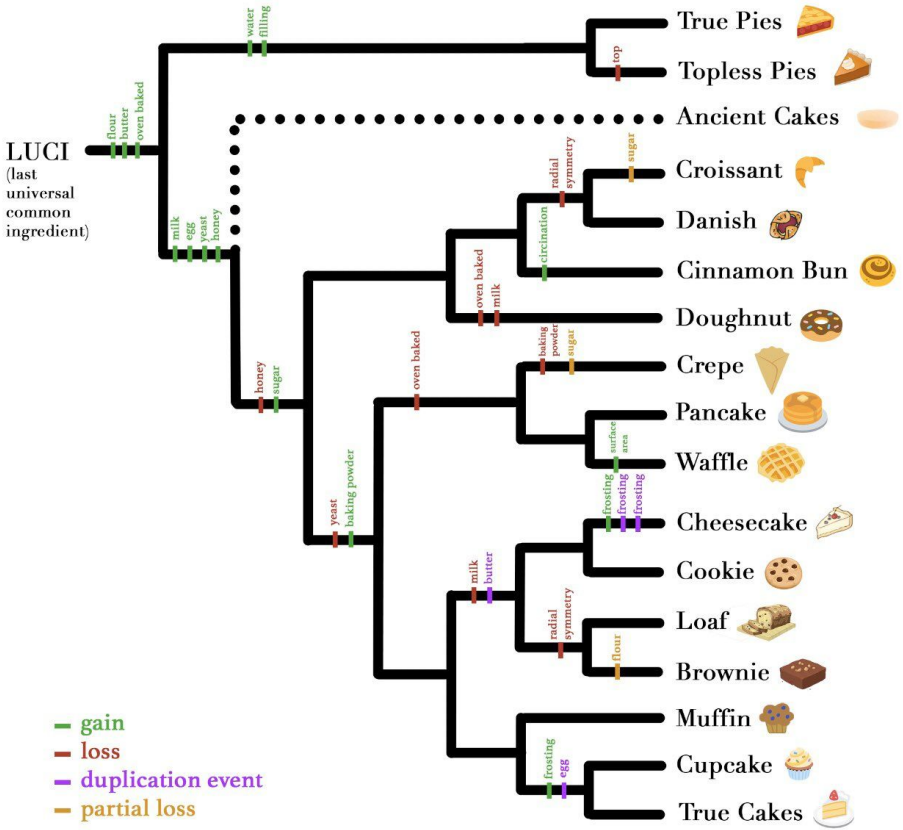
Week 4: Comparative genomics

- Whole genome alignment
  - MUMmer & Suffix trees
- Gene/species trees
  - Phylogenetic trees
  - Gene orthology & functional analysis

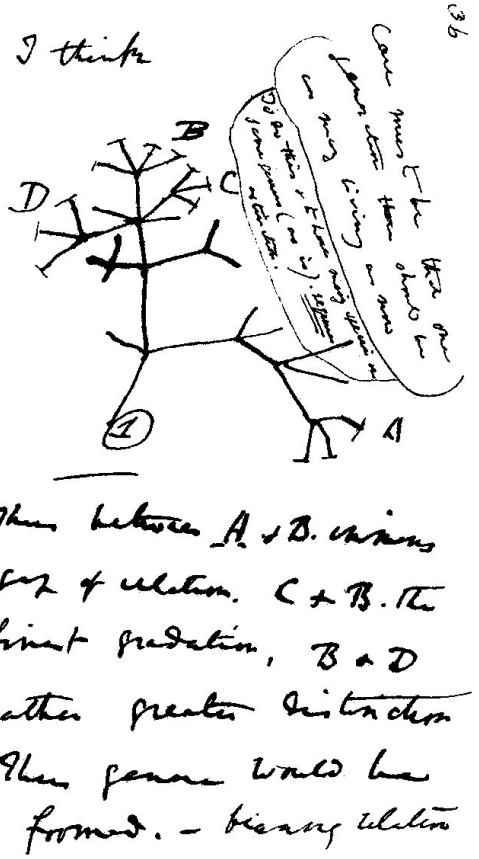
Phylogeny & Phylogenetic tree



On the Origin of Baked Goods
by Means of Natural Consumption

Phylogeny & Phylogenetic tree

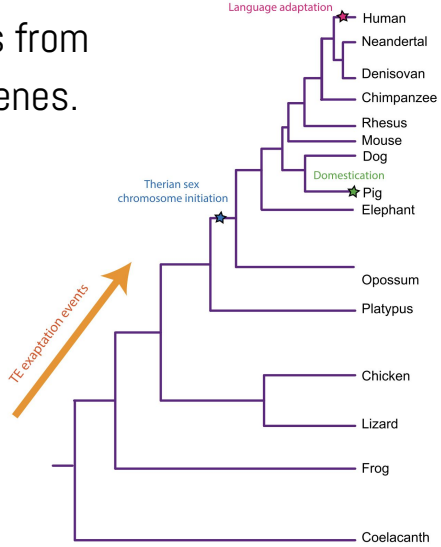
- Useful for:
  a. organizing knowledge of biological diversity,
  b. structuring classifications, and
  c. providing insight into events that occurred during evolution.
- Diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.
- Trees show descent from a common ancestor.



