Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Review Article **DOI: 10.26479/2022.0801.05**

EXPLORING BASIC BIOINFORMATIC TOOLS FOR DNA SEQUENCE ANALYSIS

Sheeba Madaan*, Amit Pandey, Ranjana Juwantha, Shailesh Pandey

Forest Pathology Discipline, Forest Protection Division,

Forest Research Institute, Dehradun, India.

ABSTRACT: Advance genomic studies lead to the opening of a new field of science to biologists. The field is a combination of Information Technology and Biology called bioinformatics. Recent development in DNA sequencing methods has led to the availability of an enormous amount of DNA sequence data. Today new species identification is in rich demand. Thus, the DNA-based sequence analysis is a powerful tool for taxon identification and to determine the evolutionary relationships among microorganisms. There are numerous software's available for sequence data analysis that needs to be understood completely for accurate interpretation of results. This article devotes special attention to three important bioinformatics tools, viz., BLAST (Basic Local Alignment Search Tool), MEGA7 (Molecular Evolutionary Genetics Analysis), and DnaSP (DNA Sequence Polymorphism) for analyzing sequence data. Further, the potential application of these bioinformatic tools in taxonomy, phylogeny, and systematics is also discussed.

Keywords: Bioinformatic tool, Sequence data, BLAST, MEGA7, DnaSP.

Article History: Received: Jan 10, 2022; Revised: Jan 22, 2022; Accepted: Jan 30, 2022.

Corresponding Author: Ms. Sheeba Madaan* Ph.D.

Forest Pathology Discipline, Forest Protection Division, Forest Research Institute,

Dehradun, India. Email Address: sheeba.madaan@yahoo.in

1. INTRODUCTION

Bioinformatics is a multifaceted discipline analyzing biological information by combining many scientific fields including computational biology, statistics, mathematics, molecular biology, and genetics. It includes the analysis and interpretation of various biological data including nucleotides, amino acids, protein motifs, domains, and structures. This field also involves the compilation and

storage of biological data generated across the globe in an orderly and controlled manner. Moreover, the development of new algorithms and statistics contributed significantly to interpreting biological information with higher accuracy. Bioinformaticists mainly focused on three aspects: experimental design, statistical analysis, and data visualization. These tools were created to develop methods able to analyze and integrate complex omics (genomics, transcriptomics, proteomics, and metabolomics) datasets. With the advent of easy DNA sequencing facilities, reports on the number of sequences have increased at a high rate in the last decade. Importantly, the improvements in the field of the central processing unit and disk storage capacity, which allow faster computation and better data storage, revolutionized the methods to access and exchange the data. This specifically explains the need for bioinformatics to store, annotate, analyze and understand biological information that has been emerging from the omic approaches [1]. There are numerous software's available for sequence data analysis that needs to be understood completely for accurate interpretation of results. Given these facts, the main objectives of this study are: (1) to become familiar with three important bioinformatics tools, viz., BLAST (Basic Local Alignment Search Tool), MEGA7 (Molecular Evolutionary Genetics Analysis), and DnaSP (DNA Sequence Polymorphism); (2) to analyze the DNA sequence data in an appropriate and user-friendly manner; (3) to interpret the taxonomy, phylogeny, and systematics of microbes.

COMPUTATIONAL SOFTWARE INVOLVED IN DATA ANALYSIS

There are numerous websites available for accessing and analyzing DNA sequence data (Table. 1). The following section describes the three important tools containing fast computational algorithms and statistical methods to conduct efficient data analyses.

BLAST (Basic Local Alignment Search Tool)

It is a sequence similarity search program that can be run alone or through a public interface site (http://www.ncbi.nlm.nih.gov/blast). This program compares the query sequence (nucleotide or amino acid) with nucleotide or protein databases [2]. The main objective behind this tool is to find a similarity sequence against a query sequence and interpret biological information about the organism/ species. It is a universal tool applied in functional genomics research useful in annotation, visualization, and analysis for samples [3]. It is an intuitive and interactive desktop application that allows monitoring and comprehension of the whole annotation and analysis process. It is a user-friendly, easy to distribute, and low-maintenance tool. This program uses specific scoring matrices (like PAM or BLOSSOM) for performing sequence-similarity searches against a variety of sequence databases, to give us high-scoring ungapped segments among related sequences. The BLAST algorithm is fast, accurate, and web-accessible. It is relatively faster than other sequence similarity search tools and it provides us with the ability to perform analysis by different types of programs. The analysis of data includes: entering query sequence, selecting the database to search, running blast search, analyzing the output, and interpreting E-values (Fig. 1). This is a powerful tool for

sequence comparison studies. For instance, suppose we aim to determine the identity of unknown fungi based on DNA sequence data. To achieve this, the BLAST tool compares similarity percentages based on Multiple Sequence Alignment of sequences available in the nucleotide database. We can see whether the sequence, which we have sequenced, is homologous (similar) or not with any of the sequences in the database of interest.

