RESEARCH PAPER

A Comprehensive Review on Phylogenetic Tree Construction: Algorithms, Optimization Techniques, and Applications

Amandeep Kaur[1] · *Manjot Kaur[2] · Navneet Kaur[3]

[1]PhD Scholar, School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

[23]Associate Professor, School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

*Corresponding author: Manjot Kaur, E-Mail: manu.sembhi@gmail.com

Abstract

Phylogenetics is a potent and integrative method in contemporary biology that brings molecular information to the evolutionary theory to construct the tree of life. The construction of phylogenetic trees attempts to find answers to some of the most fundamental questions regarding ancestry, divergence, and the passage of genetic information across time—a metaphor originally set forth by Charles Darwin. They are not mere diagrams in the contemporary scientific landscape, but explanatory hypotheses of evolutionary pathways built upon molecular evidence, statistical models, and computational methods. Emerging technologies in DNA and RNA sequencing, computational biology, and high-performance computing have significantly enhanced the accuracy and reliability of evolutionary reconstructions. This review summarizes the fundamentals of phylogenetic tree construction, covering molecular data types, sequence alignment methods, and tree formats, which lay a foundation for comprehending the range of algorithms and optimization techniques available. The review further examines practical applications in medicine, ecology, and biotechnology, demonstrating how phylogenetics has been transformed from a descriptive into a quantitative discipline. Finally, the paper describes current challenges—including computational scalability, alignment error, and horizontal gene transfer—along with promising future directions in evolutionary biology.

Keywords: DNA · RNA · Genome · Phylogenetic Tree · MEGA · Maximum Likelihood · Bayesian Inference · Optimization Algorithms

How to cite this article: Kaur A, Kaur M, Kaur N. A Comprehensive Review on Phylogenetic Tree Construction: Algorithms, Optimization Techniques, and Applications. Int J Drug Deliv Technol. 2026;16(40s): 965-986. DOI: 10.25258/ijddt.16.40s.98

1. Introduction

1.1 Origins of the Tree of Life Concept

The concept of phylogeny traces back to the nineteenth century, a period when evolutionary biology was emerging as a formal science. Charles Darwin, in the Origin of Species published in 1859, fundamentally altered the scientific understanding of biological diversity by proposing that all species share common origins through descent with modification. To illustrate this principle, Darwin drew a simple branching diagram, today regarded as the first published form of a phylogenetic tree, depicting the manner in which species diverge from common ancestors [1]. Building upon Darwin's foundation, Ernst Haeckel in the late nineteenth century developed detailed tree-of-life illustrations, organizing organisms into branching hierarchies on the basis of observable characteristics such as morphology and anatomy. Although these early diagrams incorporated no molecular evidence whatsoever, they were of considerable importance in establishing the tree metaphor as the principal scheme for understanding evolutionary history.

1.2 The Rise of Molecular Phylogenetics

The twentieth century witnessed developments of far-reaching consequence that transformed phylogenetics into a quantitative science. The determination of the structure of DNA's double helix by James Watson and Francis Crick in 1953, followed by the deciphering of the genetic code, provided a molecular basis for heredity and evolution, demonstrating that mutations in DNA and protein sequences serve as molecular records of evolutionary change [2]. The suggestion by Linus Pauling and Emile Zuckerkandl that sequence similarity and divergence among proteins could be used to establish relatedness and estimate separation times between taxa laid the groundwork for what later became known as molecular phylogenetics. The DNA sequencing techniques developed by Frederick Sanger in the 1970s and 1980s gave scientists the direct means to read and employ genetic material in phylogenetic analysis. Carl Woese's comparative

studies of ribosomal RNA sequences ultimately produced the three-domain system—Bacteria, Archaea, and Eukarya—demonstrating the power of molecular phylogenetics to reform biological classification.

1.3 The Computational Revolution

The necessity for computational approaches became evident as both DNA and protein sequences accumulated in increasing numbers. Early attempts at phylogenetic inference were manual and small in scale, but the emergence of computers in the late twentieth century transformed the field entirely. Algorithms such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean), Neighbor Joining, and Maximum Parsimony enabled researchers to construct trees from large volumes of data in a systematic and reproducible fashion. By the 1990s, Maximum Likelihood and Bayesian models had been developed, providing statistically rigorous means of testing evolutionary hypotheses [3]. The introduction of high-throughput sequencing, commonly referred to as next-generation sequencing (NGS), in the 2000s produced a volume of genomic data unprecedented in the history of biology. Whereas this opened the possibility of examining evolutionary processes at the level of whole genomes, it simultaneously demanded new optimization strategies and high-performance computing resources. The field has become increasingly interdisciplinary, with parallel computing, distributed cloud systems, and machine learning being progressively integrated into phylogenetic workflows.

1.4 Applications and Significance

Phylogenetics is now essential across a broad range of scientific and applied domains. In medicine, it is used to monitor viral epidemics including HIV and SARS-CoV-2. In ecology, it aids conservation efforts by identifying endangered lineages and quantifying biodiversity [4]. In biotechnology, it guides drug and enzyme design; in anthropology, it reveals patterns of human migration including evidence for the Out of Africa hypothesis. These applications collectively demonstrate the central role of phylogenetics in enabling molecular biology, computer science, and evolutionary theory to address questions of fundamental scientific and practical importance.

2. Fundamentals of Phylogenetic Tree Construction

2.1 Sources of Molecular Data

Phylogenetic tree construction begins with the collection of appropriate molecular data. Different data types provide unique insights into evolutionary relationships, and the choice of data is itself a critical methodological decision. Nuclear DNA offers a comprehensive and biparentally inherited source but is complicated by recombination. Mitochondrial DNA, transmitted maternally and characterized by high mutation rates, has been extensively employed to trace human migrations. Chloroplast DNA is particularly valuable in plant studies owing to its slow rate of evolution. Ribosomal RNA, owing to its high degree of conservation across all domains of life, remains a universal phylogenetic marker. Protein sequences, which reflect selection pressures at the functional level, are especially informative for reconstructing distant evolutionary relationships. Whole-genome approaches, although computationally intensive, provide the most comprehensive reconstruction of evolutionary history. Table 1 summarizes the principal advantages and limitations of these molecular data types.

Table 1. Advantages and Limitations of Molecular Data Types

Data Type	Advantages	Limitations	Example Use Case
Nuclear DNA	Comprehensive, biparental	Recombination complicates analysis	Human population genetics
Mitochondrial DNA	High mutation rate, maternal	Limited to maternal lineage	Tracing human migrations
Chloroplast DNA	Useful in plants, slow evolution	Limited scope	Plant phylogenies
Ribosomal RNA	Highly conserved, universal	Poor resolution at shallow scales	Three-domain classification
Protein Sequences	Conserved domains, functional	Requires translation from DNA	Distant relationships
Whole Genome	Most comprehensive	Computationally intensive	Phylogenomics

2.2 Sequence Alignment

Before constructing a phylogenetic tree, sequences must be aligned to identify homologous positions. Proper alignment ensures that insertions, deletions, and substitutions are accounted for, enabling meaningful evolutionary comparisons. Three principal alignment approaches are employed in practice.

Global alignment attempts to align entire sequences from end to end and is appropriate when sequences are similar in length and composition. Local alignment focuses on regions of high similarity and is suited to sequences that differ substantially in length. Multiple Sequence Alignment (MSA), which simultaneously aligns three or more sequences, represents the standard approach in phylogenetics, as it provides each sequence with its evolutionary context relative to the group.

Among the principal tools available, CLUSTALW is one of the earliest and most widely used MSA programs; MUSCLE offers improved accuracy and speed for large datasets; MAFFT is notable for its capacity to handle very large alignments with high accuracy; and T-Coffee employs consistency-based approaches specifically designed to reduce alignment errors [5].

Alignment is susceptible to error, particularly in sequences that are highly divergent. Homologous positions become difficult to establish when insertions and deletions are frequent, and repetitive regions introduce ambiguities that propagate through subsequent tree inference. Accordingly, careful alignment and validation are indispensable steps in any rigorous phylogenetic analysis.

