CONSTRUCTOR UNIVERSITY

Introduction to Bioinformatics

Phylogenetic Trees – UPGMA

Marc-Thorsten Hütt
mhuett@constructor.university

Felix Jonas
fjonas@constructor.university

Johannes Falk
jfalk@constructor.university

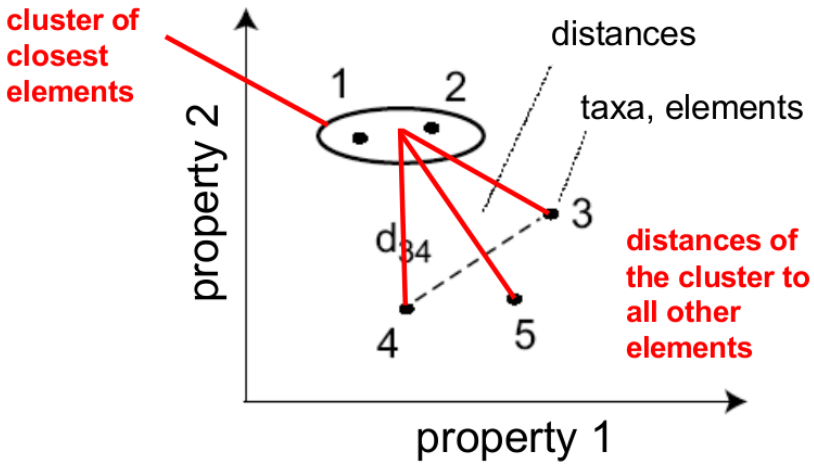
October 23, 2024

Lectures – Phylogenetics

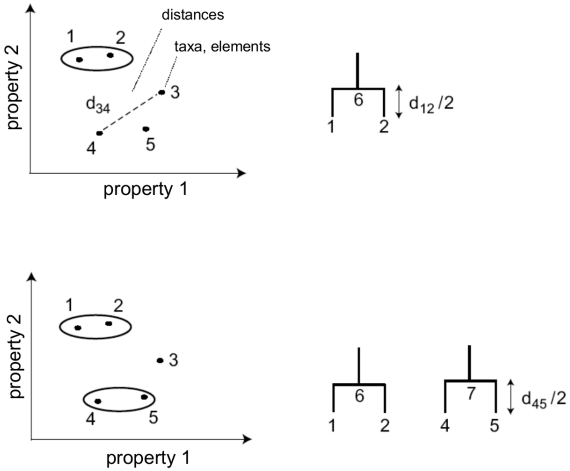
- What is this session about?
  - UPGMA algorithm is finalized.
  - Neighbor-joining algorithm is introduced.
- How can you revise the material after the session?
  - Read Baxevanis/Oullette chapters 14.1, 14.2
  - Read Durbin et al. chapter 7.3 (first half)
  - alternative reading: Hütt/Dehnert chapters 3.3.1 – 3.3.3

UPGMA (Unweighted Pair Group Mean Average)

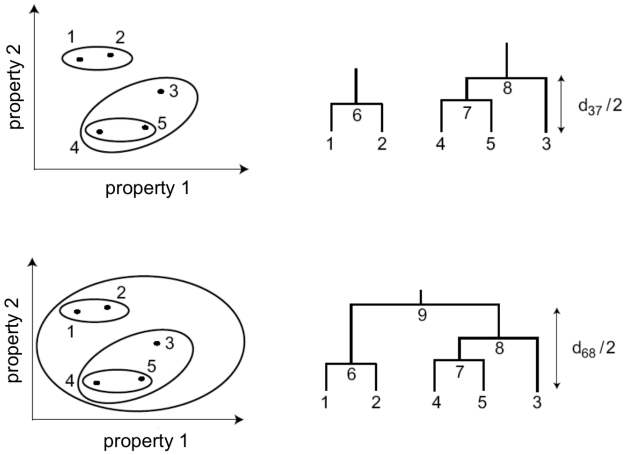
UPGMA – Visualization



UPGMA – Visualization



UPGMA – Visualization



Triangle inequality

Let d_{xy} be the distance (the dissimilarity) between two sequences.