Species tree vs gene tree

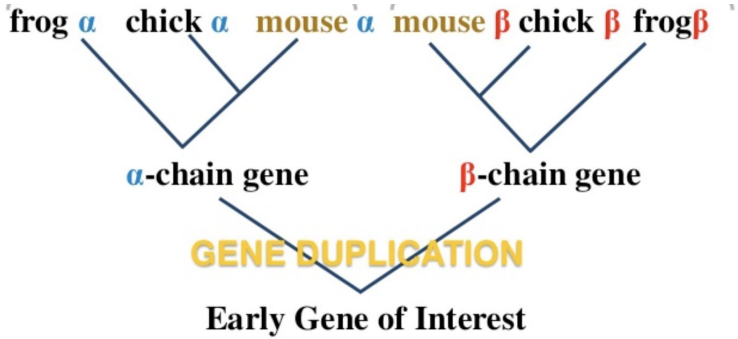
Species trees recover the genealogy of taxa, individuals of a population, etc.

- Internal nodes represent speciation or other taxonomic events.
- Species trees should Contain sequences from only orthologous genes.



Gene trees represent the evolutionary history of the genes included in the study.

- Gene trees can provide evidence for gene duplication events, as well as speciation events.
- Sequences from different homologs can be included in a gene tree; the subsequent analyses should cluster orthologs



NCBI, NAWBIS; Alföldi & Lindblad-Toh (2013) Genome Res.

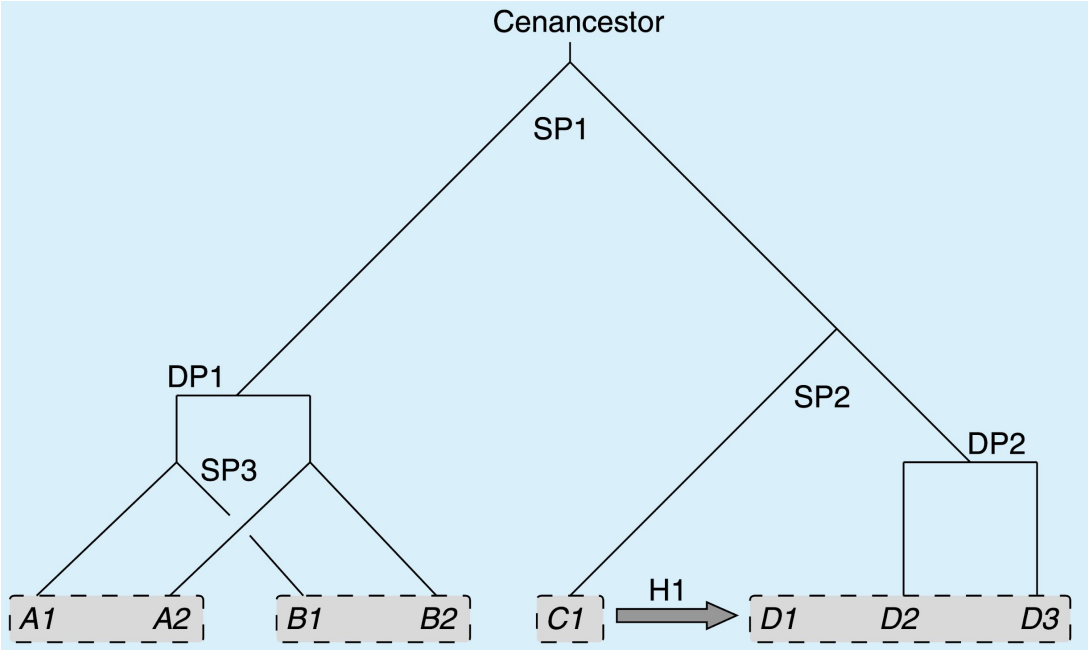
Evolutionary relationships between genes in different species

Evolutionary relationships:

- Orthologs
- Paralogs
  - Subfunctionalization
  - Neofunctionalization

Complicated evolutionary processes:

- gene fusion and fission
- horizontal gene transfer
- whole gene deletion



Webber, Ponting (2004) Curr. Biol.

