Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/66/2024.17991

How can statistical models improve the accuracy of phylogenetic tree reconstruction?

Huiyang Shi

USA College of Art and Science, Boston University, USA

shiliu@bu.edu

Abstract. When building the phylogenetic tree, current methods for calculating phylogenetic trees often lack the desired level of accuracy. This research paper explores the role of statistical models in enhancing the precision of phylogenetic tree reconstruction. The phylogenetic tree plays an important role in phylogenetic analysis and evolutionary biology. An accurate phylogenetic tree underpins our understanding of the major transitions in evolution, such as the emergence of new body plans or metabolism. Recent advancements in statistical modeling have more elaborate improvements than traditional methods. For example, Bayesian inference and maximum likelihood estimation provide frameworks for evaluating phylogenetic relationships by incorporating probability distributions and likelihood functions. These models account for the inherent uncertainties and variations in genetic data, leading to more accurate tree constructions. Moreover, using statistical models allows for the incorporation of complex evolutionary processes such as variable rates of evolution across lineages, horizontal gene transfer, and hybridization events. Techniques like Markov Chain Monte Carlo (MCMC) and bootstrap methods enhance the reliability of the inferred phylogenies by providing measures of confidence for the estimated relationships. Using these advanced statistical approaches, researchers can gain more accurate and reliable phylogenetic trees. Through a comprehensive review of current methodologies and case studies, this study aims to highlight statistical models' significant impact on evolutionary biology and the broader field of phylogenetics analysis and evolutionary biology.

Keywords: keywords. phylogenetics trees, statistical method, Bayesian inference, MCMC, bootstrap.

1. Introduction

A phylogenetic tree is a graphical representation of the evolutionary relationships between biological entities [1].

Phylogenetic trees are important tools for researchers studying the evolutionary relationships between species. These trees depict the branching patterns of evolution, and they demonstrate how different species have evolved from a series of common ancestors. For instance, phylogenetic analyses have shown that birds are descendants of dinosaurs, based on shared physical traits such as hip bones and skull structures. This capacity to reveal evolutionary links makes phylogenetic trees largely promote the study of biodiversity and evolutionary history.

When building phylogenetic trees, it is key to inferring the origin of new genes, detecting molecular adaptation, understanding morphological character evolution, and reconstructing demographic changes

DOI: 10.54254/2753-8818/66/2024.17991

in recently diverged species. Although data are ever more plentiful and powerful analysis methods are available, there remain many challenges to reliable tree building [2]. Variable rates of evolution across different lineages can complicate the reconstruction of these trees because some species might evolve faster or slower than others. Additionally, horizontal gene transfer, where genes are exchanged between unrelated species, can blur the lines of descent and can create misleading genetic signals. Besides, hybridization events, where distinct species interbreed to produce hybrid offspring, make the process more complicated. Hybridization introduces new genetic combinations, so traditional methods, such as distance-based methods, might not account for them effectively.

To overcome these challenges, researchers have increasingly turned to statistical models. These models provide robust frameworks to address the intricacies of evolutionary processes. By incorporating variability in evolutionary rates and accounting for phenomena like horizontal gene transfer and hybridization, statistical models offer improved tools for phylogenetic analysis.

2. Discussion and Result

2.1. Bayesian inference

Bayesian inference (/ˈbeɪziən/ BAY-zee-ən or /ˈbeɪʒən/ BAY-zhən) [3] is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference derives the posterior probability from two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data. In the context of phylogenetic tree reconstruction, Bayesian inference is employed to infer the most likely phylogenetic tree given the observed data, such as DNA sequences.

The process begins with selecting a model of evolution. This model describes how DNA, RNA, or protein sequences evolve and includes parameters such as substitution rates and base frequencies. Once the model is chosen, a prior distribution is established, representing the initial beliefs about the model parameters and the possible phylogenetic trees. Priors can be informed by previous knowledge or be non-informative if no prior information is available.

Next, the likelihood of the observed data given a particular phylogenetic tree and model parameters is calculated. This step involves computing the probability of observing the sequence data under the specified model of evolution for each possible tree. The likelihood function quantifies how well a particular tree explains the observed data.

Bayesian inference then applies Bayes' theorem to update the prior distribution with the likelihood of the observed data. This results in a posterior distribution, representing the probability of the phylogenetic tree and model parameters given the observed data. Mathematically, this relationship is expressed as:

$$P(\theta \mid \text{data}) \propto P(\text{data} \mid \theta) \cdot P(\theta)$$

where

- $P(\theta \mid \text{data})$ is the posterior probability,
- $P(\text{data} \mid \theta)$ is the likelihood,
- $P(\theta)$ is the prior probability.

By integrating probability theory and Bayes' theorem, Bayesian inference provides a comprehensive framework for estimating phylogenetic trees, allowing for the incorporation of uncertainty and the combination of various sources of information. This method yields a posterior distribution of phylogenetic trees, offering a probabilistic view of evolutionary relationships that accounts for the complexity of the underlying evolutionary processes.