2.3 Types of Phylogenetic Trees

Several distinct types of phylogenetic trees are employed depending on the analytical objective. Rooted trees specify an ancestral node and indicate the direction of evolutionary change. Unrooted trees demarcate connections among taxa without implying a common ancestor. Phylograms represent branch lengths in proportion to the rate of evolutionary change, whereas cladograms depict topology only, without branch length information. Consensus trees summarize the information contained in a distribution of trees, as produced by bootstrap or Bayesian analyses, and are particularly useful when multiple plausible tree topologies have been identified [6]. Table 2 provides a comparative overview of these tree types.

Table 2. Comparison of Phylogenetic Tree Types

Tree Type	Root?	Branch Lengths	Use Case Example
Rooted Tree	Yes	Optional	Evolutionary lineage tracing
Unrooted Tree	No	Optional	Similarity without ancestor assumption
Phylogram	Yes/No	Yes	Estimating divergence times
Cladogram	Yes/No	No	Taxonomic classification
Consensus Tree	Varies	Optional	Summarizing multiple inferences

2.4 General Workflow for Phylogenetic Tree Construction

The construction of a phylogenetic tree follows a sequence of well-defined steps: (i) data collection, in which appropriate molecular sequences are selected; (ii) alignment, in which MSA tools are applied to identify homologous positions; (iii) model selection, in which an appropriate evolutionary model is chosen; (iv) tree construction, in which one or more algorithmic approaches—distance-based, parsimony, likelihood, or Bayesian—are applied; (v) optimization, in which tree topology and branch lengths are refined; (vi) validation, in which bootstrap resampling or posterior probability estimates are used to assess confidence; and (vii) visualization, in which tools such as MEGA, RAxML, or MrBayes are employed to render and interpret the resulting tree.

2.5 Morphological versus Molecular Phylogenetics

Table 3. Comparison of Morphological and Molecular Phylogenetic Approaches

Aspect	Morphological Data	Molecular Data
Availability	Can be used even for fossils	Requires preserved genetic material
Resolution	Limited at deeper levels	High, across wide temporal scales
Limitations	Convergent evolution;	Requires sequencing;

Aspect	Morphological Data	Molecular Data
	subjective scoring	computationally demanding
Example	Dinosaur classification	Human migration using mtDNA

2.6 Case Studies: Fundamentals in Practice

The principles described above are illustrated by several landmark studies. The identification of Mitochondrial Eve relied upon maximum likelihood analysis of mtDNA sequences to establish a common maternal ancestor in Africa. The reconstruction of the three domains of life by Carl Woese used distance-based methods applied to ribosomal RNA. During the West African Ebola outbreak, Bayesian inference of viral genomes revealed human-to-human transmission patterns and cross-border spread. Most prominently, the COVID-19 pandemic placed phylogenetic analysis at the centre of global health response, with researchers sequencing SARS-CoV-2 genomes worldwide and constructing trees that traced the virus to a likely zoonotic origin in bats, tracked the emergence of variants including Alpha, Delta, and Omicron, and informed public health decision-making in near real time [7]. These examples collectively demonstrate that the choice of molecular data, the quality of alignment, and the rigour of tree inference are not merely academic concerns—they have direct consequences for scientific understanding and public health action.

Table 4. Examples of Phylogenetic Applications at the Fundamental Level

Case Study	Data Used	Method	Key Insight
Mitochondrial Eve	mtDNA	Maximum Likelihood	Common maternal ancestor in Africa
Y Chromosome Adam	Y chromosome	Parsimony + Bayesian	Common paternal ancestor in Africa
Three Domains of Life	rRNA	Distance + ML	Archaea distinct from Bacteria
Ebola Outbreak 2014	Viral genomes	Bayesian Inference	Human-to-human transmission patterns
SARS-CoV-2	Whole viral genomes	ML + Bayesian	Zoonotic origin; real-time variant tracking

3. Phylogenetic Algorithms

Phylogenetic tree inference is not merely a drawing exercise but a serious computational problem situated at the intersection of biology, mathematics, and computer science. Molecular sequences of DNA, RNA, and proteins contain within them the record of evolutionary relationships; however, extracting those relationships demands principled algorithms capable of navigating a combinatorially explosive space of possible topologies [8]. Three broad methodological families are recognized: distance-based methods, which compute pairwise distances between sequences; character-based methods, which examine individual nucleotide or amino acid positions; and statistical methods, comprising Maximum Likelihood and Bayesian inference, which model the probabilistic process of sequence evolution.

3.1 Distance-Based Methods

Distance-based methods were among the first computational approaches to phylogenetic inference. By mapping sequence similarity onto a distance measure representing genetic dissimilarity between any two taxa, they use clustering procedures to construct a tree from the resulting matrix [9]. Their computational simplicity and speed made them the dominant approach before high-performance computing became widely available. A recognized limitation, however, is that the reduction of sequence data to pairwise distances discards information on individual characters and may produce misleading trees when evolutionary rates differ substantially across lineages.

3.1.1 UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

UPGMA is a hierarchical clustering technique developed in the 1950s. It operates under the assumption of a strict molecular clock—that is, that all lineages evolve at the same rate—and produces ultrametric trees in which the distance from the root to each tip is identical. The algorithm identifies the two sequences with the smallest pairwise distance,

merges them into a cluster, recalculates distances between the new cluster and all remaining taxa by arithmetic averaging, and repeats until a single tree is produced. If clusters i and j are merged into cluster u, the distance to any other cluster k is calculated as $d(u,k) = (d(i,k) + d(j,k)) / 2$. UPGMA is extremely fast and produces trees that are valuable in dating studies where the molecular clock assumption is tenable; however, its performance deteriorates markedly when evolutionary rates vary across taxa, a situation that is common in practice.

3.1.2 Neighbor Joining (NJ)

Neighbor Joining, introduced by Saitou and Nei in 1987, remains one of the most widely used phylogenetic algorithms. Unlike UPGMA, it does not assume a molecular clock and therefore accommodates variable rates of evolution across lineages. Its principle is the minimization of total tree branch length: the algorithm identifies neighbor pairs whose joining minimizes a Q-matrix criterion, $Q(i,j) = (n-2)d(i,j) - \Sigma k\ d(i,k) - \Sigma k\ d(j,k)$, and iteratively joins them until the tree is complete. Branch lengths are estimated directly from the distance matrix at each step. Neighbor Joining is computationally efficient even for large datasets and commonly produces trees whose accuracy approaches that of Maximum Likelihood, particularly when the data are not highly heterogeneous. A recognized limitation is its sensitivity to noise in distance estimates, and bootstrap support values must be separately computed to assess confidence in the inferred clades.

3.2 Character-Based Methods

Character-based approaches retain the full information carried in each column of the multiple sequence alignment rather than reducing data to pairwise distances [10]. They are generally more accurate than distance methods but impose substantially greater computational demands.

3.2.1 Maximum Parsimony (MP)

Maximum Parsimony selects the tree that requires the fewest total evolutionary changes across all characters of the alignment—an application of Occam's razor to phylogenetics. The procedure involves generating candidate tree topologies, computing the minimum number of substitutions required across all sites for each topology using the Fitch algorithm, and selecting the topology with the lowest total change count. Parsimony is intuitive and requires no explicit evolutionary model; however, it is computationally infeasible for large datasets owing to the super-exponential growth of tree space, and it is subject to long-branch attraction—the spurious grouping of rapidly evolving lineages irrespective of their true relationships.

3.2.2 Maximum Likelihood (ML)

Maximum Likelihood, introduced by Felsenstein in 1981, is among the most statistically rigorous methods of phylogenetic inference. It selects the tree that maximizes the probability of observing the data under a specified substitution model. The likelihood function is $L(T, M \mid Data) = \Pi\_sites\ P(Data\_site \mid T, M)$, where T is the tree topology, M is the substitution model, and the product is taken across all aligned sites. A range of substitution models is available: the Jukes-Cantor (JC69) model assumes equal rates for all substitutions; the Kimura two-parameter (K2P) model distinguishes between transitions and transversions; and the General Time Reversible (GTR) model assigns unique rates to each type of substitution. Rate heterogeneity across sites is commonly modeled using a gamma distribution, which assigns different substitution rates to different site classes. ML is computationally demanding but provides statistically robust and biologically interpretable phylogenies, and it forms the basis for several widely used software implementations including RAxML, PhyML, and IQ-TREE.