Approaches for constructing phylogenetic trees

Distance-based methods

- UPGMA & Neighbor-Joining
- Calculate pairwise distances & then build tree

Character-based methods

- Maximum parsimony & Maximum likelihood
- Directly build tree by coupling tree proposal & scoring

Distance-based methods for constructing phylogenetic trees

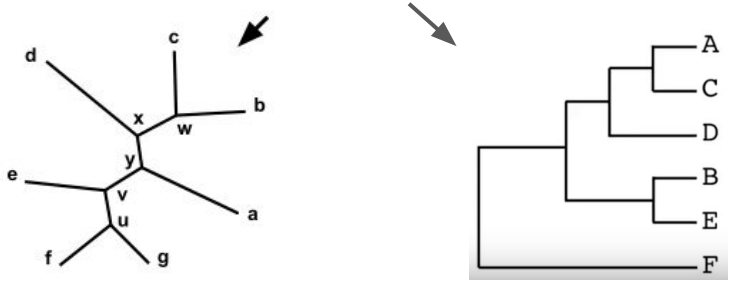
Distance-based methods

- UPGMA & Neighbor-Joining
- Calculate pairwise distances & then build tree

Multiple sequence alignment

```
B9SI54|B9SI54_RICCO_263_570     RILTNVYMYGDGIDRTIISGSKHTM-DGLPAYRTATVAVLGDGFVCKSMTIQNSATSD-K
Q01I60|Q01I60_ORYSA_160_476     YEKTNILLVGDGIGATVITASRSVGIDGIGTYETATVAVIGDGFRAKDITFENGAGAGAH
C5Y8S2|C5Y8S2_SORBI_153_466     YEKTNILLMGEGMGATVITASRSVGIDGLGTHETATVAVIGDGFRARDITFENSAGARAH
B4FRR6|B4FRR6_MAIZE_154_469     YEKANILLMGEGMGATVITASRSVGIDGLGTYETATVDVIGDGFRARDITFENSAGAGAH
D7U4G4|D7U4G4_VITVI_82_394      LEKKNVVFLGDGMGKTVITGSLNVGQPGISTYNSATVGVAGDGFMASGLTMENTAGPDEH
D7M270|D7M270_ARALY_263_574     FEKKNVVFIGDGMGKTVITGSLNAGMPGITTYNTATVGVVGDGFMAHDLTFQNTAGPDAH
Q8L7Q7|PME64_ARATH_283_601      FEKKNVVFIGDGMGKTVITGSLNVGQPGMTTFESATVGVLGDGFMARDLTIENTAGADAH
D8QSM2|D8QSM2_SELML_242_541     DSKSMIMLVGAGARKTIISGNNYVR-EGVTTMDTATVLVAGDGFVARDLTIRNTAGPELH
A9TZ89|A9TZ89_PHYPA_262_575     KQKTNLMFLGDGTDKTIITGSLSDSQPGMITWATATVAVSGSGFIARGITFQNTAGPAGR
D8SH72|D8SH72_SELML_209_529     LQKSMLMFVGDGMDKTIIRGSMSVSKGGTTTFASATLAVNGKGFLARDLTVENTAGPEGH
                                     : : * *   *:* ..       *  :  :**: * *.** .  :*..* *    :
```

	A	B	C	D	E	F
A		7	2	5	6	9
B			6	6	1	8
C				5	7	8
D					5	7
E						8
F						



Distance-based methods for constructing phylogenetic trees

UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

- Rooted tree
- Assumes constant-rate

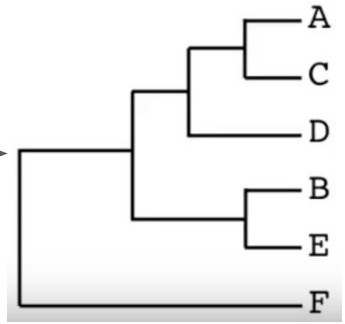
Distance b/w any two clusters A and B, each of size = the mean distance between elements of each cluster

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

$$d_{(\mathcal{A} \cup \mathcal{B}), X} = \frac{|\mathcal{A}| \cdot d_{\mathcal{A},X} + |\mathcal{B}| \cdot d_{\mathcal{B},X}}{|\mathcal{A}| + |\mathcal{B}|}$$

A	ATCGTGGTACTG
B	CCGGAGAACTAG
C	AACGTGCTACTG
D	ATGGTGAAAGTG
E	CCGGAAAACTTG
F	TGGCCCTGTATC

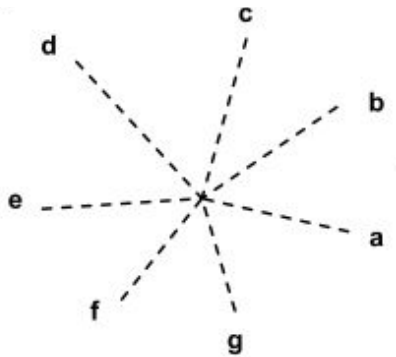
	A	B	C	D	E	F
A		9	2	4	9	10
B			9	6	2	10
C				5	9	10
D					6	10
E						10
F						



Distance-based methods for constructing phylogenetic trees

Neighbor-Joining

- Unrooted tree
- Does not assumes constant-rate



Choose x, y to merge that minimize:

