Chapter 6-Evolution and Phylogenetics - Unraveling the Tree of Life

Reza Rezazadegan
Shiraz University
www.dreamintelligent.com

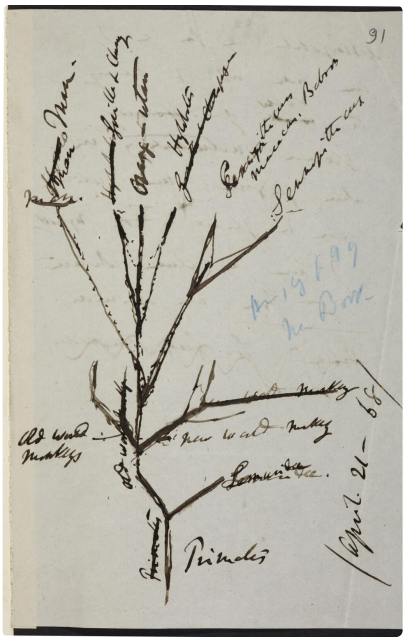
Learning Outcomes: Upon completing this chapter, students should be able to:

- **Define** molecular phylogenetics and explain its significance in reconstructing evolutionary history.
- **Explain** the fundamental principles of molecular evolution, including mutation, natural selection, and their impact on sequence divergence.
- **Explain** the differences between phylogeny based on species and genes.
- **Distinguish** between homology and homoplasy, and between orthologs and paralogs, recognizing their importance in phylogenetic analysis.
- **Interpret** the structure and different representations of phylogenetic trees (rooted, unrooted, bifurcating, multifurcating, ultrametric, additive).
- **Outline** the five main steps in phylogenetic tree construction, from molecular marker selection to tree evaluation.
- **Compare and contrast** distance-based and character-based methods for tree building, detailing algorithms like UPGMA, Neighbor-Joining, Parsimony.
- **Discuss** the role of evolutionary models and statistical measures (e.g., branch length, P-values, bootstrap values) in constructing and evaluating phylogenetic trees.
- **Utilize** Biopython tools for handling phylogenetic tree data and integrating with external programs.
- **Identify** and address practical challenges and limitations in phylogenetic inference.

1. Evolution: The Universal Principle of Biological Change

The concept of **evolution** forms the core foundation of biology and is indispensable for understanding biological data, particularly in the field of molecular phylogenetics. Evolution is fundamentally defined as the **development of a biological form from pre-existing forms**, detailing its journey from origin to its currently existing state through modifications and natural selections. Life on Earth is recognized as a complex, self-perpetuating system that is distributed across both space and time. We observe today only a snapshot of life's history, which has been proceeding for at least 3.5 billion years.

The study of **molecular phylogenetics** is the specialized field dedicated to analyzing this history. It seeks to *reconstruct the evolutionary relationships of genes and other biological macromolecules by examining the accumulated changes, or **mutations**, within their sequences and developing hypotheses about their relatedness*. The culmination of this analysis is usually presented as a **phylogenetic tree**, a diagram that summarizes the reconstructed evolutionary history.



The sketch of a phylogenetic tree by Darwin. Source: amnh.org

The Molecular Basis and Mechanism of Evolution

The hereditary blueprint for biological organisms is stored in their **genetic material**, deoxyribonucleic acid (**DNA**), which contains a message written in a four-letter alphabet. The main role of DNA is information storage, encoding all molecules necessary for life. The genetic information flows according to the **Central Dogma** of molecular biology: DNA is transcribed into RNA, which is then translated into **proteins**. Proteins are the primary working components of organisms, directing nearly all life processes, and their functional capabilities are ultimately determined by their sequences.

The Role of Mutation and Variation

Although the integrity of the genetic information in DNA is carefully protected (for instance, during replication), *errors are an inevitable occurrence within the genome*. These errors are crucial because they provide the necessary **genetic variation** upon which evolution acts. The change in the genomic nucleotide sequence is generally referred to as a **mutation**.

Mutations can occur on a range of scales:

1. **Small-Scale Changes:** These often involve just a single base being incorrectly replicated. These accumulated changes include **substitutions** (point mutations), as well as **insertions and deletions** (**indels**), which cause sequences to diverge over evolutionary time.
2. **Large-Scale Changes:** These include:
    - **Duplication:** A segment of a chromosome is copied and inserted, leading to repeated genetic material. (A major driver of evolution.)
    - **Reversals:** A segment of a chromosome is reversed end-to-end.
    - **Deletions:** a segment of genome is lost.

Mutations that are maintained and passed on to subsequent generations are referred to as **accepted mutations**. Over very long periods, the accumulation of these preserved mutations can lead to the formation of entirely new species.

Natural Selection and Selective Pressure

The fundamental driving force behind evolution is **natural selection**. Natural selection acts to preserve the forms best suited to survive in a particular environment, while eliminating "unfit" forms due to environmental conditions or sexual selection. The fate of any given mutation—whether it is retained in the population or lost—depends heavily on the **selective pressure** exerted on the organism at the time.

Note: Mutations occur to the **genotype** (the genetic material) but natural selection operates on the **phenotype** (the organism itself).

For example **thalassemia** is a genetic disorder that has been selected for immunity to malaria.

Molecular analysis provides a key way to observe this pressure by distinguishing between two types of nucleotide substitution in protein-coding regions:

- **Synonymous Substitutions:** These nucleotide changes *do not result in a change to the encoded amino acid* sequence, due to the redundancy of the genetic code. Synonymous mutations often serve as a baseline for random genetic drift.
- **Nonsynonymous Substitutions:** These changes result in an alteration to the encoded amino acid sequence.

If the rate of nonsynonymous substitution is significantly higher than the synonymous substitution rate, it suggests **positive selection** (or adaptive evolution) is at work, potentially contributing to the *evolution of new functions*. Conversely, if the synonymous rate is higher, it implies that changes at the amino acid level are generally not tolerated, suggesting the sequence is under **negative** (or purifying) **selection** for *conservation of the current structure/function*.

Molecular Evidence: Homology, Similarity, and Ancestry

The genetic material (DNA and protein sequences) contains the record of these accumulated mutations and thus serves as **molecular fossils**. By comparing the sequences of related organisms, *specific regions that are crucial for structural or functional roles tend to be consistently conserved, enabling the identification of shared ancestry.*

The process of **sequence alignment** is used to identify these patterns of conservation and variation by attempting to locate equivalent regions of two or more sequences to maximize their perceived resemblance.

A crucial distinction exists in evolutionary analysis between **similarity** and **homology**:

- **Similarity** is a direct observation or measurement of resemblance between sequences (e.g., quantified as the percentage of aligned residues that match).
- **Homology** is a qualitative statement or **inference** that the similarity observed is due specifically to **descent from a common ancestor**.

If the measured sequence similarity is sufficiently high, researchers can confidently infer a common evolutionary relationship i.e. homology (Chapter 3).

If two sequences appear similar but the resemblance is due to independent evolution (like **convergent evolution**) rather than common ancestry, this similarity is termed **homoplasy**. For example: Some insects have hemoglobin but insects do **not** descend from a common ancestor with vertebrates that already used hemoglobin as an oxygen carrier.

