Phylogenetic Trees

Phylogenetics is the study of evolutionary history and relationships among organisms. It seeks to reconstruct how species, genes, or populations diverged from common ancestors, forming the tree of life. From the earliest days of natural history (Fig. 27), biologists have sought to classify living forms and understand their relationships. Before the era of molecular biology, these efforts relied on observable features of organisms, such as morphology, anatomy, and physiology. With these data, scientists built hierarchical systems of classification, grouping organisms by shared traits that were presumed to reflect common ancestry.

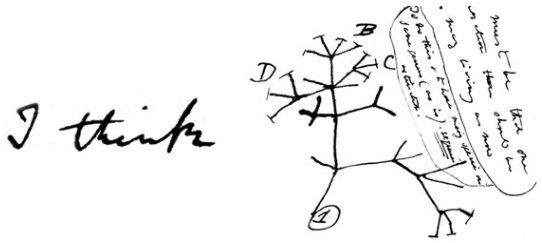

Figure 27: Charles Darwin's "I think" sketch (1837), the first known diagrammatic representation of a phylogenetic tree. Drawn in his *Notebook B*, it captures the idea that all species descend from common ancestors through branching evolution, a concept that became central to modern evolutionary biology.

The motivation for studying phylogeny extends beyond taxonomy. Understanding evolutionary relationships provides crucial insights into how traits and functions evolved, how pathogens and their hosts co-evolve, and how genetic variation shapes adaptation. In medicine, for example, phylogenetic analyses help trace the origin and spread of infectious diseases, identify reservoirs of emerging pathogens, and guide vaccine development by revealing evolutionary patterns in viral genomes.

Phylogenetic inference can draw on many types of data. Historically, biologists compared:

- *Morphological traits*, such as body shape, bone structure, or organ systems;
- *Quantitative traits*, such as size, weight, or other measurable features;

- *Physiological or behavioral traits*, including metabolic rates, temperature tolerance, or mating behavior.

These observable characteristics, though valuable, often reflect both genetic and environmental influences, which can obscure true evolutionary relationships. As a result, early phylogenies based purely on morphology were sometimes inconsistent or simply wrong.

The advent of molecular biology transformed the field. Modern *molecular phylogenetics* reconstructs evolutionary relationships using genetic data that includes DNA, RNA, or protein sequences. Because sequences carry information inherited from common ancestors, they offer a direct record of evolutionary history. Conserved genes, which change slowly over time, are particularly useful for comparing distant species, while rapidly evolving regions reveal recent divergences among closely related taxa.

In this chapter, we focus on methods that infer phylogenetic trees from molecular data. We assume that sequence alignments have been performed and that measures of similarity or distance between sequences are available. Using these distances, we will introduce two classical approaches for tree reconstruction: the *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA) and the *Neighbor Joining* method.

Phylogenetic Trees

A phylogenetic tree (*e.g.* Fig. 29) is a central concept in evolutionary biology. It is a model that represents the inferred evolutionary relationships among a set of organisms or genes. It provides a structured way to visualize how groups of organisms, known as taxa, are related through common ancestry. Each branch in the tree represents a lineage that diverges over time, accumulating genetic changes that lead to the diversity of life we observe today.

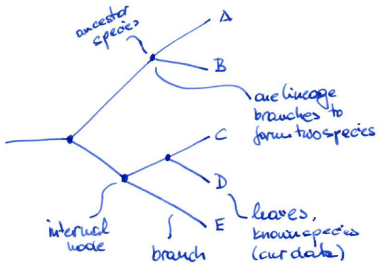

Figure 28: Elements of a phylogenetic tree.

Let us start with some terminology. A *taxon* (plural: *taxa*) refers to any group of one or more organisms that a taxonomist considers to form a unit. A taxon may represent a species, a genus, a family, or a higher-level grouping such as a phylum. The scientific discipline concerned with identifying, naming, and classifying organisms—both living and extinct—is called *taxonomy*. It provides the hierarchical framework within which phylogenetic relationships are interpreted.

In a phylogenetic tree, the *leaves* (or terminal nodes) correspond to known, extant taxa, while the *internal nodes* represent inferred ancestral taxa that no longer exist. The *branches* connecting these nodes trace the evolutionary pathways linking ancestors and descendants.

