

Phylogenetic reconstruction

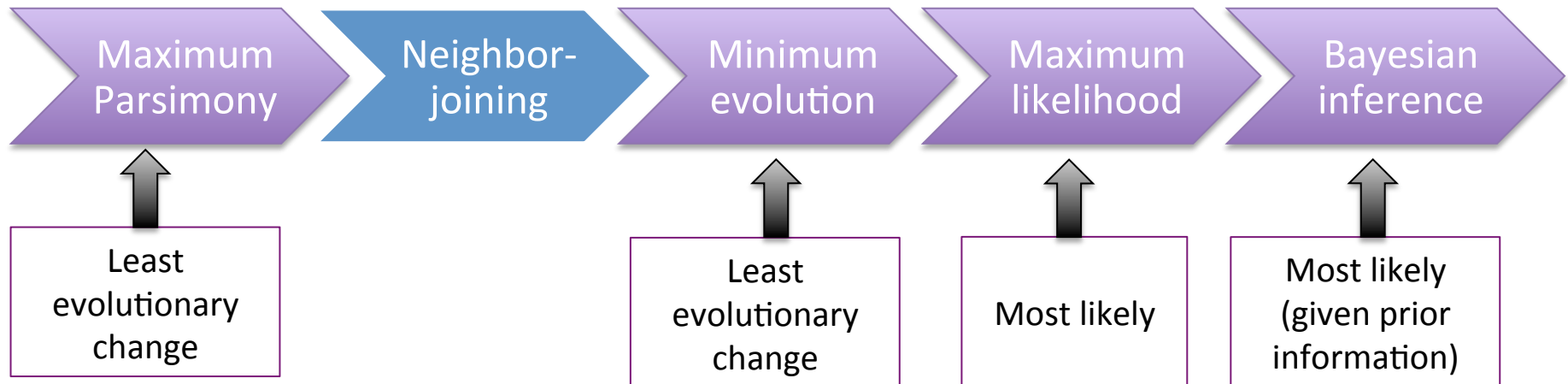
Produce a phylogenetic tree -

Describing likely descent from a common ancestral sequence of a set of aligned contemporary sequence.

How many rooted and unrooted possibilities are there?

| Number of OTUs | # rooted trees | # unrooted trees |
|----------------|----------------|------------------|
| 2 | 1 | 1 |
| 3 | 3 | 1 |
| 4 | 15 | 3 |
| 5 | 105 | 15 |
| 6 | 954 | 105 |
| 7 | 10,395 | 954 |
| 8 | 135,135 | 10,395 |
| 9 | 2,027,025 | 135,135 |
| 10 | 34,459,425 | 2,027,025 |
| | | |

Optimality criteria



Neighbor

- joining (NJ)

Neighbor-joining

- Finds pairs of taxa that minimize **internal branch length**, resulting in the **shortest** tree
- **Distance**-based tree reconstruction method
- Pros: quick, quick and **quick**
- Cons: problematic when sequences are divergent & involve many **gaps**

Neighbor-joining

- Input: **matrix** based on pairwise distance (d) between each pair of taxa (i)

| | A | B | C | D | E |
|---|---|--------|--------|--------|--------|
| A | | 0.1715 | 0.2147 | 0.3091 | 0.2326 |
| B | | | 0.2991 | 0.3399 | 0.2058 |
| C | | | | 0.2795 | 0.3943 |
| D | | | | | 0.4289 |
| E | | | | | |

Neighbor-joining

Step 1: convert pairwise distances to **net divergence** (r_i) using the formula: $r_i = \sum d_{ij}$

| | A | B | C | D | E | r |
|---|---|--------|--------|--------|--------|--------|
| A | | 0.1715 | 0.2147 | 0.3091 | 0.2326 | 0.9279 |
| B | | | 0.2991 | 0.3399 | 0.2058 | 1.0163 |
| C | | | - | 0.2795 | 0.3943 | 1.1876 |
| D | | | | - | 0.4289 | 1.3574 |
| E | | | | | - | 1.2616 |

Neighbor-joining

Step 2: calculate **rate-corrected** distance matrix (M) using: $M_{ij} = d_{ij} - [r_i + r_j]/(N-2)$

| | A | B | C | D | E | r |
|---|---------|---------|---------|---------|--------|--------|
| A | | 0.1715 | 0.2147 | 0.3091 | 0.2326 | 0.9279 |
| B | -0.4766 | - | 0.2991 | 0.3399 | 0.2058 | 1.0163 |
| C | -0.4905 | -0.4356 | - | 0.2795 | 0.3943 | 1.1876 |
| D | -0.4527 | -0.4514 | -0.5689 | - | 0.4289 | 1.3574 |
| E | -0.4972 | -0.5535 | -0.4221 | -0.4441 | - | 1.2616 |