Homology can arise through two major historical events:

1. **Speciation Events:** When a gene diverges due to the evolution of two different species from a common ancestor, the resulting homologous sequences are called **orthologous sequences** (**orthologs**). Orthologs often retain the same biochemical function. *Only orthologous sequences will identify the speciation times in a phylogenetic analysis.* Examples; let-7 RNA family, ribosomal proteins.
2. **Gene Duplication Events:** When a gene is copied *within the same genome*, the resulting related genes are called **paralogous genes** (**paralogs**). Paralogs exist within the same species and can diverge over time, potentially developing new functions. Example: A single, ancient gene existed in an early vertebrate ancestor that produced was a simple oxygen-binding protein. Later it duplicated and the two copies evolved separately and became *myoglobin* (stores oxygen in muscle tissue) and *hemoglobin* (transports oxygen in blood).

Reconstructing Evolutionary History

The divergence observed in homologous sequences, accumulated over time, constitutes the **evolutionary distance** (or genetic distance) between the sequences. This distance needs to be *quantified using various models of molecular evolution* to serve as the raw data for phylogenetic reconstruction.

Molecular phylogenetics aims to trace this history back to a **last common ancestor** and reconstruct the overall evolutionary relationship of species, often referred to as the **tree of life**. The final **phylogenetic tree** visually represents the reconstructed history, using nodes and branches to show the topology (branching pattern) of descent, providing a critical tool for evolutionary research.

2. Gene Phylogeny (Gene Tree) vs Species Phylogeny (Species Tree)

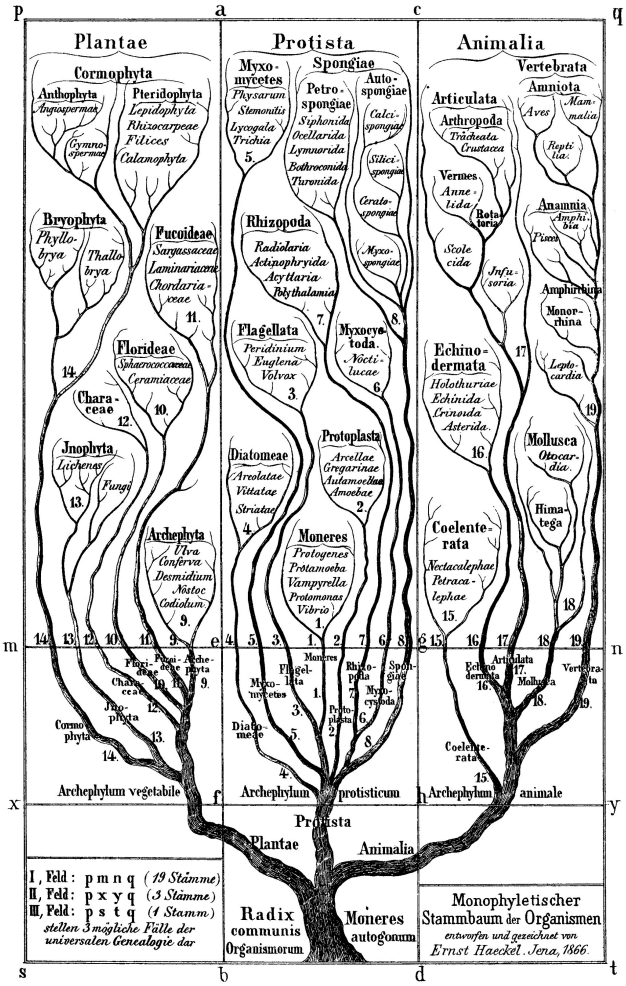
The distinction between **Gene Phylogeny** (Gene Tree) and **Species Phylogeny** (Species Tree) is critical in molecular evolution because the evolutionary history inferred from a single gene sequence does not necessarily reflect the true evolutionary history of the species from which the gene was sampled.

1. Definitions and Fundamental Differences

Species Phylogeny (Species Tree)

These trees were constructed by observing species traits. In those times:

- **Homology:** A trait shared by two species because it was inherited from a **common ancestor**. (e.g., the forearm bone structure of a human, a cat, a whale, and a bat). Homologous traits are the "true signals" of evolutionary relationship.
- **Homoplasy:** A trait that looks similar but evolved **independently** (like the wings of birds and insects). Homoplasy is "misleading noise" that early taxonomists had to carefully avoid.
- **Representation:** A species tree reconstructs the **evolutionary relationships between different species** (taxa or operational taxonomic units).
- **Internal Nodes:** The branching point at an internal node in a species tree represents a **speciation event**.
- **Construction Goal:** The goal is to accurately model how ancestral species diverged into current species. Ideally, species phylogenetic trees should be constructed using **only orthologous sequences**.



The First Darwinian Phylogenetic Tree from 1866, by Ernst Haeckel.
Source: Science Direct

Gene Phylogeny (Gene Tree)

- **Representation:** A gene phylogeny, or gene tree, describes the **evolutionary history of a particular gene or protein sequence** (the "molecular fossil"). It shows the relationships between members of a family of homologous genes or proteins.
- **Internal Nodes:** The internal nodes in a gene tree can represent two types of events:

1. **Speciation Events:** If the sequences diverged due to the evolution of two different species from a common ancestor (resulting in **orthologs**).
2. **Gene Duplication Events:** If a gene was copied within the same genome, followed by divergence of the copies within the same species (resulting in **paralogs**).

- **Construction Goal:** The tree charts the way an ancestral gene became duplicated, reduplicated, and subsequently diverged in function or sequence within genomes and between species.

2. The Lack of Correlation and Sources of Discrepancy

A phylogenetic tree inferred from a gene sequence may have an evolutionary history that is *different from, or **incongruent with**,* the evolutionary path of the species. The species evolution is the cumulative result of the evolution of multiple genes across the entire genome.

The main events that cause discrepancies between gene and species phylogenies include:

2.A. Gene Duplication and Paralogy

The most frequent source of misalignment between gene and species trees is gene duplication.

- **The Problem:** If a gene duplication event precedes a speciation event, the resulting gene tree *will show the divergence of the duplicated genes* (paralogs) rather than the split between the species (orthologs). To construct a true species tree, researchers must be able to **distinguish orthologs from paralogs**, which is often complex, especially in large protein families.

2.B. Gene Loss (and Pseudogenes)

After gene duplication, *one copy may become nonfunctional* through mutation, a process known as **gene loss**. Gene loss can occur due to mutations destroying control sequences (preventing expression) or modifying the protein sequence (rendering it inactive), potentially leading to a **pseudogene**.

- **The Problem:** If gene loss occurs in different lineages, the resulting gene tree topology may incorrectly suggest *closer relationships between species that retained the gene copies*, leading to an **erroneous conclusion** about species relationships.

2.C. Horizontal Gene Transfer (HGT)

Also known as lateral gene transfer (LGT), HGT involves the transfer of a gene from one species into the genome of another species. This is significantly different from vertical transmission from parent to offspring.