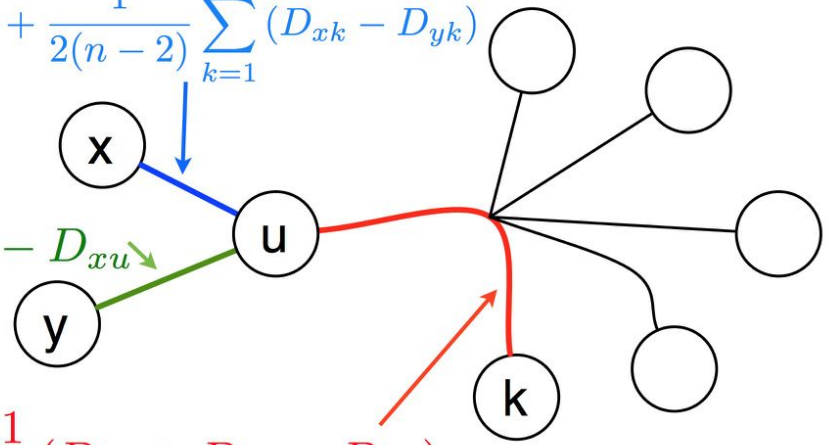
$$Q(x,y) := (n-2)D_{xy} - \left(\sum_{k=1}^{n} D_{xk} + \sum_{k=1}^{n} D_{yk}\right)$$

Update lengths:

$$D_{xu} := \frac{1}{2}D_{xy} + \frac{1}{2(n-2)}\sum_{k=1}^{n}(D_{xk} - D_{yk})$$

$$D_{yu} := D_{xy} - D_{xu}$$

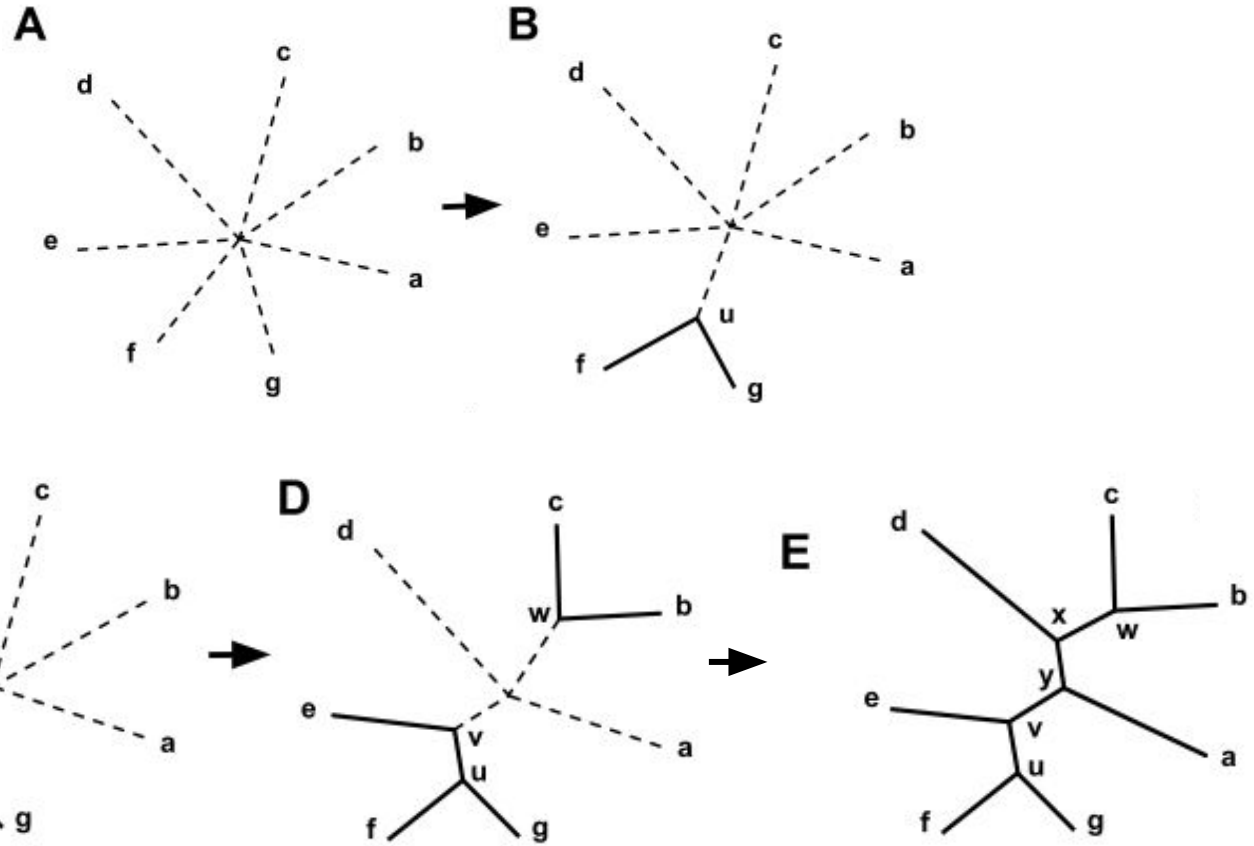
$$D_{uk} := \frac{1}{2}(D_{xk} + D_{yk} - D_{xy})$$