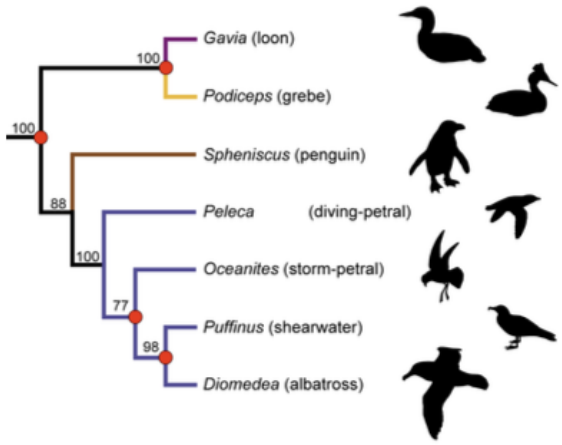

Figure 29: A part of a phylogenetic tree of pelecaniformes from osteological data, as published in Smith ND in *PLOS One* (2010).

In most such representations, the length of a branch is assumed to be proportional to evolutionary time or to the number of accumulated mutations, reflecting the concept of *evolutionary distance*. This assumption implies that mutations occur at a constant rate along all lineages—a simplifying but often useful approximation known as the *molecular clock hypothesis*. While real evolutionary processes may deviate from a perfectly constant rate, this assumption allows us to translate genetic differences into temporal relationships.

Phylogenetic trees can take different forms depending on the data and assumptions (Fig. 30). In a *binary tree*, each internal node gives rise to exactly two descendant branches, representing a series of bifurcations in evolutionary history. Binary trees are rooted, meaning they have a common ancestor at the root. In contrast, an *unrooted tree* depicts relationships among taxa without specifying a common ancestor or a direction of evolution.

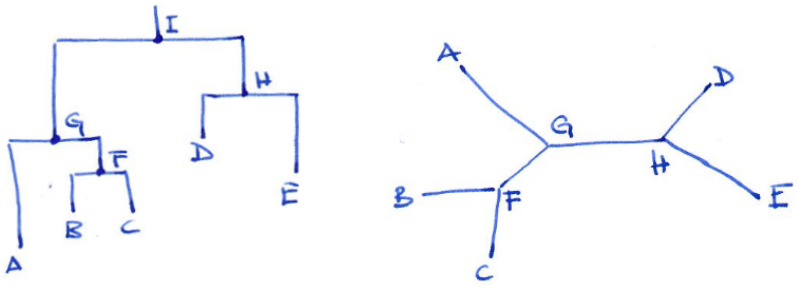

Figure 30: Rooted (left) and unrooted tree (right).

Phylogenetic Trees as Models

To construct a phylogenetic tree from molecular data, we begin with a set of pairwise distances between taxa, denoted by d_{ij} for taxa τ_i

and τ_j. These distances are obtained from aligned genetic sequences and depend on the type of sequence, alignment method, and scoring matrix. From alignments, we estimate the number of substitutions or mutations, possibly applying correction models (such as Jukes–Cantor, discussed later) to account for multiple substitutions at the same site.

A phylogenetic tree provides a model that explains these pairwise distances. If the distances are *additive*, the distance between any two taxa equals the sum of branch lengths connecting them. Thus, theoretical distances can be computed directly from the tree and compared with those derived from sequence data.

For example, consider the unrooted tree in Figure 31, with branch lengths representing evolutionary distances. The corresponding model distances are:

	τ_B	τ_C	τ_D
τ_A	4	5	6
τ_B		3	4
τ_C			3

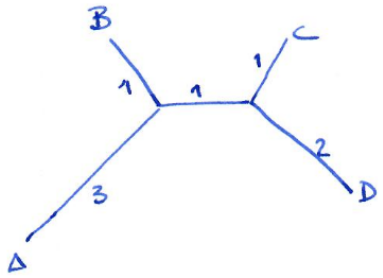

Figure 31: An example tree with edges labeled with distances.

Ideally, these model distances should match those inferred from the aligned sequences. The problem of tree reconstruction is therefore to find the topology and branch lengths such that the modeled distances best approximate the observed ones.