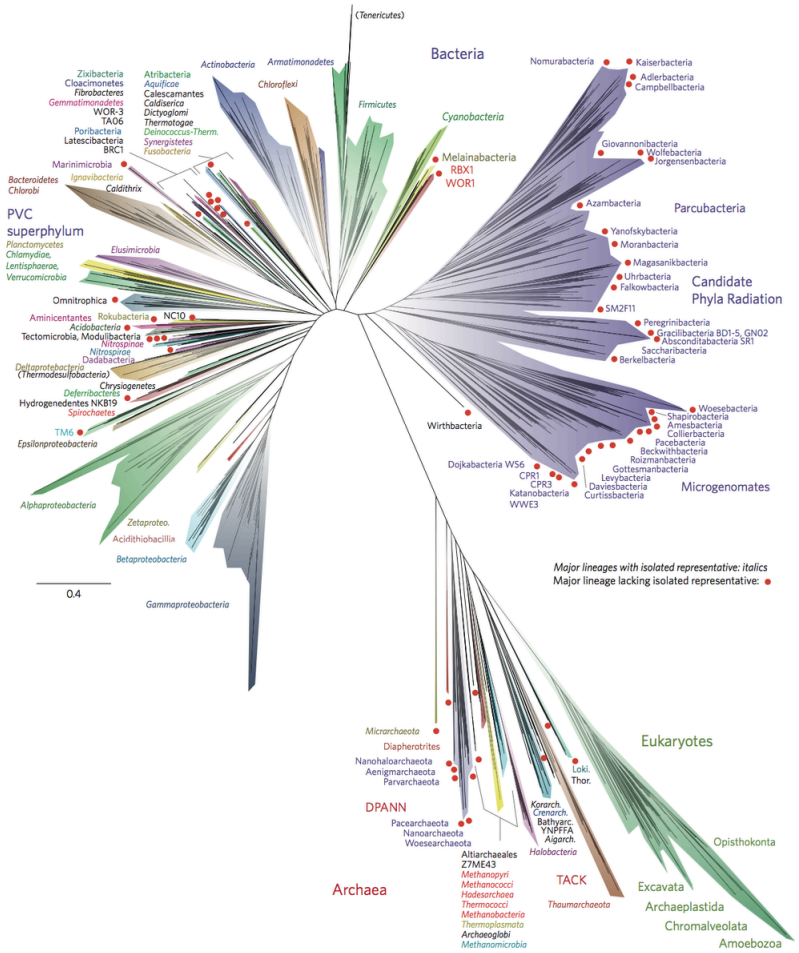
- **Prevalence:** HGT is common among prokaryotes (bacteria and archaea) but was long thought rare in eukaryotes, though *viral genes have been found in the human genome*, acquired by ancient germline infections. **Syncytin-1 and Syncytin-2**: retroviral envelope genes essential for placenta cell fusion. Syncytin-1 exists only in primates.
- **The Problem:** If a gene is transferred horizontally, the recipient species' genome carries an "**xenologous (external) gene**". If included in a standard phylogenetic analysis, the recipient species' gene will appear much more closely related to the donor species' gene than is accurate for the evolutionary history of the organisms themselves.

3. Practical Steps and Solutions

To navigate the difference between gene and species phylogenies, several approaches are used in phylogenetic analysis:

- **Ortholog Selection:** Species trees should rely only on orthologous sequences. Methods like **COG (Clusters of Orthologous Groups)** and **KOG (Eukaryotic Orthologous Groups)** use reciprocal best BLAST hits across multiple genomes to identify clusters of orthologs without necessarily constructing a full tree, assuming members of a COG/KOG share a related function.
- **Comparing Multiple Genes:** To obtain a reliable species phylogeny, phylogenetic trees from a variety of gene families must be constructed to provide an overall assessment of species evolution. If different characters yield inconsistent phylogenetic relationships, they are all dubious.
- **Reconciled Trees:** These specialized diagrams attempt to *combine gene and species trees* to clearly identify the speciation, gene duplication, and gene loss events that have occurred in the history of a gene family.
- **Molecular Markers:** The choice of molecular marker (DNA vs. protein) depends on the evolutionary distance being studied. For reconstructing relationships at the deepest levels (e.g., between bacteria and eukaryotes), **conserved protein sequences** are

preferred over rapidly evolving nucleotide sequences, as *proteins offer a higher signal-to-noise ratio in alignment.*



Tree of life obtained using ribosomal RNA protein sequences.
Source: https://www.nature.com/articles/nmicrobiol201648

3. Evolutionary Models

The development and application of **sequence evolution models** is fundamental to modern bioinformatics, particularly in the field of molecular phylogenetics. These models, often called **substitution models**, provide the mathematical framework necessary to interpret sequence differences as true evolutionary distance and estimate the time-dependent probabilities of specific mutations.

I. The Necessity of Evolutionary Models

All phylogenetic analysis and alignment scoring require a *model of how sequences change over time.* The models are essential because simply counting the observed differences between two sequences (the observed distance) severely **underestimates the true evolutionary divergence**.

A. Observed vs. True Distance

The simplest measure of distance between two aligned sequences is the p-distance (fractional alignment difference).

- **p-distance Formula (Observed Difference):** If an alignment of two sequences has L positions (excluding gaps) of which D differ, the p-distance is defined as:

$$p = \frac{D}{L}$$

- When sequences have diverged significantly, *multiple substitutions may have occurred at the same site*, or mutations *may have reverted to the original state*. (This effect, where sequence similarity is the result of parallel or convergent evolution rather than direct common ancestry, is called **homoplasy**.) Since the p-distance counts only the single observable differences, it *fails to account for these hidden events*.

- **Distance Correction:** Evolutionary models attempt to correct for homoplasy by *predicting the amount of multiple mutation that has occurred*, converting the observed p-distance into a more accurate **corrected evolutionary distance** (d).

B. Two Roles of Evolutionary Models

Evolutionary models are used in two distinct ways in phylogenetic methods:

1. **Distance Correction:** For distance-based phylogeny methods (like Neighbor-Joining), the model is used to derive an equation that converts the p-distance into a *corrected evolutionary distance* (d).
2. **Probability Calculation:** For character-based methods (like Maximum Likelihood), the model provides a formula to calculate the **probability of specific mutations** having occurred.

II. Nucleotide Substitution Models

Nucleotide models incorporate observed features of DNA evolution, such as substitution rates and base composition, to refine distance estimates.

A. Jukes–Cantor (JC) Model

The Jukes–Cantor (JC) model is one of the simplest models specifically designed for nucleotide sequences.