Distance-based methods for constructing phylogenetic trees

Neighbor-Joining

- Unrooted tree
- Does not assumes constant-rate



Distance-based methods for constructing phylogenetic trees

Multiple sequence alignment

```
B9SI54|B9SI54_RICCO_263_570    RILTNVYMYGDGIDRTIISGSKHTM-DGLPAYRTATVAVLGDGFVCKSMTIQNSATSD-K
Q01I60|Q01I60_ORYSA_160_476    YEKTNILLVGDGIGATVITASRSVGIDGIGTYETATVAVIGDGFRAKDITFENGAGAGAH
C5Y8S2|C5Y8S2_SORBI_153_466    YEKTNILLMGEGMGATVITASRSVGIDGLGTHETATVAVIGDGFRARDITFENSAGARAH
B4FRR6|B4FRR6_MAIZE_154_469    YEKANILLMGEGMGATVITASRSVGIDGLGTYETATVDVIGDGFRARDITFENSAGAGAH
D7U4G4|D7U4G4_VITVI_82_394     LEKKNVVFLGDGMGKTVITGSLNVGQPGISTYNSATVGVAGDGFMASGLTMENTAGPDEH
D7M270|D7M270_ARALY_263_574    FEKKNVVFIGDGMGKTVITGSLNAGMPGITTYNTATVGVVGDGFMAHDLTFQNTAGPDAH
Q8L7Q7|PME64_ARATH_283_601     FEKKNVVFIGDGMGKTVITGSLNVGQPGMTTFESATVGVLGDGFMARDLTIENTAGADAH
D8QSM2|D8QSM2_SELML_242_541    DSKSMIMLVGAGARKTIISGNNYVR-EGVTTMDTATVLVAGDGFVARDLTIRNTAGPELH
A9TZ89|A9TZ89_PHYPA_262_575    KQKTNLMFLGDGTDKTIITGSLSDSQPGMITWATATVAVSGSGFIARGITFQNTAGPAGR
D8SH72|D8SH72_SELML_209_529    LQKSMLMFVGDGMDKTIIRGSMSVSKGGTTTFASATLAVNGKGFLARDLTVENTAGPEGH
                                : : * *   *:* ..      *  :  :**: * *.** .  :*..* *    :
```

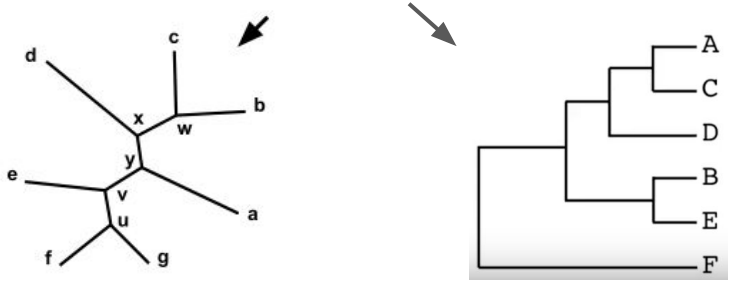
High computational efficiency (esp. NJ).

- Useful for analysing large data sets with low levels of sequence divergence.

Can perform poorly for very divergent sequences.

Need a realistic substitution model to calculate the pairwise distances. Also sensitive to gaps in the sequence alignment.

	A	B	C	D	E	F
A		7	2	5	6	9
B			6	6	1	8
C				5	7	8
D					5	7
E						8
F						



Distance-based methods for constructing phylogenetic trees

High computational efficiency (esp. NJ).

- Useful for analysing large data sets with low levels of sequence divergence.

Can perform poorly for very divergent sequences.

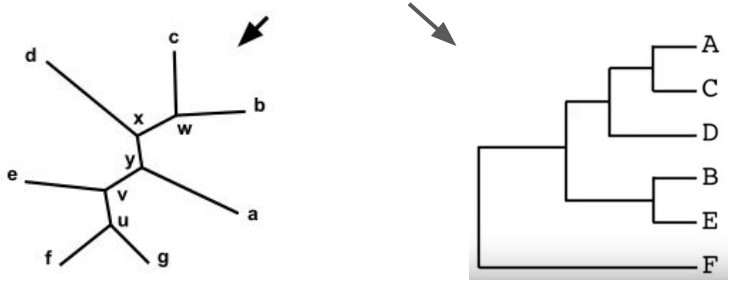
- Large distances involve large sampling errors, and most distance methods (such as NJ) do not account for the high variances of large distance estimates.

Need a realistic substitution model to calculate the pairwise distances. Also sensitive to gaps in the sequence alignment.

