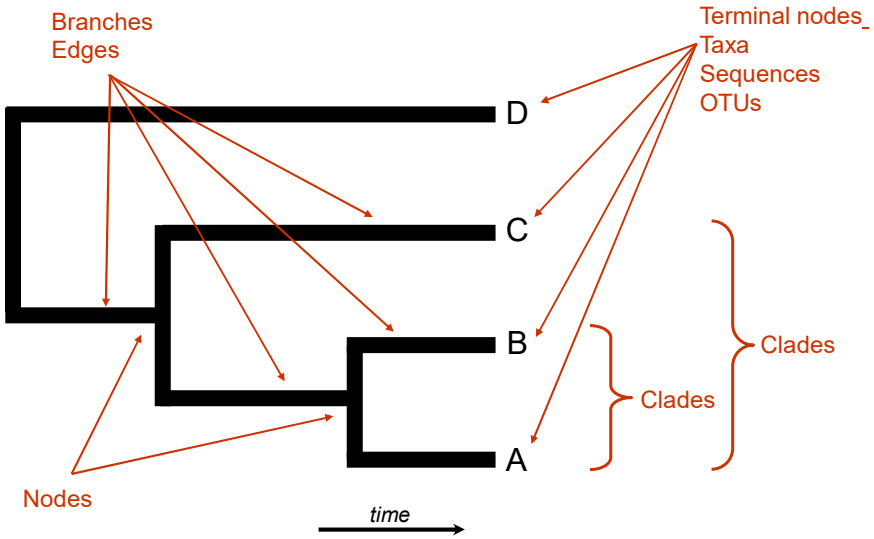
1

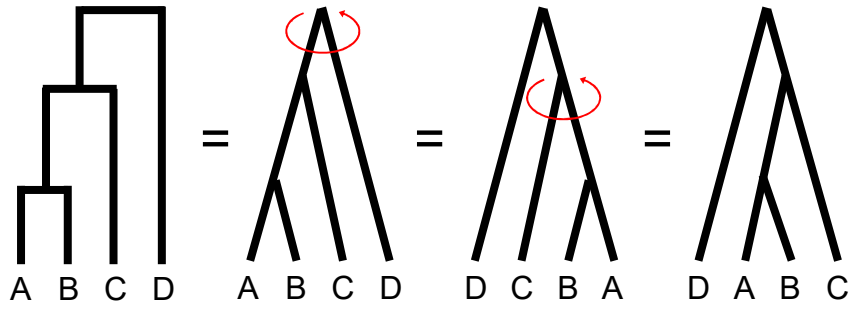
Phylogenetics

- The study of evolutionary relationships.
- Conversion of DNA or protein sequence data into a branching diagram ("tree") that shows the relationships between the sequences.