- **Assumptions:** The model assumes that all base sub=stitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, T $\leftrightarrow$ C, etc.) are **equally probable**. It also assumes equal base composition (frequency), and that *each base can mutate to any other with rate* α in a unit of time. This implies that

$$d_{JC} = 3\alpha t.$$

- We want to convert an *observed p-distance* to the *corrected JC evolutionary distance* d_{JC}. If $P(t)$ is the probability that a site has **changed** to a different nucleotide after time t, then it satisfies the differential equation:

$$dP/dt = 3\alpha(1-P) - \alpha P.$$

  Solving it we get:

$$P(t) = \frac{3}{4}(1 - e^{-4\alpha t}).$$

  If we set this equal to the p-distance and we get:

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

B. Kimura Two-Parameter (K2P) Model

Remember that A and G nucleotides (called **Purines**) have a different structure from U, T and C (which are called **Pyrimidines**).
The K2P model is an advancement that recognizes the difference mutations inside each group (called **Transitions**) and mutations from one group to another (called **Transversions**).

The K2P model recognizes that *transitions are generally much more frequent than transversions in accepted mutations.* Thus, it uses two mutation rates (parameters).

- **Distance Formula:** The corrected K2P evolutionary distance d_{K2P} is calculated using the *observed* probabilities (frequency) of transition (P) and transversion (Q):

$$d_{K2P} = -\frac{1}{2}\ln(1-2P-Q) - \frac{1}{4}\ln(1-2Q)$$

C. General Time-Reversible Model

The GTR model, also known as the **REV** model, is among the most general and *commonly used* models.
It uses *different rates of substitution for all six possible nucleotide interchanges* (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) and also permits non-uniform base compositions (i.e. each nucleotide can have its own frequency).

The model is described as *time-reversible*, meaning the probability of transition from state i to j is the same as j to i over a long period of time. *It does not mean that evolution is time-reversible.*

III. Modeling Rate Heterogeneity and Complexities

Beyond substitution types, biological evolution features *variable rates of change across the length of the sequence*, requiring more advanced modeling techniques.
The assumption that the mutation rate is constant across all positions of a sequence is often biologically unrealistic. The **Gamma distance correction** (d_G) specifically accounts for **mutation rate variation at different sequence positions**.

- **Rate Heterogeneity**: Models often incorporate a **Gamma distribution (G)** to account for variation in mutation rates across different sequence positions, as *some sites evolve faster than others.*
- This correction is often combined with other models (e.g., JC+G or REV+G).

IV. Models of Protein Sequence Evolution (Protein Distance Correction)

Specialized models are generally needed due to the 20 amino acid states, which are modeled using 20×20 rate matrices.

A simple protein model, analogous to the JC nucleotide model (assuming equal base composition and single rate for all mutations), can be analyzed to yield a distance correction equation where p is the fraction of identical sites:

$$d = -\frac{20}{19}\ln\left(1 - \frac{19}{20}p\right)$$

However this model is too simplistic. One of the common methods used for proteins today is the **LG (Le & Gascuel, 2008)** model, which is *constructed in a manner similar to PAM*, but much more diverse set of sequences.

4. Types of Phylogenetic Trees

A phylogenetic tree is a diagram that proposes a hypothesis for the reconstructed evolutionary relationships between a set of objects (taxa).
We want the distance of taxa in the tree (length of the shortest path connecting them) to be proportional to their evolutionary distance computed from sequences as above.

- **Terminology**:
  - **Taxa (Operational Taxonomic Units, OTUs)**: The individual genes, proteins, or species represented at the tips of the branches (leaves).
  - **Branches**: Lines connecting nodes, representing the evolutionary lineages. Their lengths often represent evolutionary distance or time.
  - **Nodes**:
    - **Internal nodes**: Represent hypothetical ancestral sequences or species.
    - **Root node**: The bifurcating point at the very bottom, representing the common ancestor of all members of the tree.
  - **Topology**: The branching pattern of the tree.
- **Forms of Tree Representation**:
  - **Rooted vs. Unrooted Trees**:
    - **Unrooted Tree**: Illustrates the relationships between taxa *without specifying a common ancestor for all.*
    - **Rooted Tree**: Infers the most recent common ancestor of all taxa in the tree. It implies a specific direction of evolutionary time. Rooting is often done using an **outgroup**, a sequence or group known to be evolutionarily distant from the main taxa but still related enough to be aligned.
  - **Bifurcating vs. Multifurcating Trees**:
    - **Bifurcating Tree**: A tree where each internal node splits into exactly two daughter branches. This is the common assumption in phylogenetics, implying a clear sequence of speciation events.
    - **Multifurcating Tree**: A tree with one or more internal nodes having more than two branches. This usually indicates insufficient evidence to fully resolve the tree or rapid evolutionary events.

- **Ultrametric vs. Additive Trees**:
  - **Ultrametric Tree**: A rooted tree where *all branches leading from the root to any leaf have the same total length*. This implies a constant molecular clock. Example: trees obtained using UPGMA method, below.
    - Branch length = **expected number of substitutions per site**.
  - **Additive Tree**: A tree (rooted or unrooted) where the distance between any two leaves is the sum of the lengths of the branches connecting them.
    - Branch length = **time**, often scaled to millions of years.

4. Procedure for Phylogenetic Tree Construction

Constructing a molecular phylogenetic tree typically involves five main steps :

- **Step 1: Choice of Molecular Markers (Sequence Data)**:
  - **DNA vs. Protein**: The choice depends on the divergence time of the organisms and the study's purpose.
    - **DNA Sequences**: Evolve more rapidly than proteins and are suitable for studying **closely related organisms** (e.g., individuals within a population, using noncoding mitochondrial DNA). Nucleotide sequences also contain **synonymous mutations** (which don't change amino acids) that provide additional evolutionary information.
    - **Protein Sequences**: Evolve more slowly due to stronger purifying selection. They are preferred for studying **more widely divergent groups** or when the phylogenetic signal in DNA has been lost due to extensive mutations (e.g., between bacteria and eukaryotes).
- **Step 2: Multiple Sequence Alignment (MSA)**:
  - **Critical Prerequisite**: Obtaining accurate alignment is critical for phylogenetic tree construction. A good MSA correctly identifies *homologous positions* across all sequences, which is fundamental for calculating evolutionary distances or comparing characters.
  - **Refinement**: MSA for phylogenetic analysis often requires **manual removal of ambiguously aligned regions** (e.g., highly divergent or *gap-heavy* areas) to avoid introducing noise. Any additional information, such as protein structure or key functional conserved amino acids, should be included to try to make the alignment as accurate as possible.
  - **Note:** gaps are usually considered as *missing data* in phylogenetic analysis, meaning that *they do not contribute to distances or likelihoods.*
- **Step 3: Choice of Evolutionary Model**:
  - Evolutionary models help **correct observed sequence differences into more accurate (corrected) evolutionary distances** by accounting for these phenomena.
- **Step 4: Determining a Tree Building Method**: There are two main categories of methods for reconstructing phylogenetic trees :
  - **A. Distance-Based Methods**:
    - **Principle**: These methods first calculate a matrix of evolutionary distances between all pairs of sequences (taxa) from the MSA, after correcting for multiple substitutions using an evolutionary model. The tree is then constructed based solely on this distance matrix.
  - **B. Character-Based Methods**:
    - **Principle**: These methods work directly with the sequence alignment (treating each column as a "character") rather than a distance matrix. They *search for the tree that best explains the observed character states* based on an evolutionary model.
- **Step 5: Phylogenetic Tree Validation**
  - Methods such as *Bootstrap* are used to validate phylogenetic trees. Columns are sampled from the MSA (allowing repetition). This gives us a new MSA and thus, a new phylogenetic tree. Repeating this process we can see a branch in the original tree is present in how many of the new ones. This way we can see how *stable* each branch of the original tree is.