Bayesian inference of phylogeny brings a new perspective to a number of outstanding issues in evolutionary biology, including the analysis of large phylogenetic trees and complex evolutionary models and the detection of the footprint of natural selection in DNA sequences [4].

DOI: 10.54254/2753-8818/66/2024.17991

2.2. Maximum likelihood estimation

Maximum Likelihood Estimation (MLE) is a statistical method used to infer phylogenetic trees by identifying the tree topology, branch lengths, and model parameters that maximize the likelihood of the observed sequence data. Maximum likelihood estimate (MLE) is also the most commonly used technique in statistical inference, holding many good statistical properties such as consistency and asymptotic normal distribution. [5] The process involves several steps, each aimed at determining the most probable evolutionary relationships among species or genes based on observed data.

First, a model of evolution is selected. This model describes how DNA, RNA, or protein sequences evolve and includes parameters such as substitution rates and base frequencies. With this model in place, the next step is to calculate the likelihood of the observed sequence data for a given tree topology and set of branch lengths. This calculation requires defining a probabilistic model of sequence evolution along the branches of the tree and using it to determine the probability of observing the sequences at the tips of the tree (i.e., the extant species or genes) given the sequences at the internal nodes (i.e., the ancestors).

The likelihood is computed using the Felsenstein algorithm, which efficiently calculates the probability of the data given the tree by summing over all possible ancestral states. The algorithm enables the calculation of the likelihood by considering the contribution of each branch and internal node to the overall probability.

The goal of MLE is to find the tree topology and branch lengths that maximize this likelihood. This involves searching through the space of possible tree topologies and optimizing the branch lengths and model parameters for each topology. Because the likelihood function is complex and does not have a closed-form solution, this optimization is typically performed using numerical methods.

The tree with the highest likelihood is selected as the maximum likelihood tree. This tree represents the most probable evolutionary relationships among the species or genes, given the observed data and the chosen evolution model. By maximizing the likelihood, MLE provides a robust framework for phylogenetic tree reconstruction, ensuring that the inferred tree is the most consistent with the observed sequence data and the evolutionary model used.

2.3. MCMC

Although methods such as Bayesian inference or maximum likelihood are popular in genetics and bioinformatics, these methods become largely unpredictable when the size of the data set increases. Thus, a new and more elaborate method, the jump Markov chain Monte Carlo (MCMC) algorithm, is used. MCMC has been one of the most important and popular concepts in Bayesian Statistics, especially while doing inference. [6] These algorithms are applied to artificial and high dimensional scenarios, but also to the classic mine disaster dataset inference problem. [7] The process starts with an initial tree, which could be a random tree, or one based on a simple heuristic method. At each step in the chain, a new tree is proposed by making a small change to the current tree. This change could involve rearranging branches, modifying branch lengths, or altering substitution model parameters. The proposed tree is accepted or rejected based on the Metropolis-Hastings acceptance criterion, first, calculate the likelihood of the new tree and its prior probability, and then calculate the acceptance ratio, which is the ratio of the posterior probability of the new tree to the current tree.

Accept the new tree with this ratio's probability. If the new tree is more probable, it is always accepted. If it is less probable, it is accepted with a probability proportional to the ratio; otherwise, the current tree is retained. After an initial "burn-in" period, where the chain may not yet be sampling from the true posterior distribution, the chain reaches a stationary distribution. At this point, the trees generated are samples from the posterior distribution. The MCMC process continues for many iterations, generating many trees. These samples represent the posterior distribution of trees. The sampled trees are used to construct a summary tree, often a consensus tree, which represents the most likely phylogenetic relationships. Additionally, posterior probabilities for each clade (branch) in the tree can be calculated, providing a measure of support for each relationship.

DOI: 10.54254/2753-8818/66/2024.17991

MCMC is computationally intensive, requiring significant computational resources, especially for large datasets or complex models. However, its ability to provide detailed and probabilistic insights into phylogenetic relationships makes it a valuable tool in the field of evolutionary biology and phylogenetic analysis.

By employing MCMC, researchers can obtain a more nuanced and statistically robust understanding of evolutionary relationships, considering the complexity and uncertainty inherent in the data.

2.4. Bootstrap

Apart from the MCMC method, bootstrap analysis is also a powerful statistical method used to estimate the reliability of the branches in a phylogenetic tree when facing complex evolutionary relationships. Evolutionary trees are often estimated from DNA or RNA sequence data. How much confidence should we have in the estimated trees? In 1985, Felsenstein [Felsenstein, J. (1985) Evolution 39, 783–791] suggested the use of the bootstrap to answer this question. [8] This process involves several key steps to rigorously evaluate the robustness of inferred phylogenetic relationships.

Begin with an original multiple-sequence alignment of the sequences the researchers want to analyze. This alignment arranges the sequences in a matrix format, with each row representing a sequence and each column representing a position in the sequence.

Secondly, create many new datasets (typically 100-1000) by resampling columns (sites) from the original alignment with replacement. This means that each new dataset, or "bootstrap replicate," will have the same number of columns as the original alignment but may include some columns multiple times and omit others. This resampling process introduces variability while preserving the inherent structure of the data.