3.2.3 Bayesian Inference (BI)

Bayesian inference applies Bayes' theorem to phylogenetics by combining the data likelihood with prior probabilities over tree topologies, branch lengths, and substitution model parameters to produce a posterior distribution of trees rather than a single optimal tree. Because the posterior distribution cannot be computed analytically for any dataset of practical size, Markov Chain Monte Carlo (MCMC) sampling is used to explore tree space, accepting or rejecting proposed trees according to their posterior probability. After a sufficient number of generations, the sampled trees converge on a representative sample of the posterior distribution. A consensus tree with branch support values in the form of posterior probabilities is then produced. Bayesian methods are particularly powerful for divergence time estimation when molecular clock models are combined with fossil or biogeographic calibration data, and they have become indispensable in viral epidemiology and biogeographic reconstruction.

3.3 Substitution Models and Mathematical Foundations

Substitution models describe the probabilistic process by which nucleotides or amino acids change over time and form the mathematical foundation of both likelihood and Bayesian phylogenetics. In the JC69 model, the simplest available, all base frequencies are assumed equal and all substitution rates identical; the probability that a nucleotide remains unchanged after time t is P_same(t) = ¼ + ¾ e^(−4αt), where α is the substitution rate. The K80 model distinguishes between transitions and transversions, a biologically motivated refinement since transitions occur more frequently in most organisms. The GTR model assigns an independent rate to each of the six possible types of nucleotide substitution and represents the most general reversible model in common use. Rate heterogeneity across sites, a pervasive feature of real sequence data, is routinely accounted for by a gamma distribution in which the shape parameter α governs the degree of rate variation: low values of α indicate substantial heterogeneity, while high values indicate uniformity. This correction is critical for accurate inference, especially in protein-coding genes where some sites are highly conserved while others are highly variable.

3.4 Algorithmic Search Strategies

Tree inference requires not only the evaluation of likelihood or parsimony scores but also an efficient strategy for navigating the astronomically large space of possible tree topologies. For datasets with fewer than twelve taxa, exhaustive evaluation of all possible trees guarantees an optimal solution, but this approach becomes computationally impractical as dataset size increases. Heuristic approaches are therefore employed in practice. Stepwise addition builds an initial tree by adding one taxon at a time; subsequent branch-swapping moves—including nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR)—explore the neighborhood of the current tree. Hill climbing iteratively improves tree scores but is prone to entrapment in local optima; simulated annealing addresses this limitation by permitting occasional acceptance of inferior solutions with a diminishing probability, allowing the search to escape local optima. Hybrid approaches combining these heuristics with metaheuristic strategies are increasingly employed in large-scale analyses.

3.5 Comparative Overview of Algorithms and Software

Table 5. Comparative Overview of Phylogenetic Algorithms

Method	Assumptions	Strengths	Weaknesses	Applications
UPGMA	Constant molecular clock	Very fast; ultrametric trees	Unrealistic equal-rate assumption	Symbiotic bacteria
Neighbor Joining	Variable rates allowed	Efficient; widely used	Sensitive to noise	COVID-19 early phylogenies
Max. Parsimony	Fewest changes	Simple; intuitive	Long-branch attraction; slow	Early human evolution
Max. Likelihood	Model-based probabilities	Highly accurate; flexible	Computationally expensive	Bird 10K genomes
Bayesian Inference	Priors + MCMC sampling	Captures uncertainty; divergence dating	Extremely resource intensive	SARS-CoV-2 epidemiology

Table 6. Software Implementations for Phylogenetic Analysis

Software	Algorithm	Strengths	Limitations	Best Use Case
MEGA	Distance, Parsimony, ML	User-friendly; educational	Limited scalability	Teaching; small datasets
PAUP*	Parsimony, ML	Historically important	Outdated for large data	Morphological + molecular
RAxML	ML	Speed; parallelization	Requires HPC	Genomic datasets (1000+ taxa)
PhyML	ML	Simple; moderate accuracy	Slower than RAxML	Medium datasets

Software	Algorithm	Strengths	Limitations	Best Use Case
IQ-TREE	ML	Auto model selection	Memory intensive	Medium–large datasets
MrBayes	Bayesian	Flexible; widely adopted	Long runtimes	Posterior probability inference
BEAST	Bayesian + clock	Divergence dating	Extremely expensive	Viral evolution; biogeography

4. Optimization Techniques in Phylogenetic Tree Construction

In phylogenetic inference, the central computational problem is one of optimization: searching an immense space of potential evolutionary trees, each representing a distinct hypothesis about the relationships among taxa, to identify those that best explain the observed sequence data. The space of possible tree topologies grows super-exponentially with dataset size, making exhaustive search impossible beyond small numbers of taxa. This challenge necessitates a range of heuristic and statistical optimization strategies, encompassing local tree-rearrangement heuristics, maximum likelihood parameter estimation, bootstrap resampling, Bayesian MCMC sampling, and metaheuristic approaches drawn from evolutionary computation and physics [11]. Optimization is also intimately connected with the quantification of uncertainty: bootstrap support values and Bayesian posterior probabilities provide the measures of confidence that allow phylogenetic conclusions to be stated with appropriate epistemic humility.

4.1 Heuristic Search Strategies

Because exhaustive evaluation of all possible tree topologies is computationally intractable for datasets of practical size, phylogenetic analyses rely on heuristic search strategies that systematically explore a subset of tree space. Most implementations begin with a starting tree constructed by stepwise taxon addition or a distance-based method, and then iteratively improve upon it through a series of rearrangement moves.

Nearest Neighbor Interchange (NNI) is the simplest move type, exchanging two subtrees that share an internal branch; it is computationally efficient but explores only a small neighborhood around the current tree. Subtree Pruning and Regrafting (SPR) removes a subtree from its current position and reattaches it elsewhere, exploring a broader region of tree space. Tree Bisection and Reconnection (TBR) bisects the tree into two components and re-joins them in all possible ways, providing the most thorough local search at the greatest computational cost [12]. Hill-climbing algorithms evaluate candidate trees at each step and accept only improvements; they are fast but prone to entrapment in local optima. Simulated annealing addresses this limitation by permitting occasional acceptance of inferior solutions, with the probability of such acceptance decreasing as the search progresses, thereby enabling the algorithm to escape local optima and converge toward the global optimum.

4.2 Likelihood Optimization

Maximum Likelihood inference requires optimization at two related levels: the search over discrete tree topologies and the continuous optimization of branch lengths and substitution model parameters. Given a fixed topology, branch lengths are optimized iteratively using methods such as Newton-Raphson or Expectation-Maximization, which adjust branch length values to maximize the likelihood function. Substitution model parameters—including transition-transversion ratios, base frequencies, and the gamma shape parameter controlling rate heterogeneity—are simultaneously optimized through repeated likelihood evaluations [13]. These computations are computationally demanding but yield statistically robust and biologically meaningful trees. Modern implementations such as RAxML and IQ-TREE achieve substantial speed improvements through parallelization and algorithmic innovations such as the ultrafast bootstrap approximation.

4.3 Bootstrapping and Resampling

Confidence assessment is an integral component of phylogenetic optimization. The bootstrap procedure, adapted from classical statistics, evaluates the stability of inferred clades by repeatedly resampling alignment columns with replacement, reconstructing a tree from each replicate, and computing the proportion of replicates in which each clade of the original tree is recovered as a bootstrap support value [14]. Values above 70–75% are conventionally interpreted as indicating moderate to strong support. Jackknife resampling, in which a fixed proportion of

sites is randomly removed without replacement, provides a complementary assessment of data robustness. Bayesian posterior probabilities offer a distinct measure of branch support, representing the proportion of trees in the posterior sample that contain a given clade, and are generally interpreted as providing a more direct probabilistic statement about the evidence in favor of that clade.