Phylogenetics
the anatomy of a tree



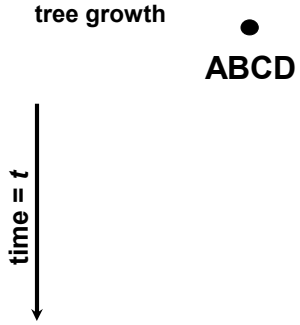
Phylogenetics
the many shapes of trees



2^{N-1} possible arrangements for a particular rooting

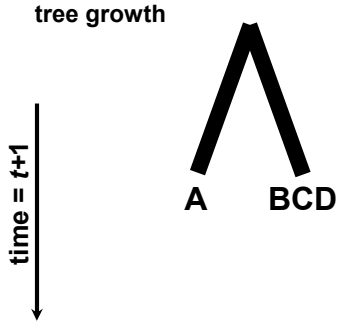
Phylogenetics

tree growth



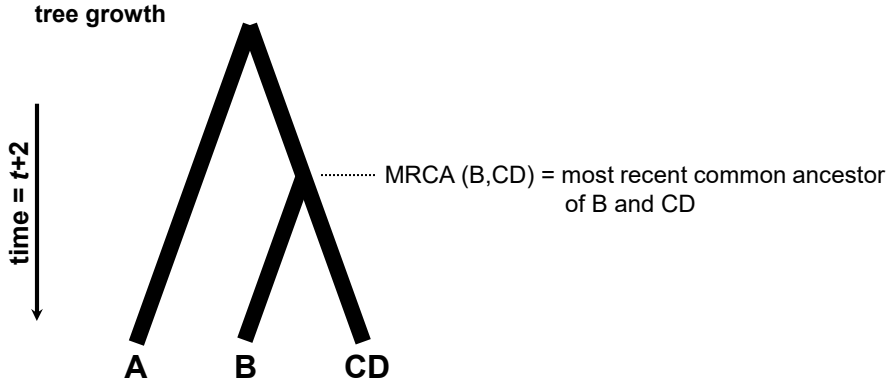
Phylogenetics

tree growth



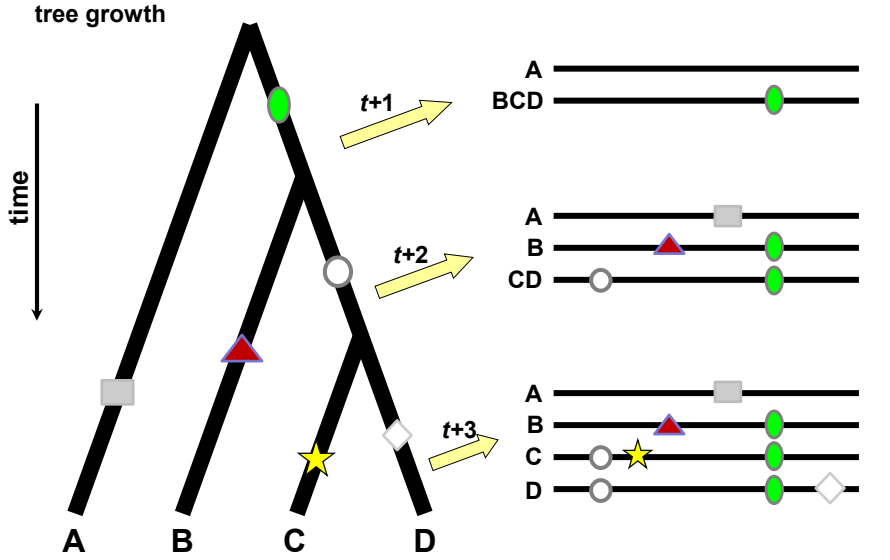
Phylogenetics

tree growth



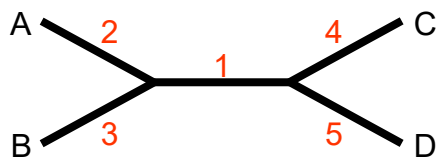
Phylogenetics

tree growth

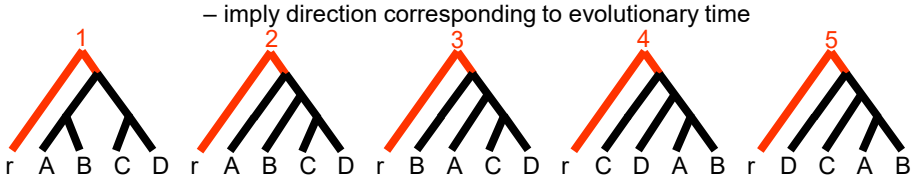


Rooting Trees

Unrooted Tree



Rooted Trees – have one node from which all other nodes descend
– imply direction corresponding to evolutionary time

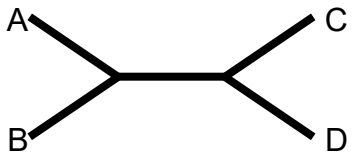


(r(A,B)(C,D)) (r(A(B(C,D)))) (r(B(A(C,D)))) (r(C(D(A,B)))) (r(D(C(A,B))))