After that, for each bootstrap replicate, construct a phylogenetic tree using the same method employed for the original dataset. Common methods include Maximum Likelihood, neighbor-joining, or Bayesian inference. Each method has its approach to estimating evolutionary relationships, but the key is to apply the same method consistently across all replicates. The process is repeated many times to produce a distribution of possible trees. [9]

Then, for each branch in the original phylogenetic tree, count how many times that branch appears in the trees generated from the bootstrap replicates. This step involves mapping the branches of the original tree to those in each bootstrap replicate, noting the frequency of occurrence.

Lastly, the frequency with which each branch appears across all bootstrap replicates is expressed as a percentage, known as the bootstrap value or bootstrap support. For instance, if a particular branch appears in 950 out of 1000 replicates, its bootstrap value would be 95%. These values provide a quantitative measure of the reliability of each branch.

Finally, higher bootstrap values indicate greater confidence in the inferred relationships, suggesting that the grouping of sequences is consistently supported by the resampled data. Conversely, lower values indicate less confidence, suggesting that the relationships may be more sensitive to the specific data sampled.

By performing bootstrap analysis, researchers can assess the robustness of the inferred phylogenetic relationships, providing a more rigorous evaluation of the evolutionary history represented by the tree. This method is crucial for validating the reliability of phylogenetic trees, especially in studies involving complex evolutionary relationships or limited data. The bootstrap values are often visualized on the phylogenetic tree, with branches annotated to reflect their support, thereby enhancing the interpretability and credibility of the phylogenetic analysis.

3. Conclusion

Phylogenetic trees are fundamental tools in phylogenetic analysis and evolutionary biology, but the traditional methods of creating these trees can be overly simplistic, potentially reducing their accuracy and leading to errors. This paper discusses four advanced statistical methods that can significantly enhance the accuracy of phylogenetic tree reconstruction: Bayesian inference, maximum likelihood estimation, the Markov chain Monte Carlo (MCMC) algorithm, and bootstrap analysis.

DOI: 10.54254/2753-8818/66/2024.17991

Bayesian inference is a powerful approach that incorporates prior knowledge and updates the probability of a hypothesis as more evidence becomes available. This method provides a robust framework for estimating phylogenies, particularly when dealing with complex evolutionary scenarios. Maximum likelihood estimation, another highly effective model, evaluates different phylogenetic trees by calculating the probability of the observed data given a particular tree structure, thus identifying the tree that best fits the data.

For more intricate evolutionary processes, the Markov chain Monte Carlo (MCMC) algorithm offers a sophisticated means of sampling from the posterior distribution of phylogenetic trees. By exploring a vast range of possible tree configurations, MCMC can more accurately infer phylogenies that reflect the true evolutionary history. Bootstrap analysis further enhances reliability by assessing the robustness of each branch in the phylogenetic tree. By resampling the data and reconstructing trees multiple times, this method provides a measure of confidence for each inferred relationship.

Incorporating these advanced statistical methods into the process of phylogenetic tree reconstruction can greatly improve the accuracy and reliability of the resulting trees. By leveraging the strengths of Bayesian inference, maximum likelihood estimation, MCMC, and bootstrap analysis, researchers can achieve more precise and credible insights into evolutionary relationships, ultimately advancing our understanding of the tree of life.

References

[1] Richard J. Edwards, in Encyclopedia of Bioinformatics and Computational Biology, 2019

[2] Kapli, P., Yang, Z. & Telford, M.J. Phylogenetic tree building in the genomic age. Nat Rev Genet 21, 428–444 (2020). https://doi.org/10.1038/s41576-020-0233-0

[3] https://en.wikipedia.org/wiki/Bayesian_inference#:~:text=Bayesian%20inference%20(%2F%CB%88be%C9%AA, evidence%20or%20information%20becomes%20available.

[4] John P. Huelsenbeck et al., Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology.Science294, 2310-2314(2001).DOI:10.1126/science.1065889

[5] Purcell, S. "Maximum likelihood estimation." Accessed 05Jun2015. Available at: http://statgen. iop. kcl. ac. uk/bgim/mle/sslike_3. html (2007).

[6] https://towardsdatascience.com/monte-carlo-markov-chain-mcmc-explained-94e3a6c8de11

[7] Andrieu, Christophe, and Johannes Thoms. "A tutorial on adaptive MCMC." Statistics and Computing 18 (2008): 343-373.

[8] Efron, Bradley, Elizabeth Halloran, and Susan Holmes. "Bootstrap confidence levels for phylogenetic trees." Proceedings of the National Academy of Sciences 93.23 (1996): 13429-13429.

[9] Carey Krajewski, Allan W. Dickerman, Bootstrap Analysis of Phylogenetic Trees Derived from DNA Hybridization Distances, Systematic Biology, Volume 39, Issue 4, December 1990, Pages 383–390, https://doi.org/10.2307/2992358