- Metric distances obey the triangle inequality:

$$D_{ik} \leqslant D_{ij} + D_{jk}$$

  In words: The dissimilarity between two sequences can not be larger than the sum of each sequence and a third.

- A uniform molecular clock assumes an ultrametric distance

$$D_{ik} \leqslant \max\{D_{ij}, D_{jk}\}$$

Inequality in ultrametric distances

$$D_{ik} \leqslant \max\{D_{ij}, D_{jk}\}$$

- We assume that we have three distances d_1, d_2, d_3, for which holds: $d_1 \leqslant d_2 \leqslant d_3$.
- The ultrametric distance is only satisfied, if $d_2 \equiv d_3$

UPGMA (Unweighted Pair Group Mean Average)

- Initialization:
  - Assign each sequence i to its own cluster C_i
  - For the tree, define one leaf for each sequence, and place at height zero
- Iteration:
  - Find the two clusters i and j for which d_{ij} is minimal.
  - Merge C_i and C_j to a new cluster C_k
  - Add a node k to the tree with daughter nodes i and j and place it at height $d_{ij}/2$
- Termination
  - When only two clusters i, j remain, place the root at height $d_{ij}/2$
- Result: a unique rooted and ultrametric tree.
- UPGMA can be used if we know that we have an ultrametric tree. (this is usually not the case).

Distance between two clusters

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i; q \in C_j} d_{pq}$$

Simplified distance

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}$$

UPGMA - Example

Example from Wikipedia (UPGMA article). JC69 (Jukes-Cantor) genetic distance matrix $\mathbf{D_1}$ from 5S ribosomal RNA sequence alignment of the bacteria:

- a) Bacillus subtilis
- b) Bacillus stearothermophilus
- c) Lactobacillus viridescens
- d) Acholeplasma modicum
- e) Micrococcus luteus

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

Source: Wikipedia contributors, "UPGMA," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/UPGMA

UPGMA - Example

- Smallest distance in D_1 is 17. Let u denote the new node that connects a and b. We set $\delta(a,u) = \delta(b,u) = D_1(a,b)/2$. That means a and b are equidistant from u.
- Update the distance matrix D_1 into a new distance matrix D_2 where a and b are merged into one row and one column:

$$
\begin{aligned}
D_2((a,b),c) &= (D_1(a,c) \times 1 + D_1(b,c) \times 1)/(1+1) &&= (21+30)/2 = 25.5 \\
D_2((a,b),d) &= (D_1(a,d) + D_1(b,d))/2 &&= (31+34)/2 = 32.5 \\
D_2((a,b),e) &= (D_1(a,e) + D_1(b,e))/2 &&= (23+21)/2 = 22.0
\end{aligned}
$$

Source: Wikipedia contributors, "UPGMA," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/UPGMA

UPGMA - Example

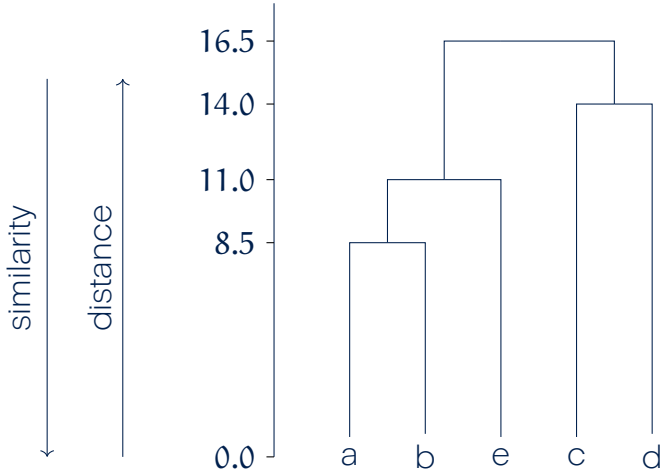
New matrix D_2:

	(a,b)	c	d	e
(a,b)	0	25.5	32.5	22
c	25.5	0	28	39
d	32.5	28	0	43
e	22	39	43	0

- Smallest distance in D_2 is 22. Let v denote the new node that connects (a, b) and e. Because of the ultrametric constraint,
  $\delta(a, v) = \delta(b, v) = \delta(e, v) = 22/2 = 11$.
  For the length between u and v holds:
  $\delta(u, v) = \delta(e, v) - \delta(a, u) = 11 - 8.5 = 2.5$.
- Update the distance matrix D_2 into a new distance matrix D_3 where (a, b) and e are merged into one row and one column
- ...

Source: Wikipedia contributors, "UPGMA," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/UPGMA

UPGMA – Example – Final ultrametric tree



Corresponding distance matrix:

	a	b	c	d	e
a	0	17	33	33	22
b	17	0	33	33	22
c	33	33	0	28	33
d	33	33	28	0	33
e	22	22	33	33	0

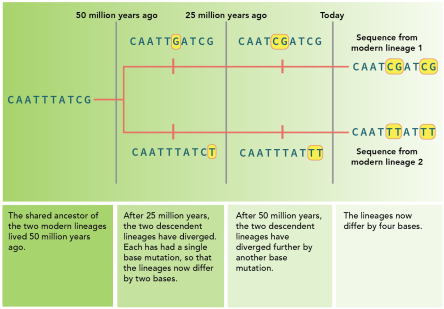
Meets the ultrametric properties: $D_{ik} \leqslant \max\{D_{ij}, D_{jk}\}$

Another example

http://www.slimsuite.unsw.edu.au/teaching/upgma/

Problems of UPGMA

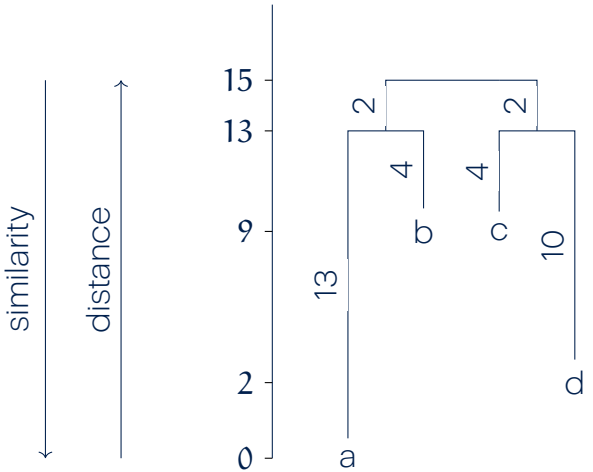
- UPGMA assumes the same evolutionary speed on all lineages: “molecular clock”.



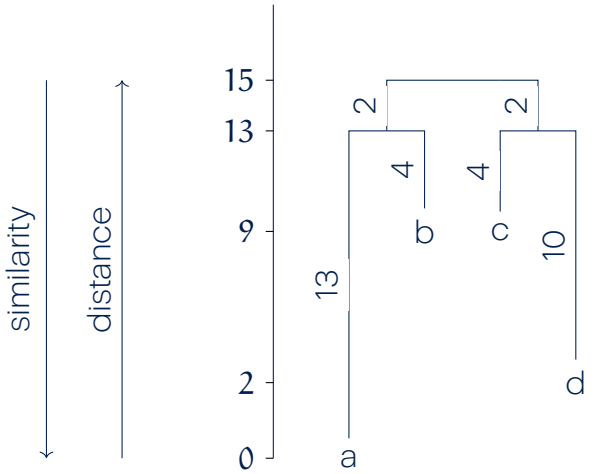
- There is evidence that substitution rates can vary considerably between species
- Turtle mitochondrial DNA has a molecular clock that is slowed down up to 14-fold compared to small mammals.
- Substitution rates depend on: Generation times, population size, intensity of natural selection, ...