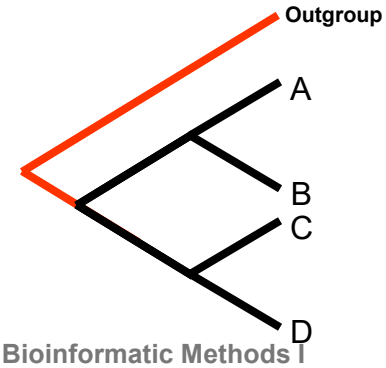
Newick representation

9

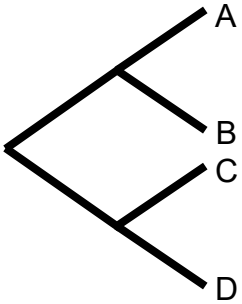
Rooting Trees



Outgroup Rooting



Midpoint Rooting



10

Phylogenetics

terminology

Ancestral State

- a.k.a. plesiomorphy

Derived State

- a.k.a. apomorphy
  - Autapomorphy = unique derived state
  - Synapomorphy = shared derived state

Homoplasy

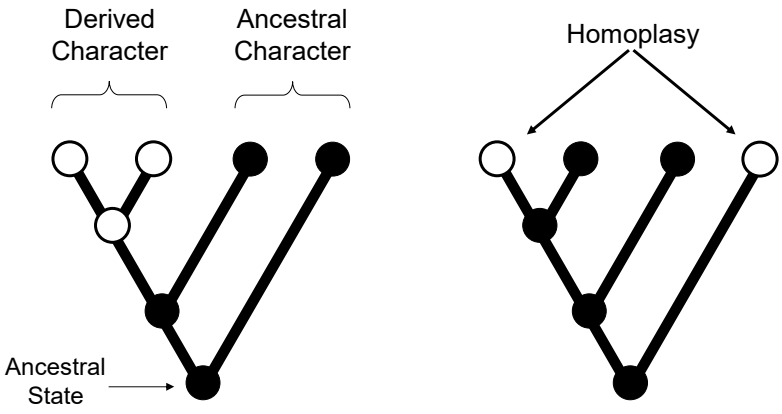
- Similarity due to parallel evolution, convergent evolution or secondary loss

Homology

- Similarity due to common ancestry

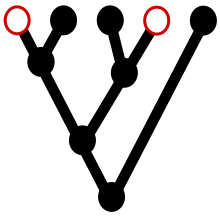
Phylogenetics

terminology



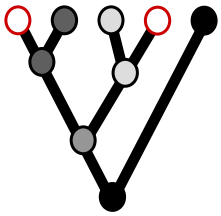
Phylogenetics

homoplasy



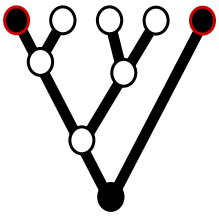
Parallel Evolution

Independent evolution of same character from same ancestral state



Convergent Evolution

Independent evolution of same character from different ancestral state

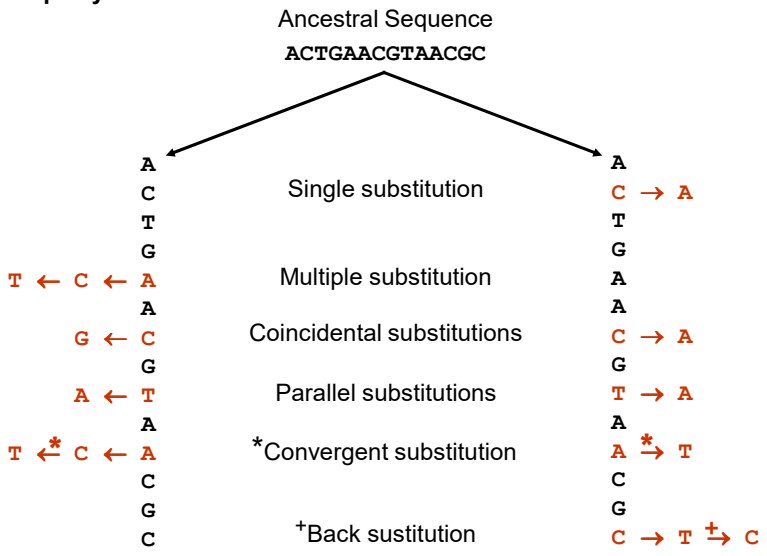


Secondary Loss

Reversion to ancestral state

Phylogenetics

homoplasy



Phylogenetics

fundamental elements

Taxa

- Sampling

Loci

- Homology
- Variation
- Independence

Analysis

- Data
- Sequence alignments
- Phylogenetic methods
- Statistical support

Phylogenetics

tree building methods

Distance methods

- UPGMA (Unweighted Pair Group Method with Arithmetic mean)
- **Neighbour-joining**

Character-based (discrete) methods

- Maximum parsimony
- **Maximum likelihood**

Phylogenetics and Recombination – how would recombination affect interpretation of a tree?

