology to drug development. Some of its applications are listed below

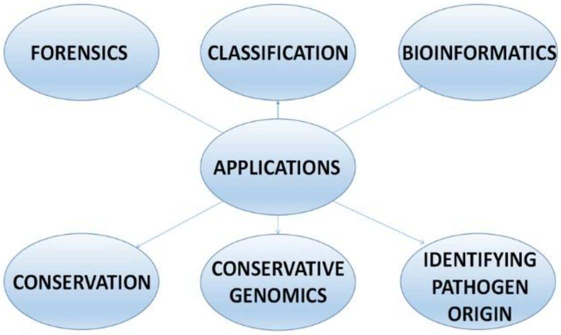

Fig. 1. Applications of phylogenetics in different fields

1. FORENSICS: - In the discipline of forensics, phylogenetics is commonly used to evaluate DNA evidence provided in court cases to inform circumstances such as when someone has committed a crime, when food has been tainted, or when the father of a child is unknown.
2. CLASSIFICATION: - Phylogenetics based on sequence data provides more accurate explanations of patterns of relatedness than was previously accessible. The Linnaean taxonomy of new species is currently based on phylogenetics.
3. BIOINFORMATICS: - Many of the phylogenetics algorithms have been utilized to create software in other domains.
4. CONSERVATION: -When conservation biologists have to make difficult decisions about which species to try to save from extinction, phylogenetics can assist them make better decisions.
5. COMPARATIVE GENOMICS: -In a phylogenetic perspective, comparisons of different species or gene sequences can provide the most useful insights into biology. Researchers in a variety of domains,

including ecology, molecular biology, and physiology, have come to this critical insight.

6. IDENTIFYING ORIGIN OF PATHOGEN: - To learn more about a new pathogen epidemic, molecular sequencing methods and phylogenetic techniques can be applied. This includes determining which species the disease belongs to and, as a result, the most likely source of transmission. This could lead to new public health policy recommendations (EMBL-EBI, n.d.) (Soltis and Soltis, 2003)

V. PROBLEMS RELATED TO PHYLOGENETIC ANALYSIS

The order of the input sequence has an impact on tree construction. Also, the amount of data used in tree construction is not always meaningful. Genetic data is also used to construct several phylogenetic trees. Various genes, on the other hand, evolve in different ways, genes can be switched between animals, and organisms can lose genes. As a result, the genes used in a phylogenetic tree must represent the scientific topics being addressed. Another feature of phylogenetic trees is that, unless otherwise stated, the branches do not account for the amount of time that evolution took place. Rather, they reflect the evolutionary divergence of lineages and the sequence in which they diverged. In other words, unless the data is directly associated to time with an evolutionary model, a long branch does not often suggest more time passed, and a short branch does not often suggest less time passed (Geosciences LibreTexts, 2019)

VI. CONCLUSION

The different methodologies for phylogenetic analysis and their specific algorithms on how they function were described in this review paper. After reviewing the benefits and drawbacks of each method, one can select the most suited approach for tree construction. The significance and applicability of this methodology in the field of biology is also discussed, and the paper concludes by outlining the recognised issues that should be addressed when constructing trees, and in future might these problems be excluded too.

VII.ACKNOWLEDGEMENT

All the authors gathered the data, analysed it and designed and contributed to the final manuscript.

VIII.CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

[1] A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. (1994). Molecular Biology and Evolution.

[2] EMBL-EBI (n.d.). Why is phylogenetics important? | Phylogenetics. [online] Available at: https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/why-is-phylogenetics-important/#:~:text=Phylogenetics%20is%20important%20because%20it.

[3] Felsenstein, J. and Sinauer Associates (2008). Inferring phylogenies. Sunderland, Mass.: Sinauer Associates, Inc., [Ca], Cop.

[4] Geosciences LibreTexts. (2019). 4.2: Limitations of Phylogenetic Trees. [online] Available at: https://geo.libretexts.org/Courses/University_of_California_Davis/UCD%3A_Geobiology_(Sumner)/Text/4%3A_Tracing_the_History_of_Evolution/4.2%3A_Limitations_of_Phylogenetic_Trees.

[5] Introduction to Phylogenetic Analysis. (n.d.). [online]. Available at: https://bip.weizmann.ac.il/education/course/introbioinfo/03/lect12/phylogenetics.pdf.

[6] Larson, D. (2021). Understanding the Importance of Phylogenetic Classification - Evolution. [online] Fossil Hunters. Available at: https://www.fossilhunters.xyz/evolution/understanding-the-importance-of-phylogenetic-classification.html [Accessed 7 Jun. 2021].