MEGA 7.0.14 version (Molecular Evolutionary Genetics Analysis)

The molecular evolutionary genetics analysis (MEGA) software is a desktop application for comparing DNA and protein sequences. This is one of the most common tools to infer the evolutionary relationship among organisms. It infers both evolution and statistical properties of genes estimating neutral and selective evolutionary divergences among sequences [4]. It is mainly relying upon:(1) performance of multiple alignments; (2) exporting of alignment in MEGA format, and (3) running distance methods or phylogeny tree construction using alignment where anything in yellow is a default parameter that can be changed. For analysis of data in the MEGA window, first, go to the alignment option. Then go to query databanks, which will directly open the NCBI site. On the same site option displaying + ADD TO ALIGNMENT type nucleotide of the desired gene (for example- "ITS *Trichoderma* which will directly open its sequence studied anywhere) by clicking on "Search", user get so many results, and on the desired result (click on fasta) after which one get DNA sequence; they are at the right side of the window, the user has an option to 'Run Blast'. Then, just copy the whole sequence and then click on 'RunBLAST'. So yes in script error (if some dialog box comes up) and paste at query sequence (copied on), click on the database for others and somewhat similar sequences in program selection of nBLAST after which one would eventually get the result (Fig.2). MEGA is popularly used by evolutionary biologists to construct the phylogenetic tree and study the evolutionary relationships among the organisms. To look for coding regions, a user may right-click on the name of sequence and then click on 'refer to GenBank', a new page will be open. Repeat this sequence for all. After this, scroll down and look for CDS (coding sequence). In alignment explorer window of MEGA, on the left side downwards, users have site option, add base number where that coding region starts e.g., 228, make sure users click somewhere in the sequence of the species which is in use, and then press enter. The user gets to know about the initial codon, i.e, ATG, which means that before this is junk DNA, so click one base before this (start codon) and a user can backspace/hold shift key and find out home button in the keypad. This will ultimately select all bases before that, press the 'delete' button. Now, the sequence starts with the start codon for this. Scroll towards the right (at end of sequences). So now gaps were seen by a user which was indicative of extra pieces of genes stuck over their evolutionary times which were truncated, otherwise, it is going to confuse the computer. Now, upstairs white boxes were seen, click on the desired white box which will select the desired sequence in that position vertical direction; Hold shift and end of the keypad which will select everything, press the 'delete'

button which will truncate everything and now you have even alignment. Thus, users can see now identical sequences among all. Now users can perform CLUSTAL W alignment (compare every 2 organisms' sequences by pairwise sequence alignment and all organisms by multiple sequence alignment) by clicking on the option. So, in white boxes, the user who has a star will show that nucleotide bases were similar, and somewhere user who has nothing (no star) will have some differences in nucleotides. Space in between shows that there is either insertional mutation over evolutionary times. So thus, particular species have inserted new bases during evolution or deletion mutation (lost 3 bases in their genome) during evolution. Similarly, at the above of this window, the user can have an option for Translated Protein Sequence, click on that, and see for protein sequence. So now users can have a screenshot of aligned sequence and while closing users can save your file whenever want, SAY YES, Save in Mega file Input title again. Click on O.K. Open in MEGA (view sequence data). Now, we can minimize and lower the same window and go for other options. Similarly, after alignment now go for the phylogeny option. There, click on Bootstrap test of phylogeny, Create an evolutionary tree and statistics by NJ (Neighbor-joining) method, Click on compute, There user have tree showing closely relatedness among different species with different ways of representation, now in view option (use best traditional way i.e- rectangular). Either save by going in image option (as EMF. or TIFF. Format) or copy to clipboard and then open in Powerpoint and paste it here. Right-click (save as picture) as JPEG. wherever applicable. Thus, phylogenetic tree construction using MEGA can be done.