5. Tree Building Methods: Distance-Based

Distance-based methods are a category of algorithms used for phylogenetic tree construction that rely on converting a multiple sequence alignment (MSA) into a *matrix of pairwise evolutionary distances*. These methods are fundamentally **phenetic**, meaning they group sequences based on overall similarity *rather than explicitly modeling the historical events* (genealogy) at each site.

As mentioned above, starting from an MSA, let $d_{i,j}$ be the *corrected evolutionary distance* between sequences i and j obtained using a choice of *evolutionary model*.

The evolutionary distances calculated using these models are then compiled into an $n \times n$ *distance matrix* D used for tree construction.

Clustering algorithms gradually build a phylogenetic tree *starting from the most similar (closest) pair of sequences*/clusters and progressively merge them. They produce a single tree and are computationally fast, making them suitable for large datasets.

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

The UPGMA method is a simple clustering technique that assumes a **constant molecular clock**.

- **Principle:** UPGMA starts by finding the two **closest taxa** in the distance matrix and merging them into a new cluster (node). This process is repeated until a single root remains.
- **Assumption:** Because it assumes a constant rate of evolution (**molecular clock**), the *distance from the root to any leaf must be the same*. It therefore produces a **rooted ultrametric tree**.
- **Cluster Distance Calculation:** The distance between a newly formed cluster C_{new} (merged from C_i and C_j) and any existing cluster C_m is calculated as the arithmetic average of all pairwise distances between their components:

$$D(C_{\text{new}}, C_m) = \frac{D_{C_i,C_m} \cdot |C_i| + D_{C_j,C_m} \cdot |C_j|}{|C_i| + |C_j|}$$

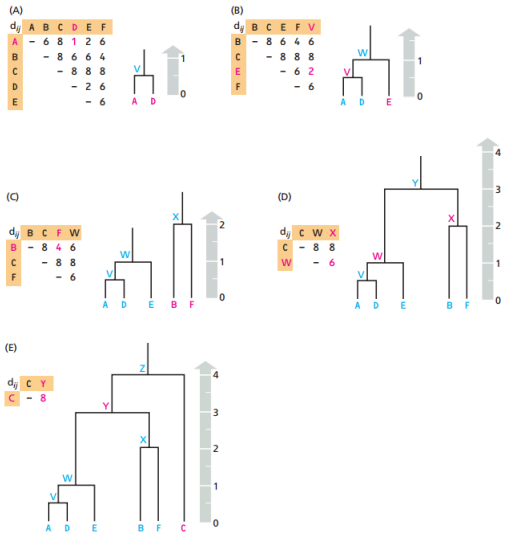
  Where $|C|$ denotes the number of leaves in cluster C.

Note: if $d(x, y)$ is a distance function (say, between sequences), then we can generalize it to a distance function between sets (of sequences) by:

$$d(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y)$$

which is the average of the distances of the points in the two sets to each other. As an exercise you can show that the above formula for D is compatible with this d.

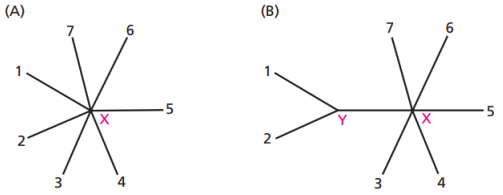
- **Node Age/Branch Lengths:** The *age (or height) of the new node* C_{new} is half the distance between C_i and C_j: $AGE(C_{\text{new}}) = D_{C_i,C_j}/2$. It signifies the **time to the common ancestor**.



Constructing a phylogenetic tree using UPGMA. Note that V is the set containing A and D, and its distance to, say, B is the average of the distance of these two points to B.

B. Neighbor-Joining (NJ)

The NJ method, developed by Saitou and Nei (1987), is a robust method that **does not assume a constant molecular clock** (constant rate of mutations for different taxa). It corrects for unequal evolutionary rates by using a conversion step.



- **Principle:** NJ is related to the concept of **minimum evolution**. It begins with a completely unresolved **star tree** (where all taxa radiate from a single node) and progressively identifies and joins pairs of neighboring leaves, decomposing the tree until all nodes are resolved. It produces an **unrooted additive tree**.
- **Neighbor-Joining Matrix (D^*):** The core idea of NJ is transforming the distance matrix D into a matrix D^* where the smallest element is guaranteed to correspond to a true pair of neighboring leaves.
  - First, calculate the total distance for each leaf i:

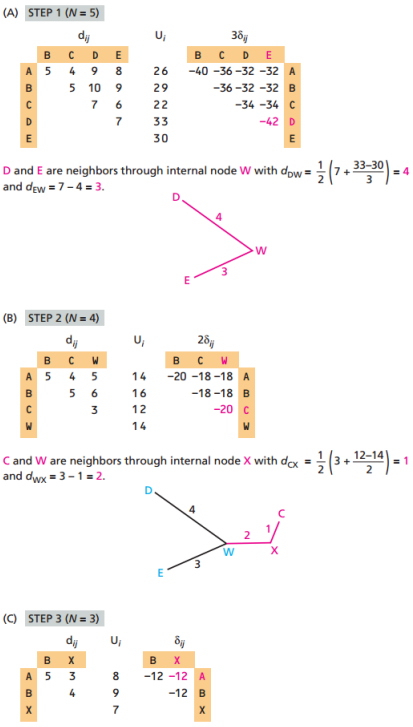
$$TD(i) = \sum_{1 \le k \le n} D_{i,k}.$$

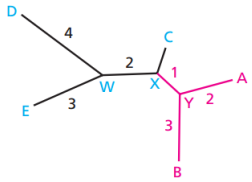
    A taxon that is generally very divergent (far from most others) will have a large TD. In NJ we subtracts $TD(i) + TD(j)$ from $D_{i,j}$ *to avoid pairing two "long" taxa just because both are far from everyone else.*
  - The neighbor-joining matrix $D^*_{i,j}$ is defined for $i \neq j$ as:

$$D^*_{i,j} = (n-2) \cdot D_{i,j} - TD(i) - TD(j)$$

- The two nodes with the lowest D^* distance are joined, into a new node Y.
- **Distance to New Node (m):** Once the neighbor pair i and j is identified, they are replaced by a new leaf Y. The distance from Y to any other leaf k is calculated as:

$$D_{k,Y} = \frac{1}{2}(D_{k,i} + D_{k,j} - D_{i,j})$$





An example of tree construction using NJ. The 3δ or 2δ are the same as D^* above.

IV. Pros and Cons of Distance-Based Methods

Distance-based methods are compared to **character-based methods** (like Maximum Parsimony and Maximum Likelihood), which use the molecular sequences directly.