Source: (figure) UC Museum of Paleontology Understanding Evolution, https://evolution.berkeley.edu/molecular-clocks/; BY-NC-SA 4.0
(turtle mDNA): J C Avise, et al., Mitochondrial DNA evolution at a turtle’s pace[...], Molecular Biology and Evolution, Volume 9, Issue 3, May 1992

UPGMA from non-ultrametric trees



UPGMA from non-ultrametric trees

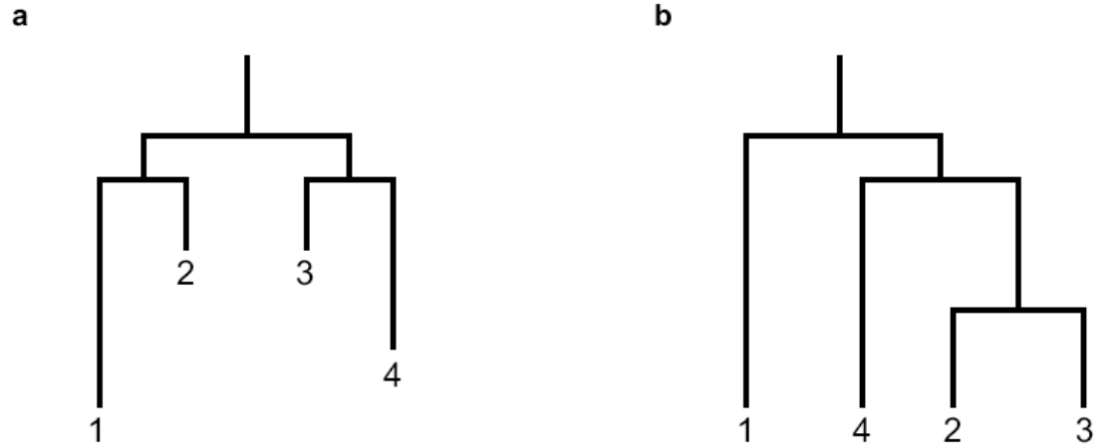


Corresponding distance matrix:

	a	b	c	d
a	0	17	21	27
b	17	0	12	18
c	21	12	0	14
d	27	18	14	0

For non-ultrametric data, UPGMA creates wrong results

Ultrametric vs. Additive Trees



If a distance matrix is to be represented by a tree, it must satisfy the four-point condition:

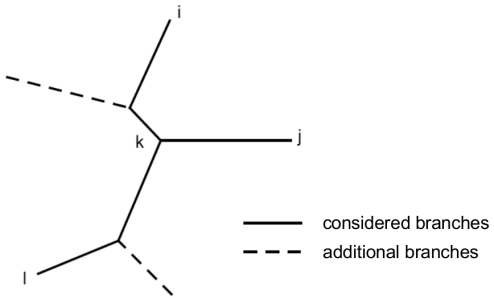
$$D_{ij} + D_{mn} \leqslant \max\{D_{im} + D_{jn}, D_{jm} + D_{in}\}$$

Neighbor-Joining

Neighbor-Joining

- Distance based method (like UPGMA)
- As for UPGMA we need an initial distance matrix $\mathbf{D}$.
- Initialization:
  - Define a tree $\mathbf{T}$ with one leaf for each sequence

Idea behind Neighbor-Joining

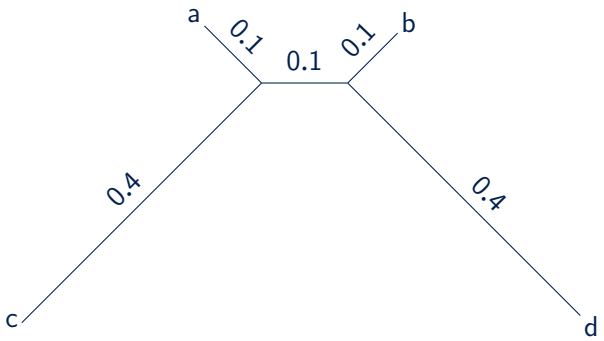


$$d_{ij} = d_{ik} + d_{kj}$$
$$d_{lj} = d_{lk} + d_{kj}$$
$$d_{li} = d_{lk} + d_{ki}$$

$$d_{kl} = \frac{1}{2}(d_{il} + d_{jl} - d_{ij}),$$

Source: Hütt M., Dehnert M., Methoden der Bioinformatik. (Second edition)

Neighbor-Joining – Which nodes to join?