Phylogenetics

distance-based methods

Relationships based upon sequence similarity.

Advantages

- Computationally fast.
- Single “best tree” found.

Disadvantages

- Assumptions
  - additive distances (always)
  - molecular clock (sometimes)
- Information loss occurs due to data transformation
- Uninterpretable branch lengths
- Single “best tree” found.

Neighbour-Joining

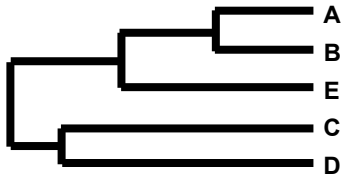
1. Calculate pairwise distances
2. Create distance matrix
3. Determine net divergence for each terminal node
4. Create rate-corrected distance matrix
5. Identify taxa with minimum rate-corrected distance
6. Connect taxa with minimum rate-corrected distance via a new node, and determine their distance from this new node
7. Determine the distance of new node from rest of taxa or nodes
8. Regenerate distance matrix
9. Return to step 2

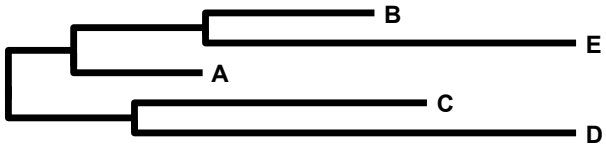
Distance-Based Phylogenetic Methods

UPGMA vs. Neighbour-Joining

Pairwise distances - upper diagonal
Rate-corrected distances - lower diagonal

	A	B	C	D	E	r_i
A	-	17	21	31	23	92
B	-48	-	30	34	21	102
C	-49	-43	-	28	39	118
D	-45	-45	-57	-	43	136
E	-50	-55	-42	-44	-	126





19

Character-Based Phylogenetic Methods

Maximum Likelihood

Attempts to answer the question:

- What is the probability of observing the data, given a particular model of evolution and evolutionary history?
  - data = MSA
  - model = transition probabilities, base frequencies, rate heterogeneity...
  - evolutionary history = phylogenetic tree

Evaluates the likelihood of every substitution of every possible tree.

All possible trees are considered, and the number of substitutions that must have occurred are calculated.

The tree with the highest likelihood is assumed to be the correct tree.

20

Likelihood
coin example

Likelihood (L) = Probability ($data_{observed}$ | model)

Data : HHTHTH

Model 1 : fair coin	Prob(H) = 0.5, Prob(T) = 0.5
Model 2 : 2-head coin	Prob(H) = 1.0, Prob(T) = 0.0
Model 3 : 2-tail coin	Prob(H) = 0.0, Prob(T) = 1.0

L (Data|Model1)
= Prob(H|Model1) * Prob(H|Model1) * Prob(T|Model1) * Prob(H|Model1) * Prob(T|Model1) * Prob(H|Model1)
= 0.5 * 0.5 * 0.5 * 0.5 * 0.5 * 0.5 = 0.0156

L (Data|Model2) = 1.0 * 1.0 * 0.0 * 1.0 * 0.0 * 1.0 = 0.0

L (Data|Model3) = 0.0 * 0.0 * 1.0 * 0.0 * 1.0 * 0.0 = 0.0

Likelihood
maximum likelihood

Find the model that maximizes the likelihood of the observed data

Data : GGACGCCTGACGCCGCTCGG

Model 1: equal base composition - 0.25, 0.25, 0.25, 0.25 – A, C, G, T, respectively
Model 2: G+C bias - 0.1, 0.4, 0.4, 0.1 – A, C, G, T, respectively
Model 3: A+T bias - 0.4, 0.1, 0.1, 0.4 – A, C, G, T, respectively