In principle, construction of the phylogenetic tree is an optimization task, where we aim to minimize the discrepancy between observed and expected distances. Since the number of possible tree topologies grows exponentially with the number of taxa, exhaustive search is infeasible. Practical methods, such as those introduced later in this chapter, use heuristic, hill-climbing approaches that iteratively improve the fit between model and data without guaranteeing a global optimum.

Unweighted Pair Group Method with Arithmetic Mean

UPGMA starts with each taxon as its own group and repeatedly merges the two groups with the smallest distance (or highest similarity) until a single group remains. It is identical to hierarchical clustering with average linkage, where for two groups of taxa G_i and G_j the intergroup distance is

$$D(G_i, G_j) = \frac{1}{|G_i|\,|G_j|} \sum_{p \in G_i} \sum_{q \in G_j} d_{pq}.$$

After merging G_i and G_j into $G_k = G_i \cup G_j$, distances to any other group G_ℓ are updated by the size–weighted average

$$D(G_k, G_\ell) = \frac{|G_i|\, D(G_i, G_\ell) + |G_j|\, D(G_j, G_\ell)}{|G_i| + |G_j|}.$$

UPGMA produces an ultrametric tree, a special kind of phylogenetic tree in which all paths from the root to any leaf have equal total length. That means all taxa are equally distant from their common ancestor, implying a constant rate of evolution.

UPGMA: a worked example

Start with the pairwise distances among taxa $\tau_A, \tau_B, \tau_C, \tau_D$:

	τ_B	τ_C	τ_D
τ_A	4	5	6
τ_B		3	4
τ_C			3

Step 1: merge τ_C and τ_D. The minimum distance is $d_{CD} = 3$, so merge $G_E = \{\tau_C, \tau_D\}$. Update distances:

$$D(\tau_A, G_E) = \tfrac{1}{2}(d_{AC} + d_{AD}) = \tfrac{5+6}{2} = 5.5,$$
$$D(\tau_B, G_E) = \tfrac{1}{2}(d_{BC} + d_{BD}) = \tfrac{3+4}{2} = 3.5.$$

The new distance matrix is:

	τ_B	G_E
τ_A	4	5.5
τ_B		3.5

Step 2: merge τ_B and G_E. The minimum is 3.5, so merge $G_F = \{\tau_B, \tau_C, \tau_D\}$. Update the distance to τ_A (sizes: $|\{\tau_B\}| = 1$, $|G_E| = 2$):

$$D(\tau_A, G_F) = \frac{1 \cdot D(\tau_A, \tau_B) + 2 \cdot D(\tau_A, G_E)}{1 + 2} = \frac{4 + 2 \cdot 5.5}{3} = \frac{15}{3} = 5.$$

The final distance matrix is:

	G_F
τ_A	5

The procedure discribed above yields an ultrametric tree with merges at distances 3, 3.5, and 5, as shown in Figure 32.

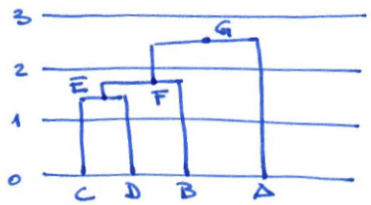

Figure 32: UPGMA tree from our worked-out example.

Problems with UPGMA

UPGMA has several limitations:

- All taxa are placed at the same present time, that is, leaves are aligned at time 0 (Fig. 33).
- It assumes a strict molecular clock (substitutions accumulate at a constant rate along all lineages).
- The fitted ultrametric tree need not be faithful to the input distances.

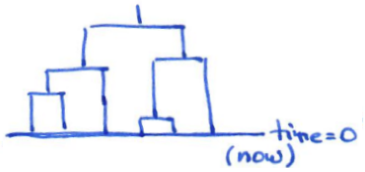

Figure 33: In ultrametric trees, all taxa are placed at the same present time (leaves aligned at time 0).

To see the lack of faithfulness, compare observed and model-implied distances. From the UPGMA tree in Figure 31, the path between τ_A and τ_B has two branches of length 2.5, so the estimated distance is

$$\hat{d}_{AB} = 2.5 + 2.5 = 5,$$

whereas the observed d_{AB} from the data is 4.