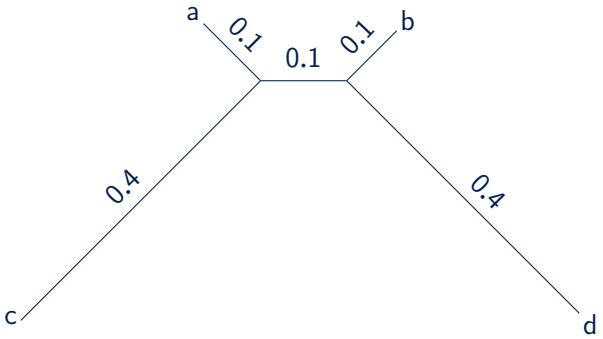
Approach:
Calculate an approx. average distance r_l to all other leaves:

$$r_l = \frac{1}{|\mathcal{L}| - 2} \sum_{i \in \mathcal{L}} d_{il}$$

(note: we do not divide by $|\mathcal{L}|$ but $|\mathcal{L}| - 2$)
Find the two leaves i and j for which $Q_{ij} = d_{ij} - r_i - r_j$ is minimal.

$$\{r_a, r_b, r_c, r_d\} = \{0.7, 0.7, 1.0, 1.0\}$$

Neighbour-Joining – Which nodes to join?



$$Q_{ij} = d_{ij} - r_i - r_j$$

Q-Matrix:

	a	b	c	d
a	0	-1.1	-1.2	-1.1
b	-	0	-1.1	-1.2
c	-	-	0	-1.1
d	-	-	-	0

Select smallest values.

$$\{r_a, r_b, r_c, r_d\} = \{0.7, 0.7, 1.0, 1.0\}$$

Neighbor-Joining

- Iteration:
  - For each leaf l, compute the approx. average distance r_l to all other leaves:

$$r_l = \frac{1}{|\mathcal{L}| - 2} \sum_{i \in \mathcal{L}} d_{il}$$

    (note: we do not divide by $|\mathcal{L}|$ but $|\mathcal{L}| - 2$)
  - Find the two leaves i and j for which $Q_{ij} = d_{ij} - r_i - r_j$ is minimal.
  - Insert a new node k to the tree. The branch lengths from k to i and j are given by:

$$d_{ik} = \tfrac{1}{2}(d_{ij} + r_i - r_j), \quad d_{jk} = d_{ij} - d_{ik}$$

  - Add k to the distance matrix and remove i and j. Update the distance matrix via:

$$d_{kl} = \tfrac{1}{2}(d_{il} + d_{jl} - d_{ij}),$$

    where l runs over all remaining entries.

Neighbor-Joining

- Termination
  - When $|\mathcal{L}| = 2$: add an edge between the last two elements of $\mathcal{L}$.
- Result: an unrooted tree
- The tree can be rooted e.g. via an outgroup

Neighbor-Joining – Example – See also UPGMA

Example from Wikipedia (UPGMA article). JC69 (Jukes-Cantor) genetic distance matrix $\mathbf{D}_1$ from 5S ribosomal RNA sequence alignment of the bacteria:

- a) Bacillus subtilis
- b) Bacillus stearothermophilus
- c) Lactobacillus viridescens
- d) Acholeplasma modicum
- e) Micrococcus luteus

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

Source: Wikipedia contributors, "UPGMA," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/UPGMA