Multiple sequence alignment

```
B9SI54|B9SI54_RICCO_263_570     RILTNVYMYGDGIDRTIISGSKHTM-DGLPAYRTATVAVLGDGFVCKSMTIQNSATSD-K
Q01I60|Q01I60_ORYSA_160_476     YEKTNILLVGDGIGATVITASRSVGIDGIGTYETATVAVIGDGFRAKDITFENGAGAGAH
C5Y8S2|C5Y8S2_SORBI_153_466     YEKTNILLMGEGMGATVITASRSVGIDGLGTHETATVAVIGDGFRARDITFENSAGARAH
B4FRR6|B4FRR6_MAIZE_154_469     YEKANILLMGEGMGATVITASRSVGIDGLGTYETATVDVIGDGFRARDITFENSAGAGAH
D7U4G4|D7U4G4_VITVI_82_394      LEKKNVVFLGDGMGKTVITGSLNVGQPGISTYNSATVGVAGDGFMASGLTMENTAGPDEH
D7M270|D7M270_ARALY_263_574     FEKKNVVFIGDGMGKTVITGSLNAGMPGITTYNTATVGVVGDGFMAHDLTFQNTAGPDAH
Q8L7Q7|PME64_ARATH_283_601      FEKKNVVFIGDGMGKTVITGSLNVGQPGMTTFESATVGVLGDGFMARDLTIENTAGADAH
D8QSM2|D8QSM2_SELML_242_541     DSKSMIMLVGAGARKTIISGNNYVR-EGVTTMDTATVLVAGDGFVARDLTIRNTAGPELH
A9TZ89|A9TZ89_PHYPA_262_575     KQKTNLMFLGDGTDKTIITGSLSDSQPGMITWATATVAVSGSGFIARGITFQNTAGPAGR
D8SH72|D8SH72_SELML_209_529     LQKSMLMFVGDGMDKTIIRGSMSVSKGGTTTFASATLAVNGKGFLARDLTVENTAGPEGH
                                : : * *  *:* ..      *  :  :**: * *.** .  :*..* *    :
```

	A	B	C	D	E	F
A		7	2	5	6	9
B			6	6	1	8
C				5	7	8
D					5	7
E						8
F						



Approaches for constructing phylogenetic trees

Distance-based methods

- UPGMA & Neighbor-Joining
- Calculate pairwise distances & then build tree

Character-based methods

- Maximum parsimony & Maximum likelihood
- Directly build tree by coupling tree proposal & scoring

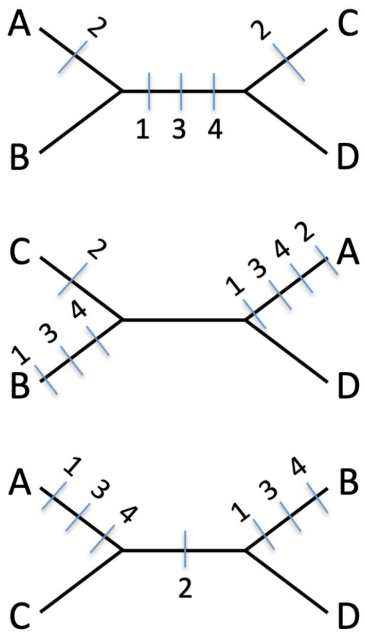
Maximum parsimony

MP minimizes the number of changes on a phylogenetic tree by assigning character states to interior nodes on the tree.

The character (or site) length is the minimum number of changes required for that site, whereas the tree score is the sum of character lengths over all sites.

The maximum parsimony tree is the tree that minimizes the tree score.

```
   1234
A: ACGTACGT
B: CCGCAAGT
C: GCGCCCAC
D: TCCCCAAC
```



Maximum likelihood

Maximum Likelihood is a:

- general statistical method
- for **estimating unknown parameters** of a probabilistic model
- by maximizing a function, so that
- **under the assumed model,**
- **the observed data is most probable**.

Likelihood of hypothesis =
Probability of data given hypothesis

- Fair or unfair coin?

$P_{head} = 0.5$	Fair
$P_{head} = 0.67$	Unfair

- Flip coin 4 times, get:

3 heads, 1 tail

	Fair	Unfair
H x H x H x T	1/2 x 1/2 x 1/2 x 1/2 = 1/16	2/3 x 2/3 x 2/3 x 1/3 = 8/81
H x H x T x H	1/2 x 1/2 x 1/2 x 1/2 = 1/16	2/3 x 2/3 x 1/3 x 2/3 = 8/81
H x T x H x H	1/2 x 1/2 x 1/2 x 1/2 = 1/16	2/3 x 1/3 x 2/3 x 2/3 = 8/81
T x H x H x H	1/2 x 1/2 x 1/2 x 1/2 = 1/16	1/3 x 2/3 x 2/3 x 2/3 = 8/81
Total	1/4 (0.25)	32/81 (0.40)

https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo15.html

Maximum likelihood

1. Given data, assume it comes from a model (e.g., normal/binomial distribution).
2. Likelihood ~ the probability of observing the data given the model: **P(Data | Model)**.
3. Examine this likelihood function to see where it is greatest (meaning, different values of the parameters of the model: e.g. μ & σ).
4. The values of the parameters at that point is the **maximum likelihood estimate** of the parameters (found numerically by some iterative optimization procedure).

MLEs have desirable asymptotic properties:

- Unbiased (expected value = true value of the parameter),
- Consistent (approach true values), &
- Efficient (have the smallest variance among unbiased estimates).