4.4 Bayesian Optimization and MCMC Sampling

Bayesian phylogenetics departs from the classical optimization paradigm by estimating a posterior distribution over trees rather than identifying a single optimal tree. The posterior distribution combines the data likelihood with prior probabilities over tree topologies, branch lengths, and substitution model parameters. Since direct computation of this posterior is analytically intractable, Markov Chain Monte Carlo (MCMC) methods are used to draw a representative sample of trees from the posterior by proposing random moves through tree space and accepting or rejecting them according to the Metropolis-Hastings criterion [15].

Convergence of the MCMC chain to the true posterior is assessed through diagnostics including the effective sample size and the agreement between multiple independent chains. After convergence, a consensus tree is produced with posterior probability support values on each branch. When combined with molecular clock models and calibration data from fossils or biogeography, Bayesian methods can simultaneously estimate phylogenetic relationships and divergence times, a capability that has proven especially valuable in viral epidemiology and historical biogeography.

4.5 Metaheuristic Approaches

For very large or highly complex datasets, conventional heuristic searches may prove insufficient to locate near-optimal trees within a reasonable time, and metaheuristic methods inspired by natural and physical processes offer powerful alternatives. Genetic algorithms evolve populations of trees through mutation, crossover, and selection, progressively favoring high-likelihood solutions while maintaining diversity to avoid premature convergence [16]. Particle Swarm Optimization models trees as particles that navigate the solution space guided by their own best-known position and the globally best-known position, providing an efficient collective search strategy. Simulated annealing, as noted above, allows temporary acceptance of sub-optimal solutions and is widely employed in conjunction with ML optimization in programs such as RAxML. Although these metaheuristics are computationally intensive, they have demonstrated the capacity to discover trees of superior quality in analyses involving hundreds or thousands of taxa where conventional hill-climbing would be insufficient.

4.6 Comparative Overview of Optimization Strategies

Table 7. Comparison of Optimization Strategies in Phylogenetic Inference

Method	Search Strategy	Parameter Focus	Output	Typical Applications
Heuristic (NNI, SPR, TBR)	Local tree rearrangements	Topology improvement	Single improved tree	Medium datasets; initial searches
Likelihood optimization	Iterative parameter refinement	Branch lengths; substitution rates	Optimized single tree	Genomic datasets
Bootstrapping	Data perturbation; resampling	Alignment columns	Branch support values	Assessing clade reliability
Bayesian MCMC	Posterior sampling	Topology; branch lengths; divergence times	Tree distribution with posteriors	Divergence dating; epidemiology
Metaheuristics (GA, PSO, SA)	Evolutionary or swarm-inspired	Combined topology and parameters	High-scoring tree(s)	Large-scale or complex datasets

5. Literature Review

Half a century of sustained research in evolutionary biology has placed phylogenetic tree construction at the centre of the discipline, providing a framework through which the relationships among organisms can be visualized and the processes underlying biological diversity can be examined. Although morphological characters dominated the early methods, the molecular revolution of the late

twentieth century redirected attention to DNA, RNA, and protein sequences as the primary source of phylogenetic information [17]. The subsequent proliferation of sequencing technologies and the continuing growth of genomics have radically expanded both the scope and the difficulty of phylogenetic inference. Optimization has accordingly become a defining element of modern phylogenetics, as researchers must evaluate enormous numbers of candidate trees and parameter combinations to identify the most probable evolutionary hypothesis.

Table 8 presents a structured and categorized summary of significant recent research works in phylogenetic tree construction, organized by methodological theme. The table covers distance and parsimony foundations, maximum likelihood optimization, Bayesian inference, scalability advances, machine learning and deep learning approaches, alignment methodology, and applied studies. It is intended to provide a consolidated reference map of the field for researchers approaching phylogenetic methods.

Table 8. Structured Literature Summary: Phylogenetic Tree Construction — Algorithms, Optimization, and Applications

A. Distance-Based and Parsimony Methods

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
Saitou & Nei [9]	1987	Neighbor Joining (NJ)	Minimizes total branch length without assuming molecular clock; fast and widely applicable	Sensitive to noise in distance estimates; no built-in support values	Protein and DNA sequence matrices
Felsenstein [8]	2004	Parsimony (Fitch/ Sankoff)	Comprehensive synthesis of parsimony, ML, and distance theory; foundational textbook for the field	Long-branch attraction under parsimony formally demonstrated	Theoretical and empirical phylogenetics
Zou et al. [17]	2024	NJ, MP, ML, Bayesian, Supermatrix in R	Systematic review and R-based implementation of all major methods; benchmarks performance on simulated and empirical gene family data	Does not address GPU-accelerated or cloud workflows	Gene families; simulated nucleotide alignments
Villalobos-Cid et al. [11]	2023	Evolutionary algorithm (parsimony-based)	Metaheuristic evolutionary algorithm for rooted phylogenetic network inference; outperforms greedy parsimony-based search on reticulate data	High computational cost relative to standard parsimony heuristics	Simulated reticulate evolution datasets

B. Maximum Likelihood Optimization

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
Felsenstein [3]	1981	Maximum Likelihood (original formulation)	Established statistical framework for phylogenetic inference; introduced pruning algorithm for computing tree likelihood efficiently	Computationally prohibitive for large datasets at the time	Small nucleotide alignments
Stamatakis [12]	2014	RAxML v8	Highly optimized ML inference with parallelization; became standard tool for large genomic datasets through the 2010s	Superseded by RAxML-NG for accuracy and model flexibility	Large empirical alignments; phylogenomics
Kozlov et al. [29]	2019	RAxML-NG	Rewrite of RAxML with improved heuristics, full GTR-derived model support, FreeRate heterogeneity, terrace detection	Slightly lower stability than IQ-TREE on some datasets	Large empirical and simulated alignments
Minh et al. [28]	2020	IQ-TREE 2	Partition models, concordance factors, mixture models, ultrafast bootstrap; comprehensive platform for genomic-era ML inference	Memory-intensive for very large datasets	Genomic datasets across all domains of life
Togkousidis et al. [14]	2025	RAxML-NG (early-stopping)	Early-stopping criterion halts optimization before over-fitting noisy likelihood surfaces; reduces runtime without accuracy cost	Benefit depends on dataset difficulty; less effective on easy datasets	Empirical genomic alignments

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
Haag et al. [33]	2022	Pythia + adaptive RAxML-NG	Gradient-boosted tree predicts dataset difficulty from alignment features before inference; adaptive resource allocation improves efficiency	Requires prior training on diverse dataset types for generalization	Multi-gene empirical alignments
De Maio et al. [34]	2023	MAPLE (pandemic-scale ML)	Parsimony-approximated likelihood for closely related genomes; infers SARS-CoV-2 phylogenies up to 1000x faster and with 100x less memory than RAxML/IQ-TREE	Optimized for epidemiological datasets; not general-purpose for divergent taxa	500,000+ SARS-CoV-2 genomes
Piñeiro et al. [24]	2024	VeryFastTree 4	Highly parallelized FastTree extension; enables ML inference on millions of sequences in hours rather than days	Lower accuracy than RAxML-NG or IQ-TREE on divergent datasets	Global SARS-CoV-2 surveillance datasets

C. Bayesian Inference and MCMC / Variational Methods

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
Ronquist & Huelsenbeck [21]	2003	MrBayes 3	First widely adopted MCMC-based Bayesian phylogenetics software; mixed models and posterior probability branch support	Long MCMC runtimes; convergence assessment challenging	Molecular systematics; mixed morphological-molecular data
Drummond & Rambaut [19]	2007	BEAST	Integrated Bayesian phylogenetics and molecular clock analysis; enables time-calibrated	Extremely computationally expensive; requires careful prior specification	Viral evolution; biogeography

