Phylogenetic Analysis: Tree Reconstruction and Evolutionary Inference

Distance Methods, Maximum Likelihood, and Bootstrap Support

Computational Phylogenomics Division
Computational Science Templates

November 24, 2025

Abstract

This comprehensive analysis presents methods for reconstructing phylogenetic trees from molecular sequence data. We cover distance-based methods (UPGMA, Neighbor-Joining), character-based approaches, and statistical support via bootstrapping. The analysis includes distance corrections for multiple substitutions (Jukes-Cantor, Kimura), tree search algorithms, and visualization of evolutionary relationships. We demonstrate phylogenetic inference using primate mitochondrial sequences and evaluate tree topology confidence through bootstrap resampling.

1 Introduction

Phylogenetics reconstructs the evolutionary history of species or sequences from observed molecular or morphological data. The goal is to infer the tree topology (branching pattern) and branch lengths (evolutionary distances) that best explain the data.

Definition 1 (Phylogenetic Tree) *A phylogenetic tree is a branching diagram representing evolutionary relationships. Tips (leaves) represent extant taxa, internal nodes represent ancestral taxa, and branch lengths represent evolutionary change.*

2 Theoretical Framework

2.1 Distance Corrections

Observed sequence differences underestimate true evolutionary distance due to multiple substitutions:

Theorem 1 (Jukes-Cantor Correction) *For equal substitution rates among nucleotides:*

$$d = -\frac{3}{4} \ln \left(1 - \frac{4p}{3}\right) \tag{1}$$

where p is the observed proportion of differences.

Theorem 2 (Kimura Two-Parameter Model) *Allowing different transition (s) and transversion (v) rates:*

$$d = -\frac{1}{2} \ln \left[(1 - 2s - v)\sqrt{1 - 2v}\right] \tag{2}$$

2.2 UPGMA Method

Definition 2 (UPGMA) *Unweighted Pair Group Method with Arithmetic Mean assumes a molecular clock (constant rate). It builds trees by iteratively clustering the closest taxa:*

$$d_{(ij)k} = \frac{n_i \cdot d_{ik} + n_j \cdot d_{jk}}{n_i + n_j} \tag{3}$$

where clusters i and j are merged and distances to taxon k are updated.

2.3 Neighbor-Joining

Theorem 3 (Neighbor-Joining Criterion) *NJ selects pairs that minimize total tree length using transformed distances:*

$$Q_{ij} = (n-2)d_{ij} - \sum_k d_{ik} - \sum_k d_{jk} \tag{4}$$

The pair with minimum Q_{ij} is joined.

Remark 1 (UPGMA vs. NJ)

- *UPGMA produces ultrametric trees (all tips equidistant from root)*
- *NJ allows variable evolutionary rates*
- *NJ is generally more accurate for real data*

3 Computational Analysis

4 Results and Analysis

4.1 Distance Matrix

4.2 Tree Statistics

Example 1 (Primate Phylogeny) *The reconstructed primate tree shows:*

- *Total tree length: ??*
- *Cophenetic correlation: ??*
- *Human-Chimp clade is most strongly supported*
- *African great apes form a monophyletic group*
- *Gibbon is sister to great apes*

5 Bootstrap Analysis

Definition 3 (Phylogenetic Bootstrap) *Bootstrap support measures clade reproducibility:*

1. *Resample characters (or distances) with replacement*
2. *Reconstruct tree from resampled data*
3. *Count frequency each clade appears*
4. *Values >70% considered strong support*

Remark 2 (Interpreting Bootstrap Values)

- *>95%: Very strong support*
- *70-95%: Moderate support*
- *<70%: Weak support*
- *Bootstrap is conservative (underestimates true support)*

6 Model Selection

6.1 Substitution Models

Choice of distance correction affects tree topology:

- **Jukes-Cantor**: Equal base frequencies, equal rates
- **Kimura 2P**: Different transition/transversion rates
- **HKY85**: Unequal base frequencies + Ti/Tv ratio
- **GTR**: General time-reversible (most parameters)

6.2 Model Testing

Likelihood ratio tests or information criteria (AIC, BIC) select best-fit model.

7 Limitations and Extensions

7.1 Model Limitations

1. **Distance loss**: Character information discarded
2. **Molecular clock**: UPGMA assumes equal rates
3. **Star tree**: Poor resolution for rapid radiations
4. **Long branch attraction**: Fast-evolving taxa cluster

7.2 Possible Extensions

- Maximum likelihood phylogenetics
- Bayesian inference with posterior probabilities
- Coalescent methods for population-level data
- Phylogenomics with whole-genome data

8 Conclusion

This analysis demonstrates phylogenetic reconstruction:

- Distance methods provide fast tree inference
- Neighbor-Joining handles rate variation better than UPGMA
- Bootstrap support quantifies clade confidence
- Primate phylogeny matches expected relationships
- Distance correction is essential for divergent sequences

Further Reading

- Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.
- Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press.
- Saitou, N. & Nei, M. (1987). The Neighbor-Joining method. *MBE*, 4, 406-425.