- Flip coin 4 times, get:

3 heads, 1 tail

	Fair	Unfair
H x H x H x T	1/2 x 1/2 x 1/2 x 1/2 = 1/16	2/3 x 2/3 x 2/3 x 1/3 = 8/81
H x H x T x H	1/2 x 1/2 x 1/2 x 1/2 = 1/16	2/3 x 2/3 x 1/3 x 2/3 = 8/81
H x T x H x H	1/2 x 1/2 x 1/2 x 1/2 = 1/16	2/3 x 1/3 x 2/3 x 2/3 = 8/81
T x H x H x H	1/2 x 1/2 x 1/2 x 1/2 = 1/16	1/3 x 2/3 x 2/3 x 2/3 = 8/81
Total	1/4 (0.25)	32/81 (0.40)

https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo15.html

Maximum likelihood for tree estimation

Model: The tree; **Parameters**: The tree's branch lengths.

Use a specific substitution model:

- Assume independent evolution of sites in the sequence → likelihood = product of the probabilities for different sites.
- Probability at any particular site = average over the unobserved character states at the ancestral nodes.

ML for tree inference is equivalent to comparing many statistical models, each with the same number of parameters.

Two optimization steps:

1. Optimization of branch lengths to calculate the tree score for each candidate tree.
2. A search in the tree space for the maximum likelihood tree.

Yang & Rannala (2012) Nat. Rev. Genet.

Maximum likelihood for tree estimation

ML is used exclusively these days for inferring deep phylogenies using conserved proteins.

- All model assumptions are explicit, so that they can be evaluated and improved.
- Availability of a rich repertoire of sophisticated evolutionary models.
  - Including models that accommodate variable amino acid substitution rates among sites or different amino acid frequencies among sites.
- Great for understanding the process of sequence evolution.
  - The likelihood ratio test can be used to:
    - Examine the fit of evolutionary models
    - Test interesting biological hypotheses (e.g. molecular clock) and selection affecting protein evolution.

Yang & Rannala (2012) Nat. Rev. Genet.

Maximum likelihood for tree estimation

There are some drawbacks!

- The attractive asymptotic properties of MLEs apply to parameter estimation when the true tree is given but not to the maximum likelihood tree.
- The likelihood calculation, particularly tree search under the likelihood criterion, is **computationally demanding**.
- The method has potentially **poor statistical properties if the model is misspecified**.

Yang & Rannala (2012) Nat. Rev. Genet.

Approaches for constructing phylogenetic trees

Distance-based methods

- UPGMA & Neighbor-Joining
- Calculate pairwise distances & then build tree

Character-based methods

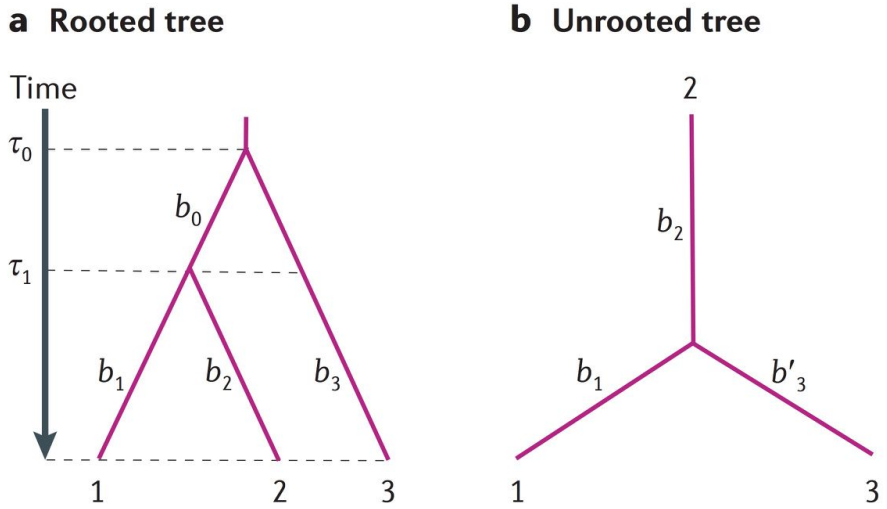
- Maximum parsimony & Maximum likelihood
- Directly build tree by coupling tree proposal & scoring

Rooted vs. Unrooted trees

Substitution rate is constant over time or among lineages ⟶ the molecular clock holds.

The tree will then have a root (inferring rooted tree is called molecular clock rooting).

- The tree will be ultrametric:
  distances from the tips of the tree to the root are all equal (b0 + b1 = b0 + b2 = b3).



A rooted tree for *s* species:

- Can then be represented by the ages of the $s - 1$ ancestral nodes.
- Involves $s - 1$ branch-length parameters.

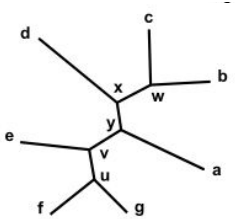
For distantly related species, the clock hypothesis should not be assumed.