Neighbor-joining

Step 3: choose as neighbors the pairs with the smallest M_{ij} and call node Y

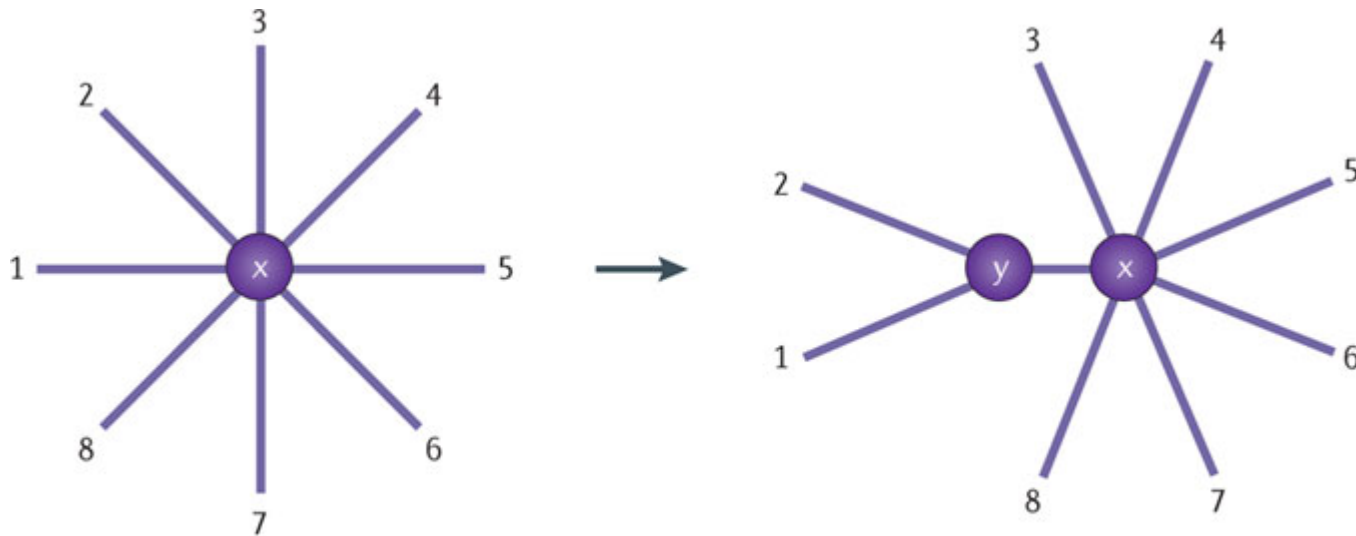
| | A | B | C | D | E | r |
|---|---------|---------|----------------|---------|--------|--------|
| A | | 0.1715 | 0.2147 | 0.3091 | 0.2326 | 0.9279 |
| B | -0.4766 | - | 0.2991 | 0.3399 | 0.2058 | 1.0163 |
| C | -0.4905 | -0.4356 | - | 0.2795 | 0.3943 | 1.1876 |
| D | -0.4527 | -0.4514 | -0.5689 | - | 0.4289 | 1.3574 |
| E | -0.4972 | -0.5535 | -0.4221 | -0.4441 | - | 1.2616 |

This is
C & D!



Neighbor-joining

Step 4: Introduce **first internal branch** & calculate length of new tree



Neighbor-joining

Step 5: Calculate **new distances** between node Y and other terminal nodes

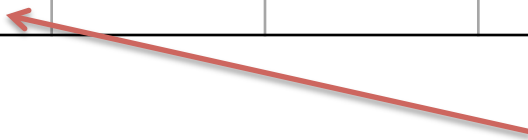
Step 6: Create a rate-corrected matrix (M) where N is now equal to 4:

| | A | B | E | Y | r |
|---|---------|---------------|---------------|---------------|--------|
| A | - | 0.1715 | 0.2326 | 0.1222 | 0.5263 |
| B | -0.3701 | - | 0.2058 | 0.1798 | 0.5571 |
| E | -0.3856 | -0.4278 | - | 0.2719 | 0.3551 |
| Y | -0.4278 | -0.3856 | -0.3701 | - | 0.5739 |

Neighbor-joining

Step 7: Choose the taxa with the **lowest distance** to Y...that would be A

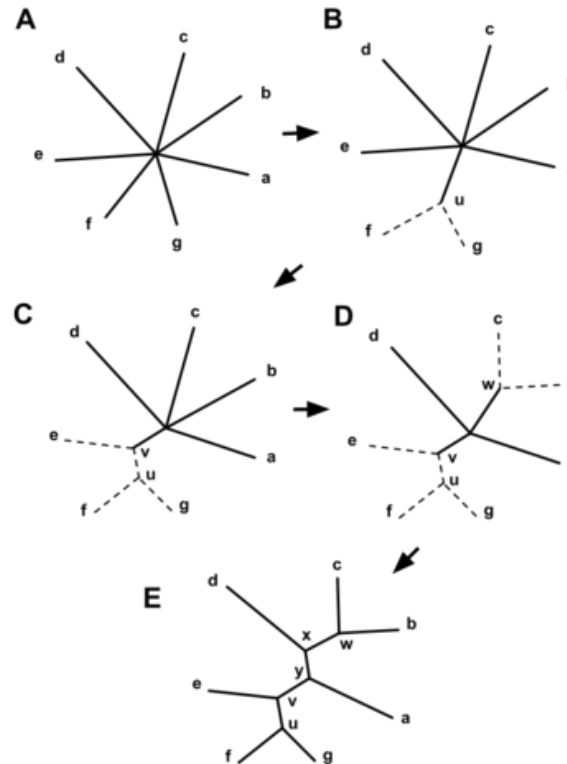
| | A | B | E | Y | r |
|---|----------------|---------|---------|--------|--------|
| A | - | 0.1715 | 0.2326 | 0.1222 | 0.5263 |
| B | -0.3701 | - | 0.2058 | 0.1798 | 0.5571 |
| E | -0.3856 | -0.4278 | - | 0.2719 | 0.3551 |
| Y | -0.4278 | -0.3856 | -0.3701 | - | 0.5739 |



This is A!