Advantage	Disadvantage
Computational Speed: Distance methods are computationally fast, capable of handling very large datasets (e.g., thousands of sequences).	**Information Loss:** The conversion of sequence alignment data into a single distance value results in the loss of detailed sequence information.
Substitution Models: They allow the use of numerous substitution models to correct distances, which accounts for hidden multiple mutations.	**Ancestral Reconstruction Impossible:** Due to information loss, distance methods **do not allow the reconstruction of ancestral sequences** at internal nodes.
Heuristic Robustness: Algorithms like Neighbor Joining are highly effective heuristics that perform well even when the evolutionary rates are unequal.	**Inaccurate for Large Datasets (Exhaustive Search):** While faster than character methods overall, optimality-based distance methods (FM, ME) become computationally prohibitive for large numbers of taxa (e.g., > 12 or more) because of the vast number of tree topologies to check.
Statistical Clarity (Optimality): Optimality methods (FM, ME) provide clear criteria (a score S) to compare alternative tree topologies.	**Molecular Clock Reliance (UPGMA):** The simplest clustering method (UPGMA) relies on the unrealistic assumption of a constant molecular clock, often producing erroneous tree topologies.

6. Example of Tree Construction using UPGMA

1) The MSA sequences (length = 12 nt each)

```
A: ATGCTAGCTAAG
B: ATGCTAGCTGAG
C: ATGTTAGCTGAG
D: ATGTTGGCTGAC
```

2) Compute pairwise differences (counted position-by-position)

A vs B

A: A T G C T A G C T A A G
B: A T G C T A G C T G A G
Differences: only at position 10 (A has A, B has G) → d = 1
p-distance = 1 / 12 = 0.083333...`

A vs C

A: A T G C T A G C T A A G
C: A T G T T A G C T G A G
Differences at positions 4 and 10 → d = 2
p-distance = 2 / 12 = 0.166666...`

A vs D

A: A T G C T A G C T A A G
D: A T G T T G G C T G A C
Differences at positions 4, 6, 10, 12 → d = 4
p-distance = 4 / 12 = 0.333333...`

B vs C
B: A T G C T A G C T G A G
C: A T G T T A G C T G A G
Difference at position 4 only → d = 1
p-distance = 1 / 12 = 0.083333...`

B vs D
B: A T G C T A G C T G A G
D: A T G T T G G C T G A C
Differences at positions 4, 6, 12 → d = 3
p-distance = 3 / 12 = 0.25`

C vs D
C: A T G T T A G C T G A G
D: A T G T T G G C T G A C
Differences at positions 6 and 12 → d = 2
p-distance = 2 / 12 = 0.166666...`

3) Distance matrix (counts and p-distance)

p-distances (d/12):

	A	B	C	D
A	0.000	0.083	0.167	0.333
B	0.083	0.000	0.083	0.250
C	0.167	0.083	0.000	0.167
D	0.333	0.250	0.167	0.000

At this stage we must compute the corrected distances using an evolutionary model. However for the sake of presentation and simplicity we omit this step here and use p-distances directly.

4) UPGMA tree construction

We merge the pair with the smallest distance, then updates distances by averaging.

Step 1 — smallest distance

The smallest pairwise distance is 1 (A–B and B–C tie). We'll pick **A & B** first.

Merge (A,B) → cluster (AB).
Height of node (AB) = distance(A,B)/2 = 1 / 2 = 0.5.

Step 2 — compute distances from (AB) to the others

UPGMA sets

$$d((AB), X) = \frac{d(A, X) + d(B, X)}{2}.$$

d((AB),C) = (d(A,C) + d(B,C))/2 = (2 + 1) / 2 = 1.5
d((AB),D) = (d(A,D) + d(B,D))/2 = (4 + 3) / 2 = 3.5

Distance matrix now (clusters): (AB), C, D:

	(AB)	C	D
(AB)	0	1.5	3.5
C	1.5	0	2
D	3.5	2	0

Step 3 — next smallest distance

Smallest is 1.5 between (AB) and C. Merge **(AB)** & **C** → cluster (ABC).
Height of node (ABC) = 1.5 / 2 = 0.75.

Branch lengths at this merge:

- The previous node (AB) had height 0.5. The new internal node (ABC) has height 0.75. So branch length from (AB) node up to (ABC) = 0.75 − 0.5 = 0.25.
- Branch length from C (a leaf with height 0) up to (ABC) = 0.75 − 0 = 0.75.

Step 4 — compute distance from (ABC) to D

Now

$$d((ABC), D) = \frac{d(A, D) + d(B, D) + d(C, D)}{3} = (4 + 3 + 2)/3 = 9/3 = 3.0.$$

Only two clusters left: (ABC) and D, distance = 3.0. Merge to root.

Root height = 3.0 / 2 = 1.5.

Branch lengths at final merge:

- Branch from (ABC) to root = 1.5 − 0.75 = 0.75.
- Branch from D (leaf) to root = 1.5 − 0 = 1.5.

5) Final tree with branch lengths

We can write the tree (ascii) with heights / branch lengths:

```
            /- A
   0.25 -|
         \- B
0.75 -|
      |           /- C
      \0.75 ----|
                \- D
```

Or more explicitly (leaf — branch length):

- A: path = A —0.5→ (AB) —0.25→ (ABC) —0.75→ root
  total path length root→A = 1.5 (as expected for ultrametric; UPGMA forces ultrametricity)
- B: same as A.
- C: branch lengths: C —0.75→ (ABC) —0.75→ root
- D: branch length: D —1.5→ root

7. Tree Construction Methods: Character-Based

I. Core Principles of Character-Based Methods

Character-based methods represent a family of phylogenetic approaches that use the **actual sequence characters** (nucleotides or amino acids) from a multiple sequence alignment (MSA) to infer evolutionary events.

In these methods, the tree is not constructed but, given an MSA, a score is assigned to all possible trees (whose leaves correspond to the taxa in the MSA) and the tree with the optimal score is chosen.

A. Foundational Assumptions

1. **Independent Evolution:** Each character (site) is treated as an individual evolutionary unit that **evolves independently** of other sites.
2. **Inference of Ancestry:** A major strength is the ability to ***infer ancestral sequences** at the internal nodes of the tree*.

There two main character based methods: **Maximum Parsimony** and **Maximum Likelihood**.