A useful diagnostic is the ultrametric condition (Fig. 34). For any triplet $\{\tau_i, \tau_j, \tau_k\}$, the distances are ultrametric iff the two largest values among $\{d_{ij}, d_{ik}, d_{jk}\}$ are equal and the third is less than or equal to them (equivalently, the strong triangle inequality $d_{ij} \leq \max(d_{ik}, d_{jk})$ with the maximum attained at least twice). In our example,

$$d_{AB} = 4, \quad d_{AC} = 5, \quad d_{BC} = 3,$$

the two largest distances (5 and 4) are not equal, so the data are not ultrametric. Hence no UPGMA (ultrametric) tree can be perfectly faithful to these input distances; the algorithm necessarily distorts some values to enforce ultrametricity.

Figure 34: The ultrametric condition, graphically.

Neighbor Joining Algorithm

Neighbor joining (NJ) is a bottom–up, agglomerative method for constructing phylogenetic trees from a distance matrix. Unlike UPGMA, NJ does not assume an ultrametric clock; it seeks a tree whose additive path lengths best match the observed distances.

Let $D = (d_{ij})$ be the current distance matrix among our investigated taxa $\{\tau_1, \ldots, \tau_n\}$. In the NJ algorithm, the matrix Q is a transformed version of the distance matrix D used to decide which pair of taxa to join next. It corrects for the overall divergence of each taxon from all others, reducing the risk of joining pairs that appear close to each other merely because they are both distant from the remaining taxa. Formally,

$$Q(\tau_i, \tau_j) = (n-2)d_{ij} - R_i - R_j, \qquad R_i = \sum_{k=1}^{n} d_{ik}.$$

Introduced by Naruya Saitou and Masatoshi Nei in 1987 (*Mol. Biol. Evol.* 4:406–425). The Q-matrix correction made neighbor joining robust to unequal evolutionary rates. Its algebraic form was so unexpectedly effective that early researchers jokingly called it "black magic."

where d_{ij} is the observed distance between taxa τ_i and τ_j, while R_i and R_j are the total distances of τ_i and τ_j to all other taxa, reflecting their overall separation in the dataset.

A note on the equation above is in order. The factor $(n-2)$ arises from the algebraic derivation of neighbor joining as a method that minimizes the total branch length of the tree. Each distance d_{ij} contributes to connections with the remaining $n-2$ taxa, so the term $(n-2)d_{ij}$ balances this contribution against the overall divergences R_i and R_j. It serves as a normalization ensuring that pair selection remains consistent as the number of taxa decreases.

The Q-matrix values are not distances themselves but *criteria* indicating how suitable each pair is for joining. The pair (τ_i, τ_j) with the smallest $Q(\tau_i, \tau_j)$ is selected for merging, since minimizing Q tends to minimize the total branch length of the inferred additive tree.

Construction of the tree by neighbour joining starts from a star topology. Initially, all taxa are attached to a central node. At each iteration it joins a pair (τ_i, τ_j) into a new internal node v (Fig. 35), computes branch lengths $L_{\tau_i v}$ and $L_{\tau_j v}$, and replaces τ_i, τ_j by v in the matrix. Distances from v to remaining taxa are then updated and the process repeats.

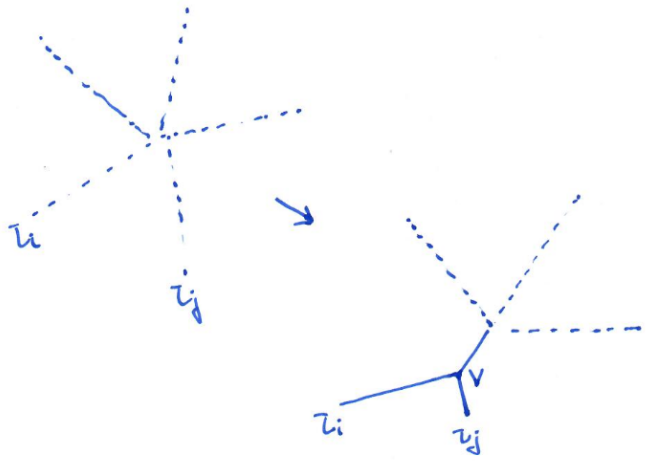

Figure 35: A single step of the NJ algorithm. Just the nodes connected to the center of the star are shown.