Neighbor-joining

Step 8: **Sequentially** introduce pairs of taxa that result in **shortest tree**...



Assessing confidence

Assessing confidence

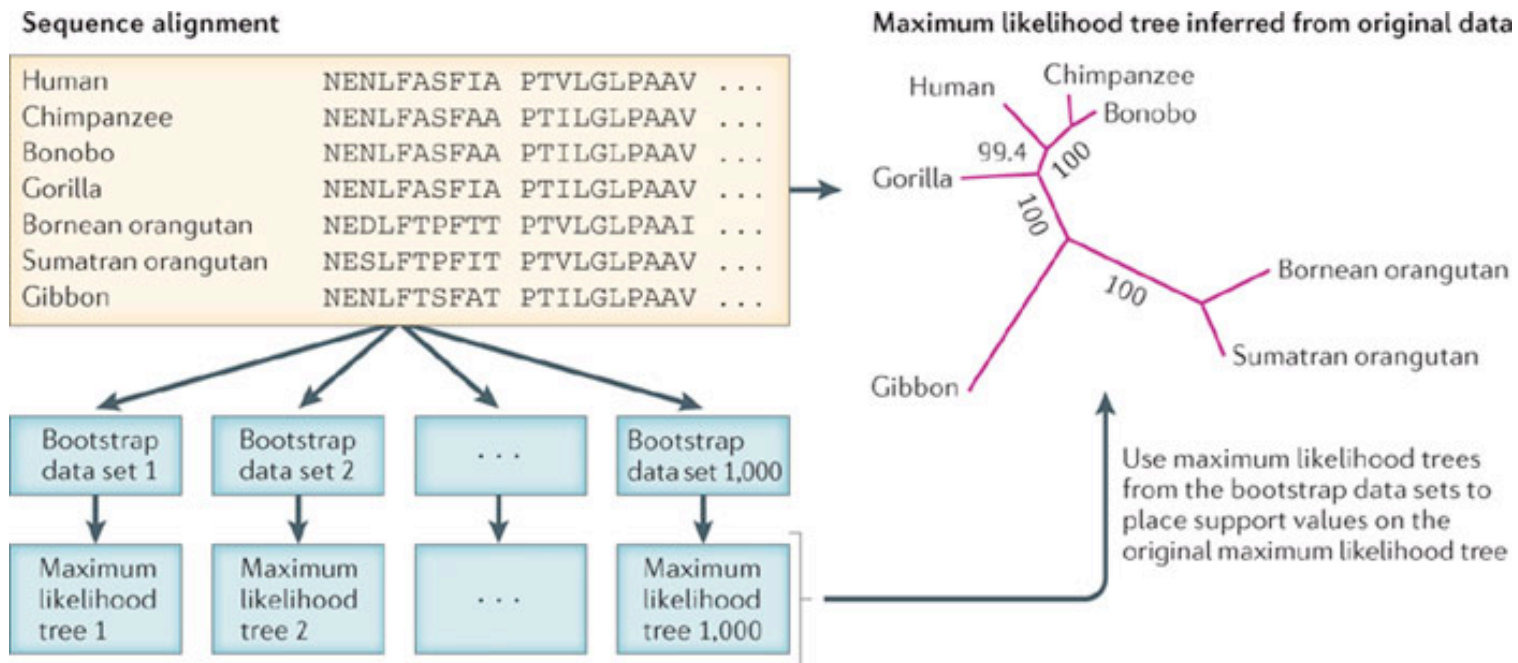
- Trees obtained by phylogenetics are subject to **error** like all other scientific hypotheses
- A tree will be generated regardless of whether there is a phylogenetic signal
- Need to **quantify** how strongly data supports each of the relationships in the tree

Bootstrapping

- Typically tackled with a statistical test called **bootstrapping**
- Assesses chances of recovering a particular clade again if we randomly re-sample our data
- Data matrix is sampled with replacement to produce **pseudo-replicate** datasets
- Measures which parts of the tree are weakly supported with a low bootstrap %

Bootstrapping

An example that uses a Maximum Likelihood (ML) tree...



Nature Reviews | **Genetics**

...bootstrapping is performed 1,000 times generating 1,000 ML trees

Bootstrap cut-offs

- Exact interpretation of bootstrap % is elusive
- Higher is better but what is a reasonable cut-off? 70%?
- Warning: bootstrapping predicts whether the same result would occur if more data were collected not whether the result is actually correct