MEGA is an integrated tool for conducting:

- Sequence alignment
- Inferring phylogenetic trees
- Estimating divergence times
- Estimating rates of molecular evolution
- Inferring ancestral sequences
- Testing the evolutionary hypothesis

DnaSP (DNA Sequence Polymorphism)

The DNA sequence polymorphism is a powerful software package for the analysis of DNA polymorphism data [5]. It measures DNA sequence variation within and between the populations in non-coding, synonymous or non-synonymous sites [6] It computes linkage disequilibrium [7] recombination, gene flow, gene conversion parameters, and Neutrality tests [8] like- Fu & Li's [9], Hudson, Kreitman & Aguadis [10], McDonald & Kreitman [11] and Tajima's test [12]. It also conducts computer simulations based on the coalescent process [13]. The Input data file for this program is MSA (Multiple Sequence Alignment) or multi- MSA file formats e.g.- .fa generated by stacks [14], .alleles & .loci , .VCF & .gVCF [15]. DnaSP is advantageous through Microsoft windows' capability for handling a large number of sequences of thousands of nucleotides each on

a microcomputer. Along with these, it can easily exchange data with other programs to perform multiple sequence alignment (MSA), phylogenetic tree analysis, or statistical analysis. Besides these, some functions like- sequence alignment, making phylogenetic inferences or manipulating trees, editing or manipulating DNA sequences, and diploid genetic information analysis was not performed by Dna-SP.

STATISTICS IN PHYLOGENY STUDIES

Phylogeny is the evolution of a genetically related group of organisms or a study of relationships between a collection of 'things' (genes, proteins, organs…) that are derived from a common ancestor [16]. The phylogenetic tree/ dendrogram is an illustration of the evolutionary relationships among a group of organisms [17]. The making of an accurate phylogenetic tree [18] requires three conditions to be successful- 1) Rarely changing characters, which creates the low chance of homoplasy due to reversals; 2) Knowledge of the characters and organisms, which recognize homoplasy due to convergence; and molecular data will build so many characters where homoplasy is overwhelmed by a majority of good data 3) Knowledge of tree reconstruction and comparison methods and tools. Minimum evolution is the basic principle for phylogenetic inference [19] There are many criteria for getting the right tree-

1) Tree reconstruction criteria- It includes two techniques:

a) Exhaustive search- here; make all possible trees and compare them to pick the best one (Fig. 4). In biological studies, the evolutionary relationship between a few species is often represented by a rooted tree (Table. 2) with labeled leaves. The leaves represent the species and the internal vertices represent the ancestors. Such trees are also used in a biogeographical analysis to represent the biological relationship between the geographical areas.

$$\text{Rooted trees} = \frac{(2n-3)!}{2^{n-2} * (n-2)!}$$

Where n= number of taxa

b) Using an algorithm to make a tree- There are many algorithms [20] available that compute the distances between all the taxa, connect the two closest and call them a new taxon. Also, recompute distances between the reduced set of taxa connect the two closest from the new set. Start with a good tree and swap branches randomly, testing each swap if the tree improves by criterion used.

2) Tree comparison criteria- It is based on:

a) Parsimony

The more parsimonious tree means the set of more linearly arrayed sequences [21]. The tree with the fewest evolutionary events is the most likely one which assumes that evolutionary events are rare. Therefore, when we try to reconstruct the past, the reconstruction with the fewest postulated events is the most likely to be accurate; Similar to the principle of Occam's razor, given two possible explanations, the simplest one is the most likely. For the same extant, taxa and ancestral state (Fig.

5). Therefore, the more preferred tree (Fig. 5a) will be more parsimonious described with less evolutionary event.

b) Model-based approaches (computer intensive)

In the realm of empirical data from actual taxa, it is not known (or knowable) how commonly Maximum-parsimony (MP), Maximum-likelihood (ML) or Bayesian-inference (BI) are inaccurate. To test the perceived need for ‘‘sophisticated’’ model-based approaches, we assessed the degree of congruence between empirical phylogenetic hypotheses generated by alternative methods applied to DNA sequence data [22]. Assume a specific model (potentially very complicated) for the probability of evolutionary events occurring and then compare trees using these models.

The two approaches are mainly used by professionals when they are really studying phylogenies.

i. ML (Maximum Likelihood) method

A maximum likelihood approach to estimate the evolutionary trees from nucleic acid sequence data is discussed. This method allows the testing of hypotheses about the constancy of evolutionary rates by likelihood ratio tests and gives a rough indication of the error for the estimate of the tree [23]. PAML is a program used to compare DNA and protein sequences and test phylogenetic trees, but their main strengths lie in the rich repertoire of evolutionary models implemented, which could be used to estimate parameters in models of sequence evolution and to test interesting biological hypotheses [24]