To compute the distances between newly introduced node v and all other taxa, we use the three-point formula. Consider the case of three taxa τ_A, τ_B, τ_C attached to a central node Z (Fig. 36) with limb lengths L_x, L_y, L_z, the observed distances satisfy

$$L_x + L_y = d_{AB},$$
$$L_x + L_z = d_{AC},$$
$$L_y + L_z = d_{BC}.$$

We can subtract d_{AB} from d_{AC} to get $L_z - L_y = d_{AC} - d_{AB}$, then add

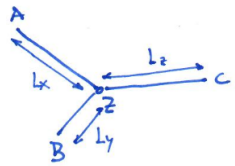

Figure 36: The topology for the derivation of the three-point formula.

$L_y + L_z = d_{BC}$ to obtain

$$L_y = \tfrac{1}{2}\big(d_{AB} + d_{BC} - d_{AC}\big),$$
$$L_x = \tfrac{1}{2}\big(d_{AB} + d_{AC} - d_{BC}\big),$$
$$L_z = \tfrac{1}{2}\big(d_{AC} + d_{BC} - d_{AB}\big).$$

In general, for joining (τ_i, τ_j), distances from the new node v to any other τ_k follow

$$d_{vk} = \tfrac{1}{2}\big(d_{ik} + d_{jk} - d_{ij}\big).$$

Branch lengths for the joined pair are then computed as follows. Using the Q-selection and the row sums R_i, R_j,

$$L_{\tau_i v} = \tfrac{1}{2} d_{ij} + \frac{R_i - R_j}{2(n-2)}, \qquad L_{\tau_j v} = d_{ij} - L_{\tau_i v}.$$

We now have all the ingredients to describe the NJ algorithm:

1. **Initialization.** Assume a star topology over $\{\tau_1, \ldots, \tau_n\}$; set $D^{(0)} = D$.

2. **Select neighbors.** For current $D^{(t)}$ with n_t taxa, compute $R_i = \sum_k d_{ik}$ and the matrix $Q(\tau_i, \tau_j) = (n_t - 2)d_{ij} - R_i - R_j$. Choose the pair (τ_i, τ_j) with the smallest Q (we use Q rather than raw d_{ij} to correct for overall proximity to the rest).

3. **Create a new node and update distances.**

   (a) Introduce a new node v joining τ_i, τ_j with branch lengths

   $$L_{\tau_i v} = \tfrac{1}{2} d_{ij} + \frac{R_i - R_j}{2(n_t - 2)}, \qquad L_{\tau_j v} = d_{ij} - L_{\tau_i v}.$$

   (b) For every remaining τ_k, set

   $$d_{vk} = \tfrac{1}{2}\big(d_{ik} + d_{jk} - d_{ij}\big).$$

   (c) Form the reduced matrix by replacing τ_i, τ_j with v.

4. **Iterate.** If more than two taxa remain, return to Step 2; otherwise connect the last two nodes and assign the final branch length by their remaining distance.

NJ thus iteratively identifies neighbors in the additive sense, computes edge lengths via the three-point formula, and builds a tree consistent with the observed distances without enforcing a molecular clock.

Neighbor Joining: an Example

Data. Consider the distance matrix $D = (d_{ij})$ for taxa a, b, c, d, e, with row sums $R_i = \sum_j d_{ij}$ shown at right:

	a	b	c	d	e	R_i
a	0	17	21	31	23	92
b	17	0	30	34	21	102
c	21	30	0	28	39	118
d	31	34	28	0	43	136
e	23	21	39	43	0	126

Initial topology. We start with $n = 5$ taxa all attached to a central node, as shown in Figure 37.

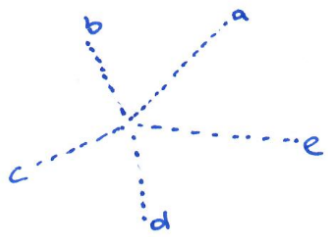

Figure 37: The initial star topology for the NJ algorithm.