II. Maximum Parsimony (MP)

The philosophy behind MP is derived from the principle of **Occam's razor**—the *simplest explanation* (the one requiring the fewest assumptions or leaps of logic) is probably the correct one. In evolution, the simplest scenario is often equated with the *fewest mutation events.*

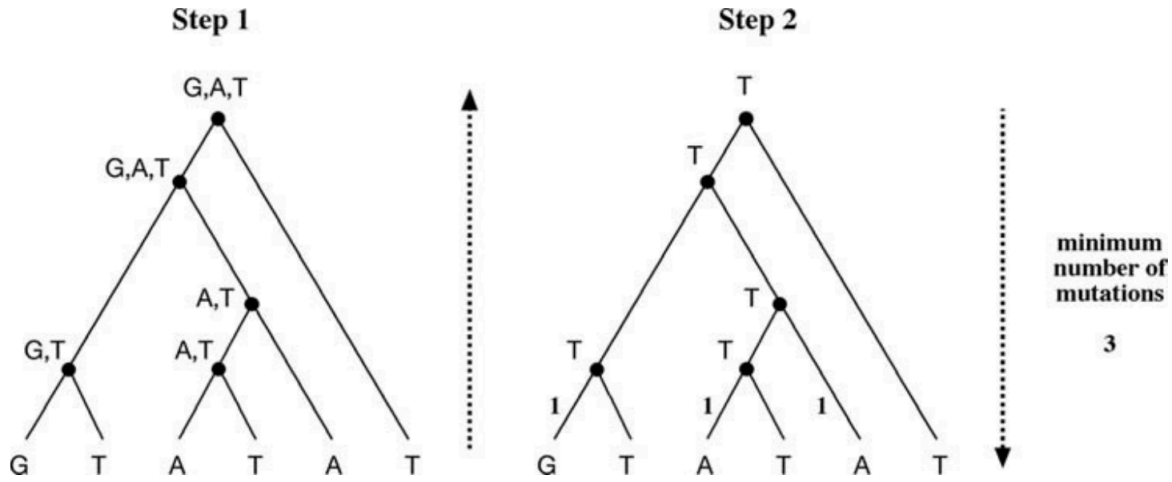
- **Parsimony Score (S):** The score of an evolutionary tree is the **total number of mutations** (substitutions) required to account for the differences observed among the extant sequences.
  - If a tree T has nodes labeled by strings of length m, the length of an edge (v, w) is the Hamming distance between the strings labeling v and w.
  - The **parsimony score** is the sum of the lengths of all edges.
  - The optimal MP tree minimizes this total score S.

Informative Sites

To reduce computational time, MP often focuses only on sites deemed **informative**.

sites / taxa	1	2	3	4	5	6	7	8
I	A	A	T	T	A	G	C	T
II	G	G	T	C	G	T	A	G
III	A	A	T	G	C	G	C	T
IV	A	G	T	A	A	G	C	A
V	A	C	T	T	C	G	C	G
VI	A	C	A	T	G	G	C	A

- An informative site is defined as one where there are at least **two different kinds of characters** (e.g., bases A and G) present, and each of those characters occurs at least **twice**.
- Sites that are constant (all the same character) or singleton sites (a unique change occurring only once) are **noninformative** because they do not help distinguish between alternative tree topologies.



Left: picking a site in the alignment, bases (characters) are associated to the leaves of the tree. Going up the tree, we associate sets of bases to each internal node.
Right: Going from the root to the leaves and assigns ancestral characters that involve minimum number of mutations. (Used only to reconstruct specific ancestral states, not needed for scoring.)

Imagine an MSA is given, we have chosen a site in it, we have a tree T and we have associated the taxa in MSA with the leaves of T. Let S be any assignment of bases to the *internal nodes* of the tree.
Then we can associate a score to each edge in the tree. The score $S(u, v)$ associated to the edge (u, v) is 0 if $S_u \cap S_v$ is nonempty and

1 otherwise (i.e. *mutation*).
The **Parsimony Score** of the tree is the sum of the scores of its edges, however we want the S which minimizes this score for this site:

$$\text{minSubs}(s, T) = \min_{S} \left(\sum_{\text{edges } (u,v)} \text{score}(S_u, S_v) \right)$$

Then we sum over all the *informative* sites in the alignment:

$$PS(T) = \sum_{s=1}^{S} \text{minSubs}(s, T)$$

The $minSubs$ can be found using the *Fitch algorithm*. Going from the bottom to the top of the tree:

1. If $S_1 \cap S_2 \neq \emptyset \Rightarrow S_{\text{parent}} = S_1 \cap S_2$
   (and **no** mutation increment)
2. If $S_1 \cap S_2 = \emptyset \Rightarrow S_{\text{parent}} = S_1 \cup S_2$
   and increment mutations by 1

(Optionally one ca go back from the root to bottom, to reconstruct specific ancestral states, not needed for scoring.)

The **Maximum Parsimony (MP)** method seeks the tree topology that requires the **minimum number of inferred evolutionary changes** or the shortest overall branch lengths to explain the observed sequence data.

$$T^* = \arg\min_{T} PS(T)$$

Example

Alignment

```
          1 2 3 4 5
Human     A T G C T
Mouse     A T G T T
Dog       C T G C T
Cat       C T G C C
```

Example tree we will use (arbitrary, not necessarily optimal)

```
((Human,Mouse),(Dog,Cat));
```

```
       Human
         |
       (HM)
         |
       Mouse
         |
       (root)
         |
       (DC)
         |
        Dog
         |
        Cat
```

Internal nodes:

- **(HM)** = ancestor of Human + Mouse
- **(DC)** = ancestor of Dog + Cat
- **(root)** = ancestor of all four

How one alignment site maps onto this tree: Site 1

```
Human   = A
Mouse   = A
Dog     = C
Cat     = C
```

Step 1 — within the clades:

- Human + Mouse → both **A**
  ⇒ ancestral state at (HM) = **A**
- Dog + Cat → both **C**
  ⇒ ancestral state at (DC) = **C**

Step 2 — Infer the root state

We have two options:

- If root = A, then the root→(DC) branch must mutate A→C (1 change)
- If root = C, then the root→(HM) branch must mutate C→A (1 change)

Either way, the site requires **one change** on the internal branch between the two clades.

So **Site 1 requires exactly one mutation** in this tree.

Example: Site 4

```
Human   = C
Mouse   = T
Dog     = C
Cat     = C
```

Clades:

- Human + Mouse → C and T (different)
  ⇒ ancestral state could be uncertain (could be C or T)
- Dog + Cat → C and C
  ⇒ ancestral state (DC) = **C**

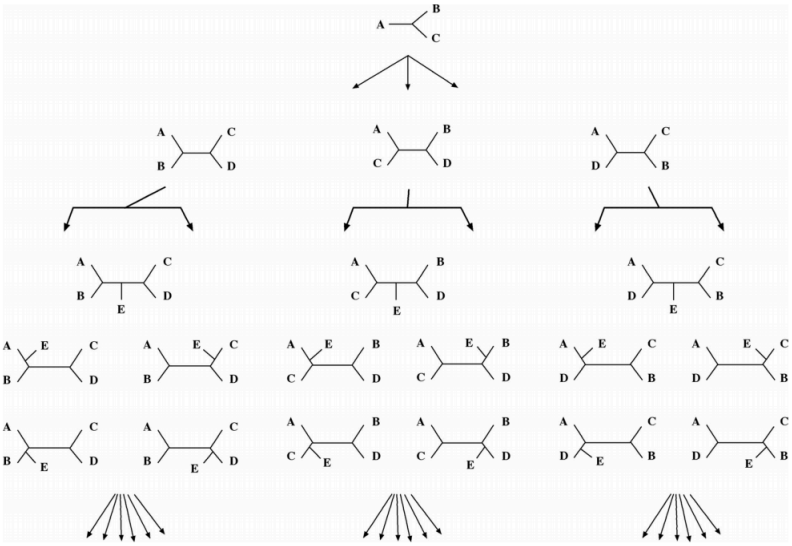
Try root = C (this usually minimizes changes):

- root = C
- (DC) = C (0 changes)
- (HM):
  - Best choice = C (minimizes changes)
  - Human = C (0 changes)
  - Mouse = T (**one C→T change** on branch to Mouse)

So **Site 4 requires 1 mutation**, located on the Mouse branch.