ii. Bayesian method

The Bayesian method is based on Markov chain simulation to study the phylogenetic relationship in a group of DNA sequences. This method is advantageous in providing estimates and corresponding measures of variability for any aspect of the phylogeny [25]. It has facilitated the exploration of parameter-rich evolutionary models. Bayesian Markov chain Monte Carlo (MCMC) analysis deals efficiently with complex models: convergence occurs faster and more predictably for complex models, mixing is adequate for all parameters even under very complex models, and the parameter update cycle is virtually unaffected by model partitioning across sites [26]. A new method for approximate Bayesian statistical inference on the basis of summary statistics is proposed. It is suited to complex problems that arise in population genetics. The method combines many of the advantages of Bayesian statistical inference with the computational efficiency of methods based on summary statistics [27] Statistics play an important role in phylogeny tree construction and evolutionary relationships between different members of the clade. It uses the probability that we would get the results we observe purely due to random factors. When the probability is low is when we accept non-random factors. Testing an observed pattern against randomness is the central idea in statistics. We want a way to see how randomness (stochastically) influences the patterns we see. The standard statistical technique to determine the validity of a tree involves re-sampling our data to see how robust/repeatable the tree is; i.e. - do we get the same tree each time. There are two

statistical methods- Bootstrapping and Jackknifing present for evaluation and distinguishing the confidence of partial hypothesis i.e- branch support present in the phylogenetic tree which is a standard method during molecular phylogenetic studies. Bootstrapping is a kind of statistical analysis to place confidence intervals on phylogenies and to test the reliability of certain branches in the evolutionary tree. In the context of tree building according to bootstrapping, each pseudo-replicate is constructed by randomly sampling columns of the original alignment with replacement until an alignment of the same size is obtained. They have used bootstrap to have confidence values for each clade in the estimation of phylogenetic trees [28]. This method is used to estimate the confidence intervals of a population mean by randomly resampling a subset of data from within a larger data set [29]. Based on computer simulations and a laboratory-generated phylogeny, bootstrapping results of parsimony analyses are used to test both measures of repeatability and accuracy [30]. Here, the replacement will create a new data set where the number of characters was picked up from the set which may duplicate some and omit others, while conceptually simulating robustness to gather even more data. Jackknifing is a statistical method of numerical resampling or sub-sampling without replacement to create a smaller data set based on deleting a portion of the original observations for each pseudo-replicate. Here, picking the number of characters from the set will omit some gathering of fewer data compare to bootstrapping. It is seen that to create pseudo-replicate, half of the columns were deleted from alignment in case of 50% jackknifing. Thus, it has become an important method to estimate support; and gives suggestions regarding the result of an infinitely repeated bootstrap analysis argues for its power [31]. It was seen that a fast tree construction approach without branch swapping is possible through the first jackknife analysis program while subsequent programs allowed a more extensive tree search. Due to more options available, question comes that what should be the optimal search strategy in jackknife analysis. Earlier studies [32] has given suggestions regarding these issues and have come to some conclusion, while the current data set provides an opportunity to extend our understanding of this procedure.

2. CONCLUSION

Bioinformatic tools are mainly used for the characterization of the gene, determination of a structure and physiological properties, phylogenetic analysis, and performing simulation to understand how the molecules behave in a living system. Thus, for all these purposes, a different type of tool is used e.g., if a user wants to know the physicochemical properties in different sequences which were analyzed based on the composition of different amino acids. Then, on that basis, one could derive a large number of information such as molecular mass, ionic strength, hydrophobicity, hydrophilicity, pI value, bulkiness, etc. Thus, the vast amount of data generated for which a user requires new approaches and that approach should have the high throughput assay. To access manually, a process requires a robotic system and high-speed computation adding great advancement in biological research. So to analyze this, we must need statistical information and bioinformatics actually help

us to understand the large amount of data created through biological research into a well-organized system to understand all this information. Gene sequence analysis helps to understand the gene product or biology. A collection of sequences doesn't by itself, increase the scientist's understanding of the biology of organisms. Sequence analysis can be used to assign a function to genes and proteins by the study of the similarities between the compared sequences [35] Genomics deals with the analysis and comparison of the entire genome of single and multiple spaces. It existed before any genomes were completely sequenced but in a very primitive stage. Thus, bioinformatics plays an important role in the field of genomics. Bioinformatics helps to acquire, manage, analyze and understand the data. The bioinformatics tools available over the internet are accessible, well developed, fairly comprehensive, and relatively easy to use. It has made it possible to test our hypothesis virtually and therefore allows us to take a better and an informed decision before launching costly experimentations. Thus, the in-silico analysis will reduce your cost, time, and energy and sequence comparison could be easily done. With the evident advancement in computational hardware and software, most of the drug discovery steps are now performed. Several software is available to perform homology modeling to get a fast structural solution of a given protein sequence as going through instruments (GC-MS / NMR) is very difficult to analyze things. It is becoming more and more difficult to do healthcare research and deliver effective healthcare without the contribution of bioinformatics. In terms of research, bioinformatics contributes by developing new methods, new ways of looking at data, new ways of being able to store data efficiently and access data, and visualizing that to make it meaningful for people. Therefore, bioinformatics helps the nascent biological field and the application of the bioinformatics tools helps a large amount of information derived from a large number of biological datasets. The future challenges for bioinformatics are the integration of more annotation sources into workflows and predicting the function of variants outside of coding regions.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The author confirms that the data supporting the findings of this research are available within the article.