Neighbor-Joining – Example

- Compute the average distances:

$$\{r_a, r_b, r_c, r_d, r_e\} = \{30.67, 34., 39.33, 45.33, 42.\}$$

$$\text{E.g. } r_a = \frac{1}{5-2}(d_{ab} + d_{ac} + d_{ad} + d_{ae})$$

- Compute the corrected distances:

	a	b	c	d	e
a	0	-47.67	-49.	-45.	-49.67
b	-47.67	0	-43.33	-45.33	-55.
c	-49.	-43.33	0	-56.67	-42.33
d	-45.	-45.33	-56.67	0	-44.33
e	-49.67	-55.	-42.33	-44.33	0

E.g.

$$Q_{ab} = 17 - 30.67 - 34 = -47.67$$

Neighbor-Joining – Example

	a	b	c	d	e
a	0	-47.67	-49.	-45.	-49.67
b	-47.67	0	-43.33	-45.33	-55.
c	-49.	-43.33	0	-56.67	-42.33
d	-45.	-45.33	-56.67	0	-44.33
e	-49.67	-55.	-42.33	-44.33	0

Minimal distance is between $\mathbf{c}$ and $\mathbf{d}$. We add a new node $\mathbf{k_1}$ that connects $\mathbf{c}$ and $\mathbf{d}$. We calculate:

$$\mathbf{d_{ck_1}} = 0.5 \cdot (\mathbf{d_{cd}} + \mathbf{r_c} - \mathbf{r_d}) = 0.5 \cdot (28 + 39.33 - 45.33) = 11$$
$$\mathbf{d_{dk_1}} = 17$$

Neighbor-Joining – Example

- We have now a new node k_1 that connects c and d in our tree.
- We now need to update the distance matrix:

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

	a	b	e	k1
a	0	17	23	12
b	17	0	21	18
e	23	21	0	27
k1	12	18	27	0

E.g. $d_{k_1 a} = 0.5 \cdot (d_{ca} + d_{da} - d_{cd}) = 0.5(21 + 31 - 28) = 12$

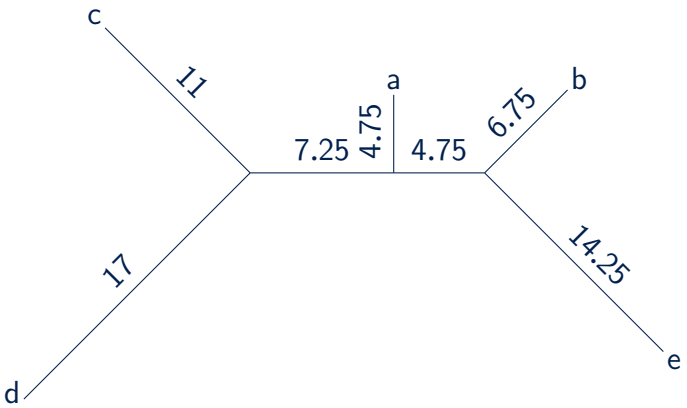
Neighbor-Joining – Example

- New corrected distances:

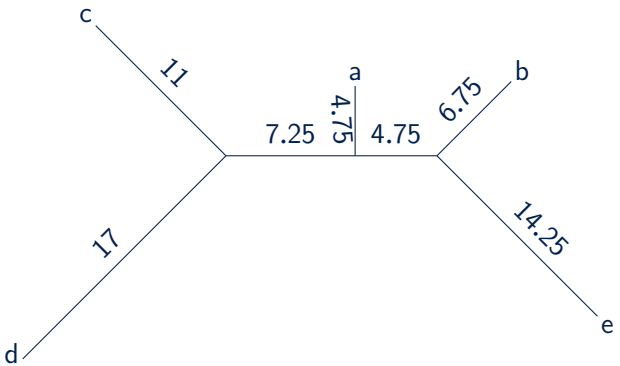
	a	b	e	k1
a	-52.	-37.	-38.5	-42.5
b	-37.	-56.	-42.5	-38.5
e	-38.5	-42.5	-71.	-37.
k1	-42.5	-38.5	-37.	-57.