Step 1: the first join. With $n = 5$, we compute the neighbour joining matrix Q with $Q(i,j) = (n-2)d_{ij} - R_i - R_j = 3d_{ij} - R_i - R_j$:

	b	c	d	e
a	-143	-147	-135	-149
b		-130	-136	-165
c			$\mathbf{-170}$	-127
d				-133

The smallest entry in this matrixis $Q(c,d) = -170$, so in the first step we join c and d into a new node u. Resulting tree topology is shown in Figure 38.

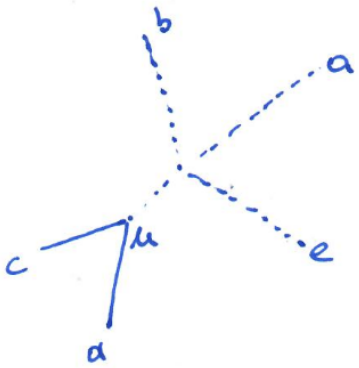

Figure 38: The tree topology after the first step of the NJ algorithm.

We now need to compute the branch lengths for the joined pair $u = (c,d)$ to the nodes c and d. With $d_{cd} = 28$, $R_c = 118$, $R_d = 136$, and $n = 5$, we get:

$$L_{cu} = \tfrac{1}{2}d_{cd} + \frac{R_c - R_d}{2(n-2)} = \tfrac{28}{2} + \frac{-18}{6} = 11,$$
$$L_{du} = d_{cd} - L_{cu} = 17.$$

In order to continue with joining, we have to update the distances from the new node u to the remaining taxa. For each remaining taxon $k \in \{a, b, e\}$, we compute:

$$d_{uk} = \tfrac{1}{2}\left(d_{ck} + d_{dk} - d_{cd}\right),$$

giving $d_{ua} = 12$, $d_{ub} = 18$, and $d_{ue} = 27$. The reduced matrix on $\{a, b, e, u\}$ is as follows, where we have also added the column for the row sums R_i:

	a	b	e	u	R_i
a	0	17	23	12	52
b	17	0	21	18	56
e	23	21	0	27	71
u	12	18	27	0	57

Step 2: the second join. We are now ready to repeat the process. With $n = 4$, we compute the $Q(i,j) = 2d_{ij} - R_i - R_j$. The adjusted distance matrix is:

	b	e	u
a	-74	-77	$\mathbf{-85}$
b		$\mathbf{-85}$	-77
e			-74

The smallest values tie at $Q(a,u) = -85$ and $Q(b,e) = -85$. We resolve the tie arbitrarily by joining (b,e) into a new node v (Fig. 39). The lengths of the new branches are:

$$L_{bv} = \tfrac{1}{2}d_{be} + \frac{R_b - R_e}{2(n-2)} = \frac{21}{2} + \frac{56-71}{4} = 6.75,$$
$$L_{ev} = d_{be} - L_{bv} = 21 - 6.75 = 14.25.$$

We now need to compute the distances to the new node v:

$$d_{av} = \tfrac{1}{2}(d_{ab} + d_{ae} - d_{be}) = \tfrac{1}{2}(17 + 23 - 21) = 9.5,$$
$$d_{uv} = \tfrac{1}{2}(d_{ub} + d_{ue} - d_{be}) = \tfrac{1}{2}(18 + 27 - 21) = 12.$$

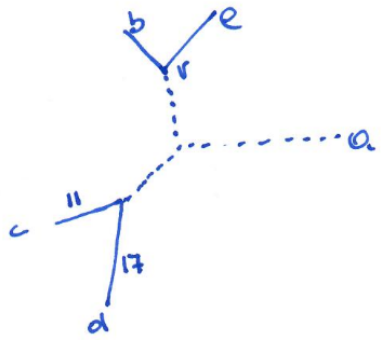

Figure 39: The tree topology after the second step of the NJ algorithm, where we just joined b and e into v.

The resulting distance matrix is:

	a	u	v
a	0	12	9.5
u	12	0	12
v	9.5	12	0

You will notice that we did not add the column for the row sums R_i this time. Why?