Finding the tree with minimum score

The problem is now to find a tree with minimum score.
For a small number of taxa, exhaustive search can be performed to build the trees recursively as in the following picture:



Since the number of possible trees increases **exponentially** with the number of sequences (taxa), exhaustive search is computationally prohibitive, and the Large Parsimony Problem is known to be **NP-Complete**. Thus, efficient algorithms and heuristics are required.

The Branch-and-Bound Method

The **branch-and-bound** method is a full search method that finds the **exact optimal solution** while dramatically increasing efficiency compared to a brute-force approach.

1. **Establish a Bound:** An initial upper limit, or **upper bound** (S_m), for the total number of substitutions is set. This bound is usually determined by quickly constructing an initial distance tree (like NJ or UPGMA) and calculating its total minimum substitution score.
2. **Stepwise Construction:** The tree is built using **stepwise addition**.
3. **Bounding:** As partial trees are constructed, their current parsimony score is calculated. The central principle exploited is that the total number of substitutions **cannot decrease** when a new sequence is added to a partial tree.
4. **Pruning:** If the score of a partial tree exceeds the current best score found for a complete tree (S_m), the algorithm immediately stops exploring that topological branch.
5. **Optimization:** If a new complete tree is generated with a smaller score S than the current S_m, this new tree becomes the current optimal tree, and S_m is lowered accordingly.

By limiting the tree growth based on the upper bound, the branch-and-bound method significantly reduces computing time while **guaranteeing** to find the most parsimonious tree.

Limitations of Maximum Parsimony

The primary drawbacks of MP are related to its speed and its simple underlying model:

1. **No Evolutionary Models:** MP does not employ complex evolutionary models to correct for multiple substitutions. This leads to an **underestimation of the true number of mutations** when sequences are highly divergent.
2. **Long-Branch Attraction (LBA):** MP is particularly sensitive to LBA, a phylogenetic artifact where two **rapidly evolving taxa (long branches) are incorrectly clustered together** regardless of their true phylogenetic position. LBA can arise even with perfect data.
    - **Figure 8.21** visually illustrates this problem, showing how the method prefers the incorrect topology (B) over the true tree (A) in specific cases involving unequal branch lengths.

III. Maximum Likelihood (ML)

The **Maximum Likelihood (ML)** method is considered the **most rigorous** among phylogenetic approaches because it is based on sound statistical foundations.

ML calculates the **probability (likelihood)** that a specific tree topology T, defined by branch lengths and a chosen evolutionary model, would produce the observed sequence data D.

- **Criterion:** The optimal tree is the one that has the **maximum likelihood** (L).

- **Model Dependence:** ML explicitly uses an **evolutionary substitution model** (e.g., Jukes-Cantor) that provides *time-dependent probabilities for sequence changes.*
- **Data Usage:** ML considers **every position** in the multiple alignment, not just the informative sites.

5. Biopython for Molecular Phylogenetics: A Computational Toolkit

Biopython provides robust support for working with phylogenetic trees, offering a consistent API for I/O and manipulation of tree objects.

- `Bio.Phylo` **Module**: Introduced in Biopython 1.54, this module is specifically designed to process, analyze, and visualize phylogenetic trees.
  - **I/O Functions**: `Bio.Phylo.parse()` and `Bio.Phylo.read()` handle various standard phylogenetic file formats (e.g., Newick, Nexus, phyloXML).
  - **Tree and Clade Objects**: `Tree` objects serve as containers for recursive sub-trees, holding global phylogeny information. `Clade` objects store node- and clade-specific information like branch length and lists of descendent clades.
  - **Visualization**: `Bio.Phylo` supports various ways to view and export trees, including ASCII art for console display and integration with Graphviz for graphical output. Specific features allow coloring branches (e.g., in phyloXML output).
  - **Integration with PAML**: The `Bio.Phylo.PAML` module supports parsing output from PAML programs (like CODEML), which perform Maximum Likelihood analysis of DNA and protein evolution.
  - `Bio.Nexus` **Port**: Some advanced features, like calculating a consensus tree, might still reside in the older `Bio.Nexus` module if not yet ported to `Bio.Phylo`.
- **Practical Workflows**: Biopython can automate steps like:
  - Reading multiple sequence alignments (e.g., from `Bio.AlignIO`).
  - Preparing data for external phylogenetic programs (e.g., converting to PHYLIP format).
  - Parsing the output from these programs (e.g., WebPhylip for distance, parsimony, or Bayesian methods).
  - Analyzing the resulting `Tree` objects to compare topologies, retrieve branch lengths, and evaluate reliability.

6. Challenges and Limitations in Phylogenetic Inference

Despite the sophistication of current methods, several challenges remain in reconstructing accurate phylogenetic trees:

- **Computational Complexity**: For a large number of taxa, the number of possible tree topologies increases astronomically. Exhaustive searches are often impossible, necessitating heuristic approaches.
- **Model Assumptions**: All phylogenetic methods rely on evolutionary models, which are simplifications of complex biological reality. Violations of these assumptions (e.g., non-independent evolution of sites, incorrect molecular clock assumption) can lead to incorrect trees.
- **Data Quality**: The accuracy of the tree depends heavily on the quality of the input multiple sequence alignment and the absence of sequence errors.
- **Homoplasy**: Convergent evolution and other forms of homoplasy can obscure true evolutionary signals, leading to similar sequences that are not genuinely homologous by descent.
- **Gene Tree vs. Species Tree Discrepancies**: As discussed, the evolution of a single gene may not perfectly reflect the evolution of the entire species, especially due to events like horizontal gene transfer. This means "caution is needed in extrapolation of phylogenetic results".
- **Parameter Choice**: The choice of evolutionary model, substitution matrix, and tree-building method can significantly influence the resulting tree. Justifying these choices is an important part of the analysis.

Conclusion:
Molecular phylogenetics is an indispensable tool for understanding the evolutionary tapestry of life. By leveraging the information embedded in molecular sequences, rigorous computational methods allow us to reconstruct ancestral relationships, infer speciation and gene duplication events, and gain profound insights into the history of life on Earth. While challenges remain in managing computational complexity and interpreting results with biological realism, the continuous development of algorithms and user-friendly tools like Biopython empowers researchers to increasingly refine our understanding of evolution.

Exercises

1. Cluster the following points using agglomerative clustering, similar to UPGMA: 0.5, 1, -0.2, 1.5, -1, 3, -0.7.
2. Implement a simplified distance calculation (e.g., p-distance, Jukes-Cantor) from a multiple alignment.
3. Implement a basic UPGMA or Neighbor-Joining algorithm.