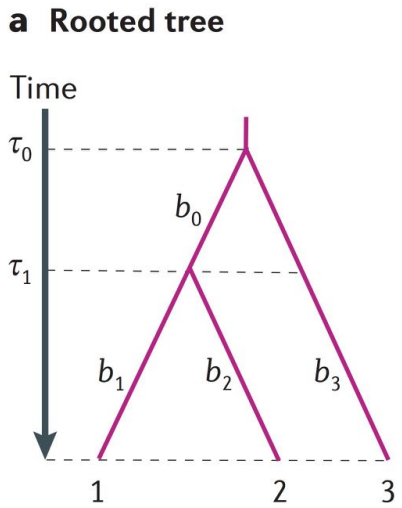
Yang & Rannala (2012) Nat. Rev. Genet.

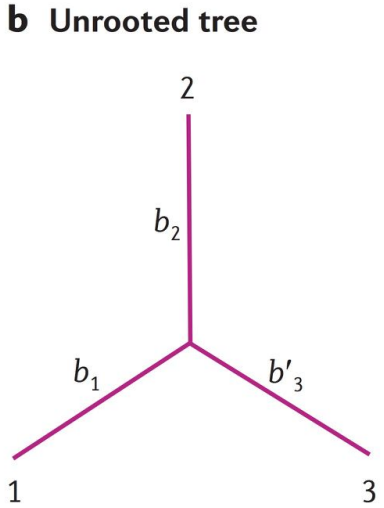
Rooted vs. Unrooted trees

If every branch on the tree is allowed to have an independent evolutionary rate ⟶ unrooted trees.

An unrooted tree for s species has $2s - 3$ branch length parameters.





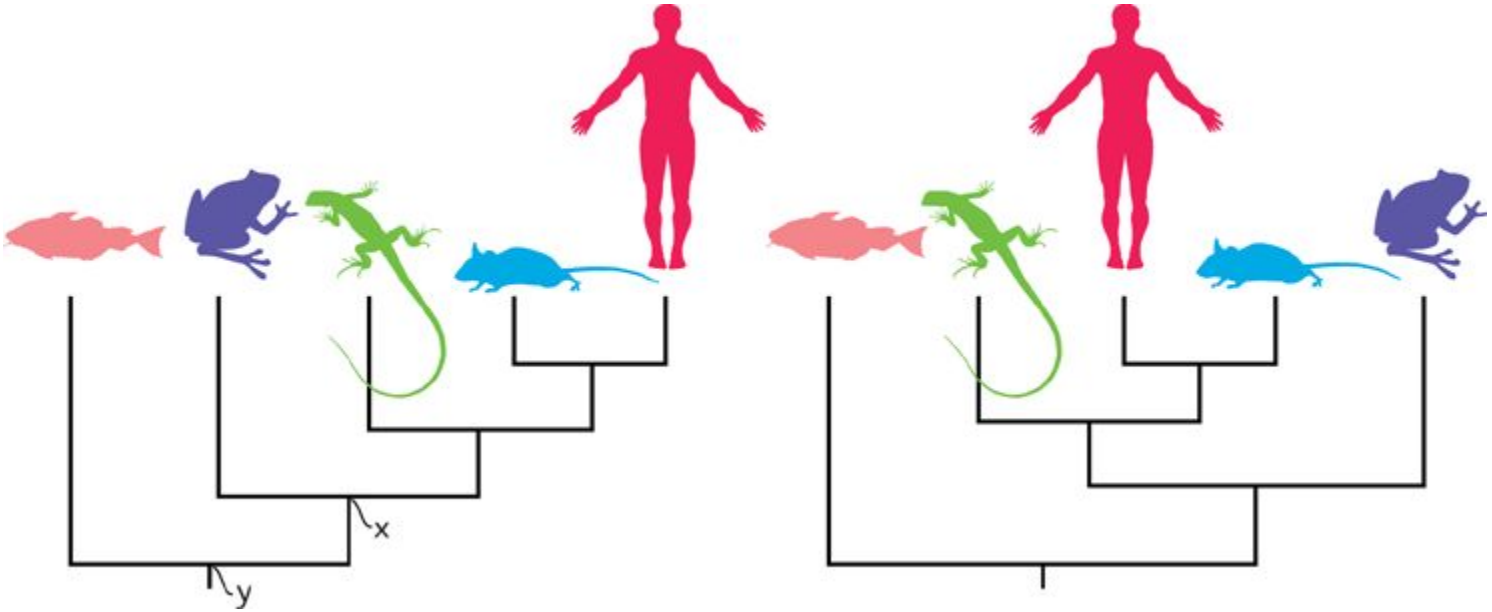


Rooting a tree using outgroup rooting:

- Include outgroup species (a species/genes known to be more distantly related than the species/genes of interest).
- Root is located along the branch that leads to the outgroup so that the tree for the ingroup species is rooted.

Yang & Rannala (2012) Nat. Rev. Genet.

Interpreting a tree



https://science.sciencemag.org/content/310/5750/979