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
			phylogenies and demographic inference		
Baridoo-Sottani et al. [20]	2024	BEAST2 (guidelines)	Systematic practical guidelines for BEAST-based inference; addresses prior choice, clock models, convergence diagnostics, and result interpretation	Primarily didactic; introduces no new algorithmic advances	Viral datasets; species diversification
Zhang & Matsen [22]	2024	Variational Bayesian (SBN)	Subsplit Bayesian networks provide expressive variational family; stochastic gradient ascent replaces MCMC; competitive accuracy at lower cost	Variational approximation may underestimate posterior variance in complex models	Benchmark biological phylogenetic datasets
Karcher et al. [32]	2024	Variational supertrees	Divide-and-conquer variational Bayesian supertrees; scalable uncertainty quantification for large taxon sets without full MCMC	Accuracy depends on quality of subtree decomposition strategy	Large-scale phylogenomics

D. Machine Learning and Deep Learning Applications

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
Azouri et al. [16]	2021	ML-guided heuristic search (RAxML)	Machine learning classifier trained on simulated data guides NNI/SPR moves; substantially accelerates convergence to optimal tree	Requires pre-training; performance depends on similarity between training and target datasets	Simulated and empirical alignments
Wang et al. [23]	2023	Deep autoregressive model	End-to-end neural network for phyloge	Evaluated on simulated data;	Simulated nucleotide alignments

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
			netic tree topology prediction from MSA; competitive accuracy at fraction of ML runtime	generalization to divergent empirical data not fully established	
Lambert et al. [31]	2023	Neural network (birth-death models)	Deep learning from phylogenies for diversification analysis; infers rates without explicit likelihood; generalizable to analytically intractable models	Accuracy sensitive to the realism of simulation-based training data	Bird and primate diversification phylogenies
Thompson et al. [30]	2024	CNN viral phylogeography	Convolutional neural networks match likelihood methods for migration rate and reproduction number estimation in viral phylogeography; orders of magnitude faster	Performance converges with likelihood methods only under well-specified models	Influenza and SARS-CoV-2 genomic surveillance

E. Sequence Alignment Methods and Alignment-Free Inference

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
Edgar [5]	2004	MUSCLE	High-accuracy progressive MSA; substantially outperforms CLUSTALW on benchmark alignments while maintaining comparable speed	Performance declines for highly divergent sequences	Protein and DNA sequence databases
Nute et al. [35]	2022	MAFFT + RAxML-NG / IQ-TREE comparative evaluation	Systematic evaluation of alignment tools and ML programs for viral epidemiology; MAFFT+RAxML-NG produces most accurate branch lengths for epidemiological distance calculations	Findings specific to closely related viral datasets; may not generalize to divergent taxa	HIV, HCV, Ebola simulated datasets
Dylus et al. [27]	2024	Read2Tree	Infers phylogenies directly from raw sequencing reads without explicit alignment; demonstrated accuracy	Less accurate than alignment-based ML on highly divergent sequences	Bacterial genomes; eukaryotic datasets

Reference	Year	Method / Tool	Key Contribution	Limitations Noted	Dataset / Domain
			for bacterial and eukaryotic genomes		

F. Applications — Epidemiology, Conservation, and Evolution

Reference	Year	Focus Area	Method Applied	Key Finding	Domain
Worobey et al. [7]	2022	SARS-CoV-2 origin	ML + molecular clock (BEAST)	Phylogenetic and geographic analysis established Huanan Seafood Market as early epicenter of COVID-19; zoonotic spillover confirmed	Pandemic genomic epidemiology
De Maio et al. [34]	2023	Pandemic surveillance	MAPLE (ML parsimony approximation)	Inferred phylogenies for 500,000 SARS-CoV-2 genomes; enabled real-time variant tracking at unprecedented scale	Genomic epidemiology; public health
Barido-Sottani et al. [20]	2024	Viral and diversification phylogenies	BEAST2 Bayesian inference	Demonstrated best practices for time-calibrate	Virology; macroevolution

Reference	Year	Focus Area	Method Applied	Key Finding	Domain
				d phylogenetic analysis; directly applicable to pandemic and macroevolutionary contexts	
Thompson et al. [30]	2024	Viral phylogeography	CNN deep learning	Deep learning provides accurate, rapid estimation of migration dynamics for rapidly evolving viruses; suitable for real-time outbreak response	Influenza; SARS-CoV-2 surveillance
Lambert et al. [31]	2023	Diversification rate inference	DL from phylogenies (BD models)	DL enables diversification inference under models previously inaccessible due to likelihood intractability; broadens macroevolutionary analysis	Avian and primate evolutionary biology

5.1 Classical Foundations: Distance-Based and Parsimony Methods

The systematic application of computational methods to phylogenetic inference began in earnest during the latter half of the twentieth century, when the accumulation of protein and, subsequently, nucleotide sequence data created the need for algorithms capable of processing large volumes of molecular information. Distance-based approaches were among the first to be formalized. The UPGMA algorithm, rooted in hierarchical agglomerative clustering, was among the earliest methods applied to molecular data [9]. Although conceptually straightforward, UPGMA operates under the assumption of a strict molecular clock—that all lineages evolve at a constant rate—a condition that is rarely met in practice and that compromises the method's reliability when rate heterogeneity exists across branches.

The introduction of Neighbor Joining by Saitou and Nei in 1987 represented a significant practical improvement, producing trees that minimize total branch length without assuming rate constancy [9]. NJ proved fast enough to analyze the datasets available at the time, and its accuracy proved acceptable for a wide range of applications. The method continues to be employed today for rapid preliminary analyses and for datasets where computational efficiency is paramount. Subsequent evaluations have confirmed that NJ trees, when distances are corrected for multiple substitutions using models such as Jukes-Cantor or Kimura two-parameter, perform comparably to more elaborate methods for closely related taxa, although performance deteriorates in datasets characterized by long branches or substantial rate variation [17].

Parsimony-based methods, which seek the tree requiring the fewest evolutionary changes, were developed and championed primarily in the context of morphological systematics before being adapted to molecular data. The application of parsimony to nucleotide sequences is grounded in algorithms such as the Fitch and Sankoff methods for computing minimum substitution counts at each site. Parsimony has the conceptual appeal of Occam's razor and requires no explicit model of sequence evolution; however, Felsenstein's landmark 1978 demonstration of long-branch attraction—the spurious clustering of rapidly evolving lineages irrespective of their true relationships—identified a systematic bias that limits the method's reliability in the presence of rate heterogeneity [8]. Despite this limitation, parsimony-based heuristics including NNI, SPR, and TBR continue to provide the core search logic for many likelihood-based programs, reflecting the continued practical relevance of the underlying algorithmic approaches.

5.2 The Maximum Likelihood Framework and Its Optimization

The introduction of Maximum Likelihood as a principled framework for phylogenetic inference by Felsenstein in 1981 transformed the field by providing a coherent statistical basis for comparing competing evolutionary hypotheses [8]. The ML approach evaluates candidate trees according to the probability of the observed sequence data given a specified substitution model and tree, selecting the tree that maximizes this probability. This criterion is statistically well-founded and, in contrast to parsimony, explicitly accounts for the stochastic nature of sequence evolution. The computational challenge is formidable: for any dataset of practical size, the exact optimization of the likelihood surface over all possible tree topologies, branch lengths, and model parameters is computationally intractable, necessitating heuristic search strategies.

A critical component of the ML framework is the selection of an appropriate substitution model. Kalyaanamoorthy et al. developed ModelFinder, a fast model-selection method implemented in IQ-TREE that evaluates a large space of substitution models using information criteria such as BIC to identify the best-fitting model for a given alignment [28]. ModelFinder is reported to be ten to one hundred times faster than earlier programs such as jModelTest and ProtTest, while achieving superior or comparable accuracy. The importance of model selection has been further underscored by studies demonstrating that model misspecification—the use of an overly simplified model—leads to consistently biased parameter estimates and incorrect tree topologies, particularly for datasets characterized by substantial among-site rate heterogeneity [13].