L (Data|Model1) = Prob(G|Model1)*Prob(G|Model1)*Prob(A|Model1)*...* Prob(G|Model1) = 0.25^{20} = 9.1×10^{-13}

L (Data|Model2) = $0.4^{16} * 0.1^{4} = 4.3 \times 10^{-11}$ ← maximum likelihood

L (Data|Model3) = $0.1^{16} * 0.4^{4} = 2.6 \times 10^{-18}$

Likelihood

maximum likelihood models in phylogenetics

Find the tree topology with the highest likelihood given a particular evolutionary model

Nucleotide substitution evolutionary models typically have 2 components

- Composition
  - nucleotide proportions
- Process
  - how the nucleotides change over time

Maximum Likelihood

Advantages of ML methods

- Based on explicit evolutionary models.
- Permits statistical evaluation of the likelihood of specific tree topologies.
- Often returns many equally likely trees.
- Usually outperforms other methods.

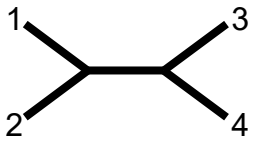
Disadvantages

- Computationally very intensive.
- Often returns many equally likely trees.

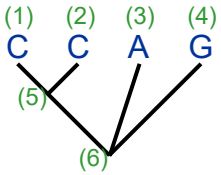
Maximum Likelihood

```
   1.........j........N
1  C...GGACACGTTTA...C
2  C...AGACACCTCTA...C
3  C...GGATAAGTTAA...C
4  C...GGATAGCCTAG...C
```

Unrooted tree for the 4 taxa



Arbitrarily rooted tree for site *j*



Maximum Likelihood

L_j = P(C C A G; A, A) * P(C C A G; G, A) * P(C C A G; C, A) * P(C C A G; T, A)
* P(C C A G; A, G) * P(C C A G; G, G) * P(C C A G; C, G) * P(C C A G; T, G)
* P(C C A G; A, C) * P(C C A G; G, C) * P(C C A G; C, C) * P(C C A G; T, C)
* P(C C A G; A, T) * P(C C A G; G, T) * P(C C A G; C, T) * P(C C A G; T, T)

Maximum Likelihood

Likelihood of the tree = product of the likelihoods for each site.

$$L = L_1 \times L_2 \times \ldots \times L_N = \prod_{j=1}^{N} L_j$$

Usually evaluated as the sum of the log likelihoods.

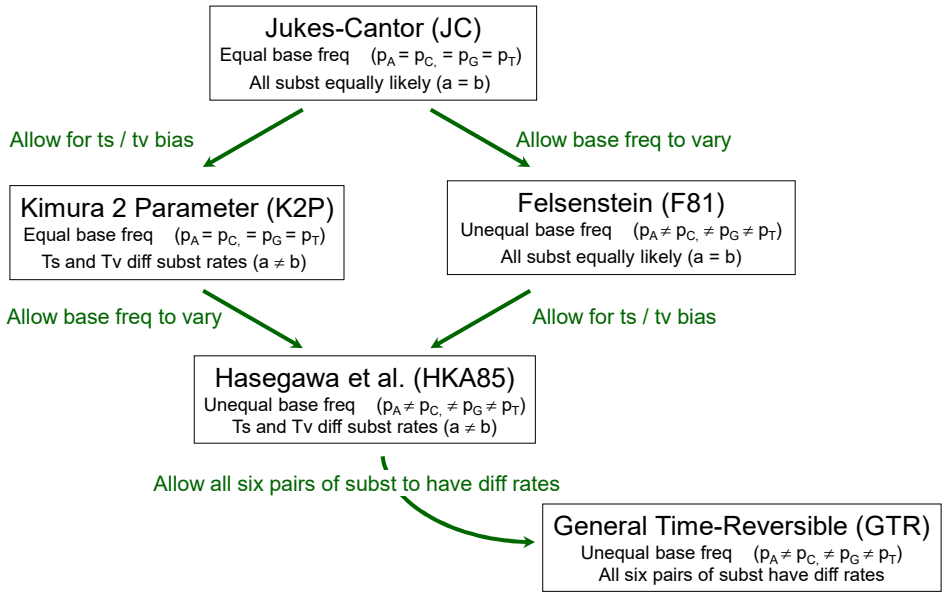
$$\ln L = \ln L_1 + \ln L_2 + \ldots + \ln L_N = \sum_{j=1}^{N} \ln L_j$$

ML evaluates:

- all possible ancestral states
  - at all variable sites
    - in all possible tree topologies

→The most likely (best) tree is the topology that has the highest overall likelihood.