FUNDING

None.

ACKNOWLEDGEMENT

The financial support by the Indian Council of Forestry Research and Education, Dehradun is gratefully acknowledged.

CONFLICT OF INTEREST

No conflict of interest exists.

REFERENCES

1. Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. Bioinformatics for omics Data. 2011:3-0.
2. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic acids research. 2008 Apr 24;36(suppl_2):W5-9.
3. Spaniol C. Integrated Analysis and Application Pipelines for Complex Disease Data.
4. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in bioinformatics. 2008 Jul 1;9(4):299-306.
5. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003 Dec 12;19(18):2496-7.
6. Rozas J, Rozas R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics (Oxford, England). 1999 Feb 1;15(2):174-5.
7. Rozas J, Gullaud M, Blandin G, Aguadé M. DNA variation at the rp49 gene region of Drosophila simulans evolutionary inferences from an unusual haplotype structure. Genetics. 2001 Jul 1;158(3):1147-55
8. Rozas J. DNA sequence polymorphism analysis using DnaSP. bioinformatics for DNA sequence analysis 2009 (pp. 337-350). Humana Press.
9. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993 Mar 1;133(3):693-709.
10. Kreitman M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature. 1983 Aug;304(5925):412-7.
11. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991 Jun;351(6328):652-4.
12. Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989 Nov 1;123(3):597-601.
13. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular biology and evolution. 2017 Dec 1;34(12):3299-302
14. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis toolset for population genomics. Molecular ecology. 2013 Jun;22(11):3124-40

15. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8.
16. Brooks DR, McLennan DA, McLennan DA. Phylogeny, ecology, and behavior: a research program in comparative biology. University of Chicago Press; 1991.
17. Avise JC. Molecular markers, natural history, and evolution. Springer Science & Business Media; 2012 Dec 6.
18. Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approach es. Nature reviews genetics. 2003 Apr;4(4):275-84
19. Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. International Workshop on Algorithms in Bioinformatics 2002 Sep 17 (pp. 357-374). Springer, Berlin, Heidelberg
20. Gusfield D. Algorithms on strings, trees, and sequences: Computer science and computational biology. Acm Sigact News. 1997 Dec 1;28(4):41-60.
21. Fitch WM. On the problem of discovering the most parsimonious tree. The American Naturalist. 1977 Mar 1;111(978):223-57.
22. Rindal E, Brower AV. Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data. Cladistics. 2011 Jun;27(3):331-4.
23. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981 Nov 1;17(6):368-76
24. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 2007 Aug 1;24(8):1586-91.
25. Li S, Pearl DK, Doss H. Phylogenetic tree construction using Markov chain Monte Carlo. Journal of the American Statistical Association. 2000 Jun 1;95(450):493-508.
26. Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey J. Bayesian phylogenetic analysis of combined data. Systematic biology. 2004 Feb 1;53(1):47-67.
27. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002 Dec 1;162(4):2025-35.
28. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. evolution. 1985 Jul;39(4):783-91.
29. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC Press; 1994 May 15.
30. Hillis DM, Bull JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic biology. 1993 Jun 1;42(2):182-92.
31. Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG. Parsimony jackknifing outperforms neighbor-joining. Cladistics. 1996 Jun 1;12(2):99-124.

32. DeBry RW, Olmstead RG. A simulation study of reduced tree-search effort in bootstrap resampling analysis. Systematic Biology. 2000 Mar 1;49(1):171-9.
33. Mort ME, Soltis PS, Soltis DE, Mabry ML. Comparison of three methods for estimating internal support on phylogenetic trees. Systematic Biology. 2000 Mar 1;49(1):160-71.
34. Salamin N, Hodkinson TR, Savolainen V. Towards building the tree of life: a simulation study for all angiosperm genera. Systematic biology. 2005 Apr 1;54(2):183-96.
35. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic acids research. 1998 Jan 1;26(1):320-2