The development of computationally efficient ML software has been a sustained focus of the field. RAxML, originally developed by Stamatakis, established a benchmark for speed in large-scale ML analysis and has undergone successive improvements [12]. Kozlov et al. presented RAxML-NG, a re-implementation that incorporates improved heuristic search algorithms, support for all GTR-derived substitution models, the FreeRate model of rate heterogeneity, and scalable

parallelization [29]. Benchmark comparisons demonstrated that RAxML-NG consistently recovers higher-scoring trees than its predecessor on taxon-rich datasets. IQ-TREE 2, developed by Minh et al., introduced additional innovations including partition models that assign distinct substitution models to different genes or genomic regions, concordance factors that quantify the proportion of loci supporting each branch, and mixture models for capturing heterogeneous evolutionary processes [28]. The ultrafast bootstrap approximation implemented in IQ-TREE provides branch support values at speeds approximately one hundred times faster than standard bootstrap resampling, making rigorous support assessment feasible for large genomic datasets.

A notable recent contribution to scalable ML inference is MAPLE, presented by De Maio et al. (2023), which reworks Felsenstein's pruning algorithm specifically for epidemiological genomic datasets [34]. MAPLE was demonstrated to infer SARS-CoV-2 phylogenies more accurately than existing ML approaches while operating up to thousands of times faster and requiring at least one hundred times less memory on large datasets comprising millions of genomes. This development addressed a critical bottleneck in pandemic surveillance, where the use of conventional tools such as RAxML or IQ-TREE on global SARS-CoV-2 datasets would require years of computation per tree update.

5.3 Bayesian Phylogenetics: From MCMC to Variational Inference

Bayesian inference entered phylogenetics in the late 1990s and rapidly gained widespread adoption owing to its capacity to quantify uncertainty across the full posterior distribution of trees rather than identifying a single optimal topology. The implementation of MCMC for Bayesian phylogenetics, first realized in programs such as MrBayes, allowed researchers to sample the posterior distribution over tree topologies, branch lengths, and substitution model parameters simultaneously, with posterior probability values on branches providing a direct probabilistic measure of clade support [21]. BEAST, developed by Drummond and Rambaut, extended the Bayesian framework to include molecular clock models and fossil or biogeographic calibration priors, enabling the estimation of divergence times alongside phylogenetic relationships and making time-calibrated phylogenies a standard tool in evolutionary biology [19].

Barido-Sottani et al. (2024) provided a systematic guide to Bayesian phylogenetic inference using BEAST, addressing the selection of priors, the specification of clock models, the assessment of MCMC convergence through diagnostics including the effective sample size and Gelman-Rubin statistic, and the interpretation of posterior distributions [20]. This work emphasized that the reliability of Bayesian phylogenetic conclusions depends critically on the adequacy of MCMC convergence, a point that is frequently underappreciated in applied studies. Bouckaert et al. documented the advances incorporated in BEAST 2, which introduced a modular package architecture enabling users to combine diverse evolutionary models and facilitating the rapid development and testing of new methodological innovations [19].

A fundamental limitation of MCMC-based Bayesian inference is its computational cost: analyses of large genomic datasets may require weeks or months on high-performance computing clusters, and the sequential nature of the MCMC chain places practical limits on the datasets that can be analyzed. Zhang and Matsen IV addressed this limitation by developing a variational approach to Bayesian phylogenetic inference, in which subsplit Bayesian networks serve as an expressive family of variational distributions over tree topologies, and structured amortization is applied to branch lengths [22]. The variational approximation is trained by stochastic gradient ascent, and the resulting approach was demonstrated to provide competitive performance relative to MCMC while requiring substantially fewer iterations. Karcher et al. extended this framework to the supertree setting, developing variational supertrees that enable scalable Bayesian phylogenetic inference for large taxon sets by combining a divide-and-conquer decomposition with variational optimization [32]. These developments represent a significant step toward making Bayesian uncertainty quantification computationally accessible for phylogenomic datasets.

5.4 Sequence Alignment as a Prerequisite and Its Effect on Tree Accuracy

Sequence alignment is a prerequisite for the vast majority of phylogenetic analyses, and it is well established that alignment error propagates into tree inference, potentially biasing both topology and

branch length estimates. The performance of multiple sequence alignment programs has been extensively benchmarked, with MAFFT and SATé consistently achieving the lowest alignment error rates across diverse test sets, while CLUSTALW and related programs have been shown to produce substantially higher error rates, particularly for divergent sequences [17]. Nute et al. demonstrated in an evaluation of phylogenetic workflows for viral molecular epidemiology that MAFFT consistently outperformed MUSCLE and Clustal Omega in both accuracy and runtime across simulations of HIV, HCV, and Ebola datasets [35]. The study further established that while FastTree 2, IQ-TREE, RAxML-NG, and PhyML achieve similar topological accuracy given a correct alignment, RAxML-NG produces the most accurate branch lengths and pairwise distances—a finding with direct relevance to applications where branch lengths carry biological meaning, such as divergence time estimation and epidemiological distance calculations.

The recognition that alignment uncertainty itself constitutes a source of phylogenetic error has motivated the development of methods that co-estimate alignment and tree simultaneously, most notably the SATé algorithm, which iterates between alignment and tree refinement. More recent work has taken an alignment-free approach: Dylus et al. (2024) introduced Read2Tree, a method that infers phylogenies directly from raw sequencing reads without performing explicit multiple sequence alignment [27]. Read2Tree was shown to produce accurate phylogenies for both bacterial and eukaryotic genomes, providing a practical pathway for rapid phylogenetic inference when alignment is computationally prohibitive or technically challenging.

5.5 Coalescent Models and Gene Tree-Species Tree Reconciliation

The recognition that individual gene trees may differ from the species tree owing to biological processes including incomplete lineage sorting, hybridization, and horizontal gene transfer has motivated the development of phylogenetic methods that explicitly model this discordance. ASTRAL, introduced by Mirarab et al. and subsequently improved in a series of versions, estimates species trees by maximizing the number of shared quartet topologies across a set of estimated gene trees, a criterion that is statistically consistent under the multispecies coalescent model [34]. ASTRAL has been shown to outperform concatenation-based ML methods in datasets characterized by high levels of incomplete lineage sorting, while remaining efficient enough to handle datasets with thousands of genes and hundreds of taxa.

The challenge of gene tree-species tree discordance is particularly acute in phylogenomics, where the availability of genome-scale data provides both the opportunity and the obligation to account for variation in evolutionary histories across loci. Zou et al. (2024) described and benchmarked supermatrix and supertree approaches within the R computational environment, providing practical guidance on the choice between concatenation and coalescent-based methods and illustrating the circumstances under which each is appropriate [17]. Their analysis confirmed that concatenation frequently outperforms summary coalescent methods when ILS levels are low or gene trees are estimated from short alignments with high uncertainty, while coalescent methods provide superior accuracy when discordance is extensive. This interaction between data properties and methodological performance underscores the importance of informed method selection in phylogenomic studies.

5.6 Scalability and Adaptive Optimization Strategies

As datasets in molecular systematics, epidemiology, and phylogenomics have grown to encompass thousands of taxa and millions of aligned sites, the computational demands of phylogenetic inference have become a central concern. Piñeiro et al. (2024) developed VeryFastTree 4, a highly parallelized implementation of the approximate ML algorithm that extends the original FastTree 2 to handle datasets comprising millions of sequences [24]. VeryFastTree 4 was applied to global SARS-CoV-2 datasets and demonstrated the capacity to construct phylogenies in hours that would require days or weeks with conventional tools, directly enabling real-time genomic surveillance of the evolving pandemic.

Haag et al. (2022) introduced Pythia, a gradient-boosted tree classifier trained to predict the difficulty of a phylogenetic analysis from properties of the multiple sequence alignment before inference begins [33]. Pythia's difficulty predictions were used to design an adaptive RAxML-NG strategy that allocates computational resources in proportion to the complexity of the search landscape, reducing

total runtime without compromising accuracy on easy datasets while investing additional effort in difficult ones. This work represented a conceptual advance in that it treats the allocation of computational resources as an optimization problem in its own right, one that can be addressed using machine learning applied to dataset characteristics. Togkousidis et al. (2025) proposed an early-stopping criterion for ML tree inference that detects when further optimization is unlikely to improve the tree score and terminates the search, avoiding the over-optimization of noisy likelihood surfaces that can actually reduce accuracy on empirical datasets [14].