Models of Sequence Evolution



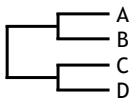
Phylogenetic Confidence

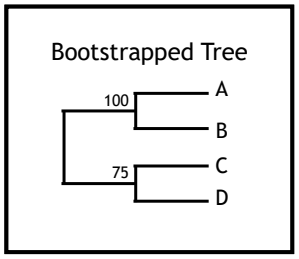
Bootstrapping

Original Sequence and Tree

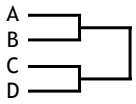
```
seqA    AGGCTCCAAA
seqB    AGGTTCGAAA
seqC    AGCCCCGAAA
seqD    ATTTCCGAAC

pr1     1310110012
pr2     1000222003
pr3     0120401200
...
pr1000  1010220112
```



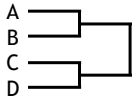


Pseudo-Replicated Data and Trees



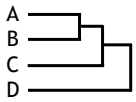
pseudo-replicate 1

```
seqA    AGGGGTCAAA
seqB    AGGGGUCAAA
seqC    AGGGCCCAAA
seqD    ATTTTCCACC
```



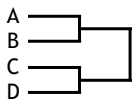
pseudo-replicate 2

```
seqA    ATTCCCCAAA
seqB    ATTCCGGAAA
seqC    ACCCCGGAAA
seqD    ACCCCGGCCC
```



pseudo-replicate 3

```
seqA    GGGTTTTCAA
seqB    GGGTTTTGAA
seqC    GCCCCCCGAA
seqD    TTTCCCCGAA
```



⋮

pseudo-replicate 1000

```
seqA    AGTTCCAAAA
seqB    AGTTCCAAAA
seqC    ACCCCCAAAA
seqD    ATCCCCAACC
```


Phylogenetic Confidence

Bootstrapping

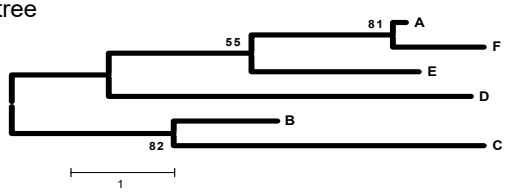
Assumptions

- Data size is large enough to accurately reflect the true error distribution
- The data are identically and independently distributed

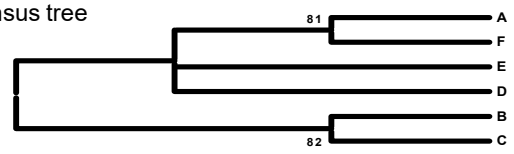
Phylogenetic Confidence

Bootstrapping

Bootstrapped NJ tree



Bootstrap consensus tree

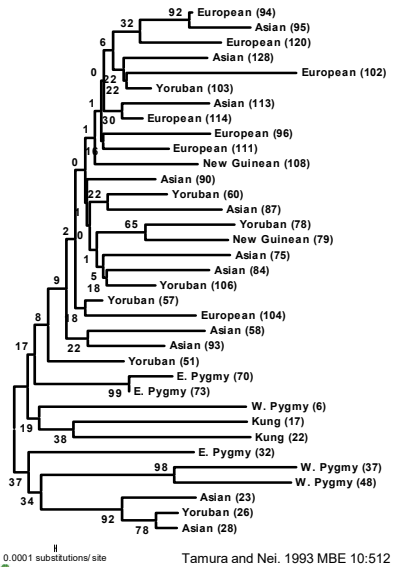


Bootstrap values

> 90%	strongly supported
70 > 90%	well supported
50 > 70%	weakly supported
< 50%	not supported

Phylogenetic Confidence

Bootstrapping



Tamura and Nei. 1993 MBE 10:512