Step 3: no more joins, just computations of branch lengths. With three nodes left, we can now compute the branch lengths, that is, distances to the center z by the three-point formula:

$$d_{az} = \tfrac{1}{2}(d_{au} + d_{av} - d_{uv}) = \tfrac{1}{2}(12 + 9.5 - 12) = 4.75,$$
$$d_{uz} = \tfrac{1}{2}(d_{au} + d_{uv} - d_{av}) = \tfrac{1}{2}(12 + 12 - 9.5) = 7.25,$$
$$d_{vz} = \tfrac{1}{2}(d_{av} + d_{uv} - d_{au}) = \tfrac{1}{2}(9.5 + 12 - 12) = 4.75.$$

The final tree is shown in Figure 40. Great. Done. We have constructed a phylogenetic tree from the distance matrix using the neighbour joining algorithm. But is the tree faithful to our initial distance data? Let us check, and use out tree model to estimate the distance between c and e:

The final tree is shown in Figure 40. We have constructed a phylogenetic tree from the distance matrix using the neighbor joining algorithm. Great, done! But is the tree faithful to the initial distances?

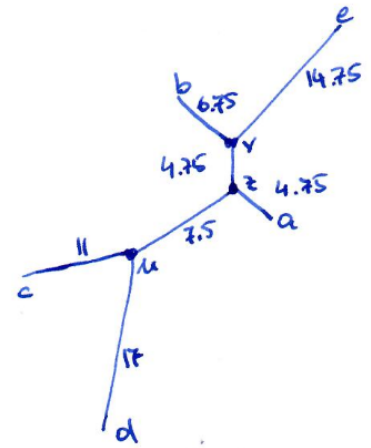

Figure 40: The final tree topology after the NJ algorithm.

As a check, let us estimate the model distance between c and e:

$$\begin{aligned}\hat{d}_{ce} &= L_{cu} + d_{uz} + d_{vz} + L_{ev} \\ &= 11 + 7.25 + 4.75 + 14.25 \\ &= 37.25.\end{aligned}$$

From the data, $d_{ce} = 39$, so the tree underestimates by $39 - 37.25 = 1.75$. This illustrates that NJ yields a close, but not perfectly faithful, additive fit when the input distances are not exactly additive.

Open Questions

The neighbor joining algorithm provides an efficient and widely used approach to reconstruct phylogenetic trees, but several important questions remain regarding the quality and interpretation of the resulting tree.

Faithfulness of the reconstruction. A tree is said to be *faithful* (or *additive*) if the distances measured along the tree exactly reproduce the observed pairwise distances between taxa. For neighbor joining, faithfulness holds when the input distance matrix is itself additive, that is, when it satisfies the four-point condition:

$$d_{ij} + d_{kl} \leq \max(d_{ik} + d_{jl},\, d_{il} + d_{jk}) \quad \text{for all distinct } i, j, k, l.$$

When this condition is met, NJ reconstructs the true additive tree exactly. In the example above, the distances approximately satisfy but do not perfectly meet additivity, so the resulting tree is a close approximation but not strictly faithful. Small deviations from additivity can lead to slight distortions in branch lengths or topology.

Estimating stability: bootstrap. To assess the reliability of inferred clades, one can use the *bootstrap* method. In this approach, the original sequence alignment is resampled with replacement to generate many pseudo-replicate datasets. For each replicate, a tree is reconstructed (using NJ or another method). The frequency with which a given branch or grouping appears across replicates provides a measure of its statistical support, often expressed as a bootstrap percentage. High bootstrap values (typically above 70–80%) indicate well-supported branches; low values suggest uncertainty in the inferred relationships.

Position of the root. The neighbor joining algorithm produces an *unrooted* tree, showing relationships among taxa without specifying the direction of evolution. To assign a root, one typically introduces

an *outgroup*—a taxon known to be distantly related but still part of the broader evolutionary context. The root is then placed on the branch connecting the outgroup to the rest of the taxa, orienting the tree and allowing inference of ancestral–descendant relationships.

Comparison with UPGMA. Unlike UPGMA, neighbor joining does not assume a constant rate of evolution and can therefore accommodate non-ultrametric data. It is generally more faithful to the observed distances when the molecular clock assumption is violated. The fit of any reconstructed tree to the data can be quantified, for instance, by the *least-squares error* between observed and tree-implied distances:

$$E = \sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2.$$

Lower E indicates a tree that better reproduces the observed distances. Comparing E_{UPGMA} and E_{NJ} for the same dataset allows an objective assessment of which method yields a more faithful model of the evolutionary relationships.