5.7 Machine Learning and Deep Learning for Phylogenetic Inference

The application of machine learning to phylogenetics has progressed rapidly from proof-of-concept studies to practical tools with demonstrated advantages over classical methods in specific contexts. Azouri et al. (2021) showed that guidance from a machine learning classifier trained on simulated data could substantially accelerate the heuristic tree search in RAxML, reducing the number of NNI and SPR moves required to reach a tree of given quality [16]. This approach exploited the observation that not all topological moves are equally likely to improve the tree score, and that this likelihood can be estimated from local features of the tree using a machine learning model.

Wang et al. (2023) developed a deep learning framework for phylogenetic tree inference that processes multiple sequence alignments through a neural network architecture to directly predict tree topologies, demonstrating competitive accuracy with ML methods on simulated datasets while operating at substantially reduced computational cost [23]. Lambert et al. (2023) extended deep learning approaches to the problem of diversification analysis, training neural networks on phylogenetic trees generated under birth-death models and using them to infer diversification rates without the need for an explicit likelihood formula [31]. This approach is particularly valuable for diversification models whose likelihood is analytically intractable or computationally prohibitive to evaluate. Thompson et al. (2024) applied convolutional neural networks to the problem of viral phylogeography, demonstrating that deep learning estimates of migration rates and reproduction numbers converge on the same values as likelihood-based methods in well-specified models while being orders of magnitude faster, with direct implications for real-time outbreak response [30].

5.8 Applications in Pandemic Surveillance and Viral Evolution

The COVID-19 pandemic provided an unprecedented test of phylogenetic methods at scale, with millions of SARS-CoV-2 genomes sequenced within two years of the virus's emergence. The rapid accumulation of genomic data created both opportunities and computational challenges: while the data provided an extraordinarily detailed view of viral evolution, the volume exceeded the capacity of conventional ML and Bayesian tools to analyze without subsampling, which reduced resolution and statistical power [34]. De Maio et al. addressed this through MAPLE, while tools such as Nextstrain, which combines ML phylogenetics with geospatial visualization, were adapted for continuous real-time analysis, tracking the emergence and spread of variants of concern including Alpha, Delta, and Omicron as they arose [34].

Phylogenetic analyses played a central role in establishing the zoonotic origin of SARS-CoV-2, identifying bat coronaviruses as the closest known relatives of the pandemic virus and evaluating the roles of various intermediate hosts. Time-calibrated Bayesian analyses using BEAST estimated the date of the most recent common ancestor of sampled SARS-CoV-2 sequences, providing evidence consistent with emergence in late 2019 and informing public health policy regarding the timing and extent of early transmission. The practical importance of these analyses established phylogenetics as an essential component of pandemic preparedness infrastructure, with investment in scalable and real-time phylogenetic methods now recognized as a public health priority.

5.9 Applications in Conservation Biology, Agriculture, and Human Diversity

Optimized phylogenetic trees provide the foundation for phylogenetic diversity metrics that are increasingly employed by conservation planners to guide the prioritization of species for protection. By estimating the evolutionary history represented by different assemblages of species, these metrics quantify the evolutionary heritage that would be lost through extinction and allow conservation priorities to be set in a manner that maximizes the preservation of unique evolutionary lineages. Studies of amphibians, mammals, and plants have demonstrated that phylogenetic diversity-based

priorities frequently diverge substantially from those based on species richness alone, making the accuracy of the underlying phylogenetic optimization directly relevant to conservation decision-making [26].

In agriculture, phylogenetic optimization has been applied to resolve the domestication histories of major crops including maize, rice, wheat, and soybean. These analyses have documented complex domestication events involving multiple independent origins, hybridization, and introgression from wild relatives, providing insights that guide plant breeding programs and the identification of wild germplasm as a source of agriculturally valuable traits. In human evolutionary biology, phylogenetic analyses of ancient DNA—including the landmark sequencing and analysis of Neanderthal and Denisovan genomes—have revealed the timing and extent of archaic admixture into modern human populations, establishing that a small but detectable proportion of the genomes of non-African individuals today traces to archaic human lineages [25]. These analyses required the development of specialized substitution models that account for the DNA damage patterns characteristic of ancient specimens, illustrating the continued need for methodological innovation in response to new data types.

5.10 Cancer Phylogenetics and Emerging Applications

An active and growing application of phylogenetic inference is the reconstruction of clonal evolutionary histories within tumors. Sequencing of multiple spatially or temporally separated samples from a single tumor allows the construction of a phylogenetic tree representing the clonal diversification of the cancer cell population over the course of disease progression. These analyses have revealed the heterogeneity of tumor evolution, the timing and order of driver mutation acquisition, and the mechanisms by which resistant subclones emerge and expand under the selective pressure of treatment [26]. Likelihood and Bayesian methods adapted from classical phylogenetics have been successfully applied to tumor phylogeny inference, with modifications to accommodate the high mutation rates, copy number variation, and mixed-ancestry sequencing reads characteristic of cancer genomic data.

Beyond cancer, phylogenetic methods are finding application in an expanding range of fields. In microbiology, phylogenomic analyses of thousands of bacterial and archaeal genomes have resolved the deep relationships among prokaryotic lineages and placed horizontal gene transfer within a quantitative evolutionary framework. In linguistics and cultural evolution, phylogenetic approaches have been applied to reconstruct the history of language families, the diffusion of technologies, and the transmission of cultural practices across human populations. These diverse applications share the fundamental computational and statistical challenges of phylogenetic inference, and methodological advances in one domain frequently prove transferable to others.

6. Emerging Themes and Future Directions

6.1 Reticulate Evolution and Network-Based Optimization

Conventional phylogenetic inference assumes strictly bifurcating trees, but real evolutionary histories are frequently more network-like due to horizontal gene transfer, hybridization, introgression, and recombination. Horizontal gene transfer, common in bacteria and archaea, breaks vertical inheritance patterns and can mislead tree-based optimization, necessitating phylogenetic network models that incorporate reticulation nodes, albeit at substantially increased computational cost. In plants—where polyploid lineages arising from hybridization are common—optimization must account for multiple ancestral contributions and estimate the timing of hybridization events. Viral and bacterial recombination generates genomic regions with distinct evolutionary histories; recombination-aware optimization methods, which identify mosaic regions and apply separate models to each, are increasingly employed to avoid artifact-ridden inferences.

6.2 The Multi-Species Coalescent

A major theoretical advance in phylogenomics has been the incorporation of the multi-species coalescent (MSC) model into optimization frameworks. The MSC accounts for the expectation that individual gene trees will differ from the species tree owing to incomplete lineage sorting, which is particularly pronounced during rapid diversification events. Under this model, phylogenetic inference simultaneously estimates a species-level tree and a set of embedded gene trees, a hierarchical optimization problem of considerable complexity. Tools such as ASTRAL summarize concordance

among gene trees to produce a species tree estimate, while BEAST jointly estimates gene trees, species tree topology, and divergence times under a fully Bayesian framework. These approaches have proven essential for resolving shallow radiations in plants and vertebrates where gene tree discordance is pervasive.

6.3 Artificial Intelligence and Deep Learning

The application of deep learning to phylogenetic optimization represents one of the most actively developing frontiers in the field. Neural networks trained on simulated evolutionary data have demonstrated the capacity to rapidly approximate near-optimal tree topologies, providing starting points that substantially reduce the computational burden of subsequent likelihood refinement. Reinforcement learning agents that learn effective strategies for navigating tree space combine the adaptability of machine learning with the statistical power of stochastic optimization. Variational inference approaches—including variational Bayesian phylogenetics and differentiable tree inference methods—offer the prospect of performing Bayesian phylogenetic analysis at scales that are inaccessible to MCMC-based methods, opening new possibilities for phylogenomic analyses of very large datasets.

6.4 Structural and Functional Constraints in Optimization

Classical phylogenetic optimization relies exclusively on nucleotide or amino acid sequences, but structural and functional constraints on molecular evolution carry phylogenetically relevant information that can improve inference. For ribosomal RNA, conserved secondary structures impose correlated substitution patterns; structure-aware likelihood models that preserve base-pairing constraints have been shown to reduce both alignment error and the incidence of incorrect tree topologies. In protein phylogenetics, substitution models that incorporate structural and functional constraints assign biologically motivated weights to amino acid changes, improving accuracy particularly for enzymes and immune proteins that are under strong functional selection.