[7] Mount, D.W. and Harbor, S. (2006). Bioinformatics: sequence and genome analysis. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press, [Post], Cop.

[8] Müller, Sebastian. (2015). Re: What are the criteria for selecting a sequence for constructing a phylogenetic tree? Available at: https://www.researchgate.net/post/What_are_the_criteria_for_selecting_a_sequence_for_constructing_a_phylogenetic_tree/54eae8d0d4c118b0098b468b/citation/download.

[9] Munjal, G., Hanmandlu, M. and Srivastava, S. (2019). Phylogenetics Algorithms and

Applications. Advances in Intelligent Systems and Computing, pp.187–194.

[10] Nixon, K.C. (2001). Phylogeny. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780123847195001088.

[11] Phylogeny - Taxonomic systems | Britannica. (2019). In: Encyclopædia Britannica. [online] Available at: https://www.britannica.com/science/phylogeny/Taxonomic-systems.

[12] Ramirez-Flandes, Salvador. (2015). Re: What are the criteria for selecting a sequence for constructing a phylogenetic tree?. Available at: https://www.researchgate.net/post/What_are_the_criteria_for_selecting_a_sequence_for_constructing_a_phylogenetic_tree/54e9ce0ed4c118d96a8b465b/citation/download

[13] Rzhetsky, Andrey & Nei, Masatoshi. (1991). A Simple Method for Estimating and Testing Minimum-Evolution Trees. Molecular Biology and Evolution. 9. 10.1093/oxfordjournals.molbev.a040771.

[14] Saitou2, N. and Nei, M. (n.d.). The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees'. [online]. Available at: http://www.saitou-naruya-laboratory.org/assets/files/pdf/Saitou_MBE87.pdf [Accessed 6 Jun. 2021].

[15] Sarma, H., Pradhan, S., Mattaparthi, V.S.K. and Kaushik, S. (2019). Phylogenetic Analysis: Early Evolution of Life. Encyclopedia of Bioinformatics and Computational Biology, pp.938–952.

[16] Sciencedirect.com. (2014). UPGMA - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/agricultural-and-biologicalsciences/upgma.

[17] Scq.ubc.ca. (2006). INTRODUCTION TO PHYLOGENETICS. | SCQ. [online] Available at: https://www.scq.ubc.ca/introduction-to-phylogenetics/.

[18] Shreya Feliz (2017). UPGMA. [online] Available at: https://www.slideshare.net/shreyacathe/upgma [Accessed 6 Jun. 2021].

[19] Singh, R. (n.d.). 404 – Not Found. [online] Bioinformatics India. Available at: https://bioinformaticsindia.com/phylogenetic-reconstruction-tree-analysis/ [Accessed 6 Jun. 2021].

[20] Soltis, D.E. and Soltis, P.S. (2003). The Role of Phylogenetics in Comparative Genetics. Plant Physiology, [online] 132(4), pp.1790–1800. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC526274/.

[21] Theoretical foundation of the minimum-evolution method of phylogenetic inference. (1993). Molecular Biology and Evolution.

[22] Treehozz.com. (2020). What is neighbor joining method? [online] Available at: https://treehozz.com/what-is-neighbor-joining-method [Accessed 6 Jun. 2021].

[23] Wikipedia. (2021). Computational phylogenetics. [online] Available at: https://en.wikipedia.org/wiki/Computational_phylogenetics#Fitch-Margoliash_method [Accessed 6 Jun. 2021].

[24] Wikipedia. (2021). Neighbor joining. [online] Available at: https://en.wikipedia.org/wiki/Neighbor_joining#Advantages_and_disadvantages [Accessed 6 Jun. 2021].

[25] www.bionity.com. (n.d.). Distance_matrices_in_phylogeny. [online] Available at: https://www.bionity.com /en/ encyclopedia/Distance_matrices_in_phylogeny.html [Accessed 6 Jun. 2021].

[26] www.cs.gettysburg.edu. (n.d.). Building Fitch-Margoliash Phylogenetic Trees. [online] Available at: http://www.cs.gettysburg.edu/~ilinkin/projects/bio/phylo-fm/fm.html [Accessed 7 Jun. 2021].

[27] www.sequentix.de. (n.d.). Neighbor-Joining Method. [online] Available at: https://www.sequentix.de/gelquest/help/neighbor_joining_method.htm.

[28] www.sequentix.de. (n.d.). UPGMA Method. [online] Available at: https://www.sequentix.de/gelquest/help/upgma_method.htm [Accessed 6 Jun. 2021].