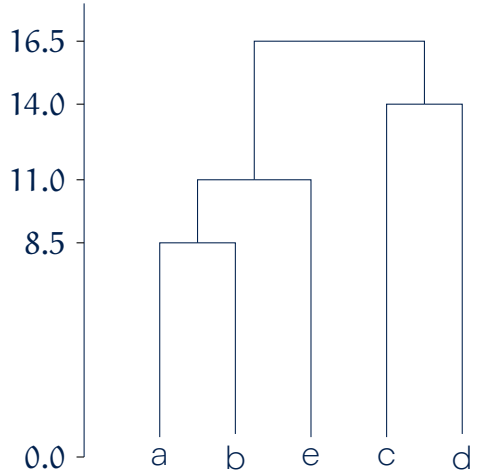
- The smallest distances are between $\mathbf{b}, \mathbf{e}$ or $\mathbf{a}, \mathbf{k_1}$.
- Select two (e.g. $\mathbf{a}$ and $\mathbf{k_1}$) and proceed ...

Neighbor-Joining – Example – Final Tree

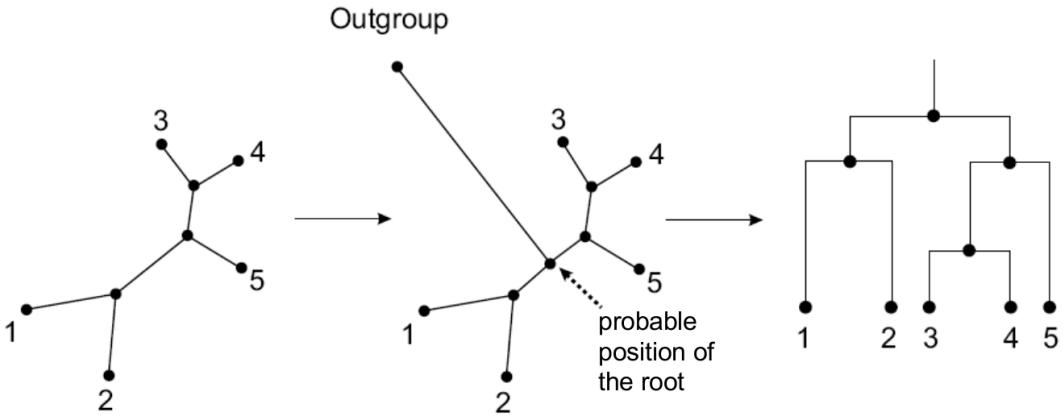


Neighbor-Joining – Comparison with UPGMA





Rooting



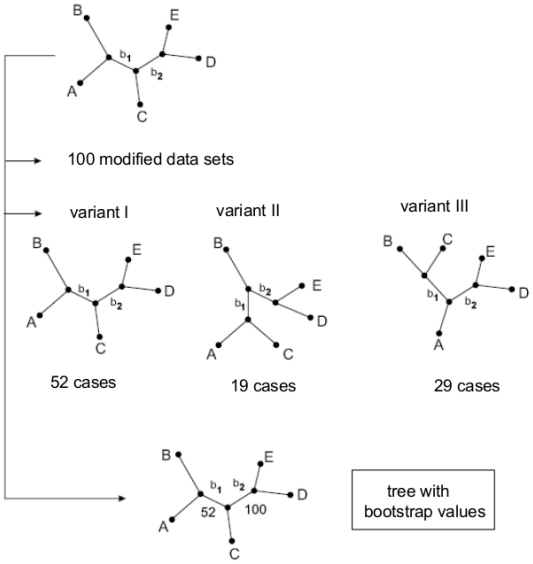
radial tree from neighbor-joining

inclusion of an outgroup

corresponding rooted tree

Source: Hütt M., Dehnert M., Methoden der Bioinformatik. (Second edition)

Bootstrap



Source: Hütt M., Dehnert M., Methoden der Bioinformatik. (Second edition)