6.5 Future Research Directions

Several directions appear particularly promising for the advancement of phylogenetic optimization. First, the development of scalable Bayesian methods—through variational inference and approximate computation—will be essential to extend the statistical rigour of Bayesian inference to phylogenomic datasets comprising thousands of taxa. Second, the integration of machine learning with conventional optimization offers substantial potential for improving both speed and accuracy, particularly in real-time surveillance applications where rapid inference is critical. Third, the explicit modeling of reticulate evolution in optimization frameworks will be necessary to accurately reconstruct the evolutionary histories of groups—including bacteria, archaea, and many plant lineages—for which tree-like inheritance is an inadequate approximation. Fourth, improved tools for the visualization and interpretation of optimized phylogenies, including interactive platforms that communicate uncertainty in statistically appropriate terms, will be required to make the outputs of phylogenetic analysis accessible to the broad scientific and medical communities that increasingly depend upon them.

7. Conclusion

The construction of phylogenetic trees has been transformed from a descriptive diagrammatic exercise into a sophisticated computational discipline at the intersection of molecular biology, statistics, and computer science. Contemporary methods—spanning distance-based clustering, maximum parsimony, maximum likelihood, Bayesian inference, coalescent modeling, and emerging metaheuristic and AI-guided approaches—extract evolutionary information from complex molecular data while simultaneously estimating branch lengths, substitution parameters, and their associated uncertainties. The applications of optimized phylogenetic inference are far-reaching, encompassing pandemic surveillance, conservation genomics, cancer evolution, paleogenomics, agriculture, and cultural evolution, and continue to expand as new biological systems and questions are addressed through the lens of evolutionary history.

Significant challenges remain. Model misspecification, alignment error, reticulate evolution, and computational scalability continue to impose limits on the accuracy and accessibility of phylogenetic inference, particularly as dataset sizes grow. The integration of artificial intelligence and machine learning with conventional statistical optimization offers a promising route toward analyses that are simultaneously faster, more

accurate, and better calibrated. As sequencing technology continues to reduce costs and high-performance computing becomes more widely accessible, the capacity to apply phylogenetic optimization to the full breadth of biological diversity appears within reach, promising continued fundamental contributions to evolutionary biology and its applications.

Declarations

Funding: The authors declare that they have not received any external funding for this study.

Conflict of Interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of Data and Material: Data sharing is not applicable to this article.

Code Availability: Code sharing is not applicable to this article.

Authors' Contributions: All authors read and approved the final manuscript.

References

1. Darwin, C.: On the Origin of Species by Means of Natural Selection. John Murray, London (1859)

2. Watson, J.D., Crick, F.H.C.: Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171(4356), 737–738 (1953). https://doi.org/10.1038/171737a0

3. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17(6), 368–376 (1981). https://doi.org/10.1007/BF01734359

4. Woese, C.R., Fox, G.E.: Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA 74(11), 5088–5090 (1977)

5. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5), 1792–1797 (2004). https://doi.org/10.1093/nar/gkh340

6. Felsenstein, J.: Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39(4), 783–791 (1985)

7. Worobey, M., et al.: The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. Science 377(6609), 951–959 (2022). https://doi.org/10.1126/science.abp8715

8. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Sunderland, MA (2004)

9. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4), 406–425 (1987)

10. Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M.: Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (eds.) Molecular Systematics, 2nd edn, pp. 407–514. Sinauer Associates, Sunderland (1996)

11. Villalobos-Cid, M., et al.: An evolutionary algorithm based on parsimony for the inference of rooted phylogenetic networks. Swarm Evol. Comput. 81, 101349 (2023). https://doi.org/10.1016/j.swevo.2023.101349

12. Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9), 1312–1313 (2014). https://doi.org/10.1093/bioinformatics/btu033

13. Yang, Z.: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39(3), 306–314 (1994)

14. Togkousidis, A., et al.: Accelerating maximum likelihood phylogenetic inference via early-stopping to evade over-optimization. Syst. Biol. (2025). https://doi.org/10.1093/sysbio/syaf043

15. Zhang, R., et al.: Fast Bayesian inference of phylogenies from multiple continuous characters. Syst. Biol. 73(1), 102–120 (2024). https://doi.org/10.1093/sysbio/syad063

16. Azouri, D., Abadi, S., Mansour, Y., Mayrose, I., Pupko, T.: Harnessing machine learning to guide phylogenetic-tree search algorithms. Nat. Commun. 12, 1983 (2021). https://doi.org/10.1038/s41467-021-22073-8

17. Zou, Y., Zhang, Z., Zeng, Y., Hu, H., Hao, Y., Huang, S., Li, B.: Common methods for phylogenetic tree construction and their implementation in R. Bioengineering 11(5), 480 (2024). https://doi.org/10.3390/bioengineering11050480

18. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q.: IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32(1), 268–274 (2015). https://doi.org/10.1093/molbev/msu300

19. Drummond, A.J., Rambaut, A.: BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214 (2007). https://doi.org/10.1186/1471-2148-7-214

20. Barido-Sottani, J., et al.: Practical guidelines for Bayesian phylogenetic inference using BEAST. Open Res. Eur. 3, 204 (2024). https://doi.org/10.12688/openreseurope.16679.2

21. Ronquist, F., Huelsenbeck, J.P.: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12), 1572–1574 (2003). https://doi.org/10.1093/bioinformatics/btg180

22. Zhang, C., Matsen IV, F.A.: A variational approach to Bayesian phylogenetic inference. J. Mach. Learn. Res. 25, 1–56 (2024)

23. Wang, Z., et al.: A framework for phylogenetic tree inference via deep learning. Nucleic Acids Res. 51(22), e113 (2023). https://doi.org/10.1093/nar/gkad807

24. Piñeiro, C., et al.: Efficient phylogenetic tree inference for massive taxonomic datasets: VeryFastTree 4. Algorithms Mol. Biol. 19, 14 (2024). https://doi.org/10.1186/s13015-024-00264-4

25. Penn, M.J., et al.: Leaping through tree space: continuous phylogenetic inference for rooted and unrooted trees. arXiv preprint arXiv:2306.05739 (2023)

26. Mimori, T., Hamada, M.: GeoPhy: differentiable phylogenetic inference via geometric gradients of tree topologies. arXiv preprint arXiv:2307.03675 (2023)

27. Dylus, D., et al.: Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. Nat.

Biotechnol. 42, 139–147 (2024). https://doi.org/10.1038/s41587-023-01753-4

28. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R.: IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37(5), 1530–1534 (2020). https://doi.org/10.1093/molbev/msaa015

29. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A.: RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35(21), 4453–4455 (2019). https://doi.org/10.1093/bioinformatics/btz305

30. Thompson, A., Liebeskind, B.J., Scully, E.J., Landis, M.J.: Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong. Syst. Biol. 73(1), 183–206 (2024). https://doi.org/10.1093/sysbio/syad074

31. Lambert, S., Voznica, J., Morlon, H.: Deep learning from phylogenies for diversification analyses. Syst. Biol. 72(6), 1262–1279 (2023). https://doi.org/10.1093/sysbio/syad044

32. Karcher, M.D., Zhang, C., Matsen IV, F.A.: Variational supertrees for Bayesian phylogenetics. Bull. Math. Biol. 86, 109 (2024). https://doi.org/10.1007/s11538-024-01338-5

33. Haag, J., Höhler, D., Bettisworth, B., Stamatakis, A.: From easy to hopeless—predicting the difficulty of phylogenetic analyses. Mol. Biol. Evol. 39(12), msac254 (2022). https://doi.org/10.1093/molbev/msac254

34. De Maio, N., et al.: Maximum likelihood pandemic-scale phylogenetics. Nat. Genet. 55, 746–752 (2023). https://doi.org/10.1038/s41588-023-01368-0

35. Nute, M., et al.: An evaluation of phylogenetic workflows in viral molecular epidemiology. Virus Evol. 8(1), veac082 (2022). https://doi.org/10.1093/ve/veac082