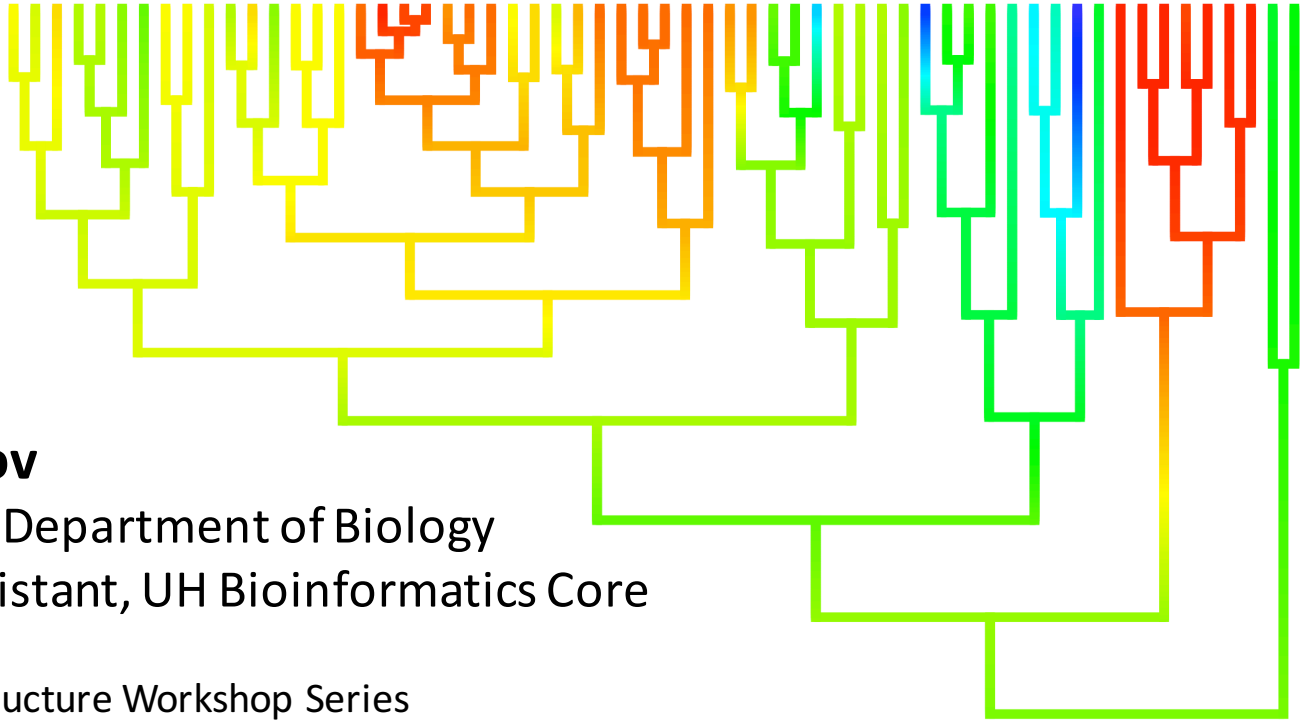
An Introduction to Molecular Phylogenetic Inference



Kirill Vinnikov
PhD Student, Department of Biology
Graduate Assistant, UH Bioinformatics Core

ITS Cyberinfrastructure Workshop Series
September 9, 2016

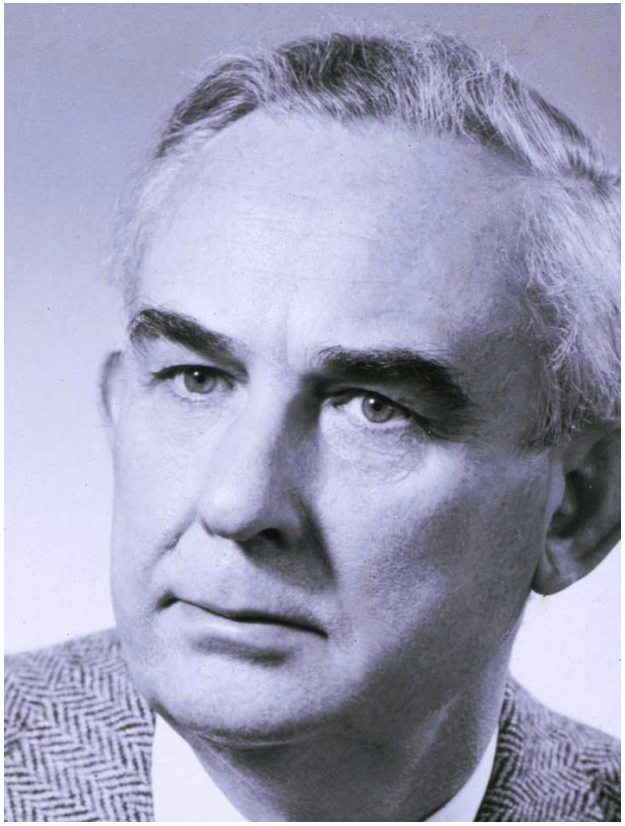
3. Introduction to phylogenetic methods

Outline

- Short historical overview
- Maximum parsimony
- Distance-matrix methods
- Parameters of the evolutionary models
- Introduction to probabilistic methods
- Relaxing clocks and tree calibration

Hennig's Auxiliary Principle

Shared derived character states (synapomorphies) should be considered as evidence of monophyly unless they contradict to much stronger evidence obtained from other characters (homoplasy).



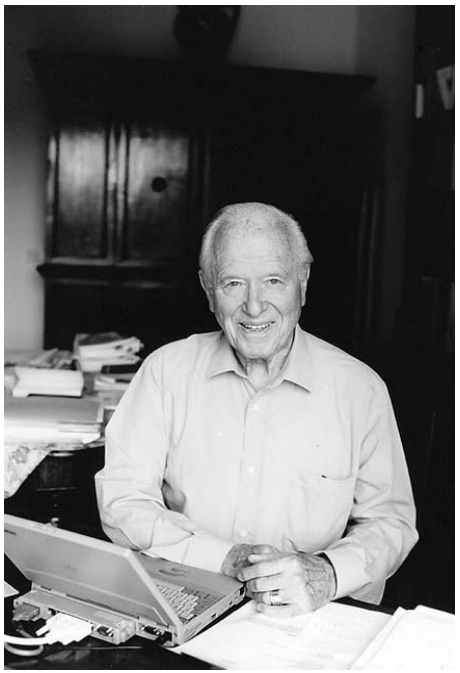
Willi Hennig (1913-1976)

Parsimony Principle

Phylogenetic tree must represent "the minimum net amount of evolution"

Edwards & Cavalli-Sforza, 1963, 1964

Anthony Edwards
(b. 1935)



Luca Cavalli-Sforza
(b. 1922)

Ockham's Razor Principle

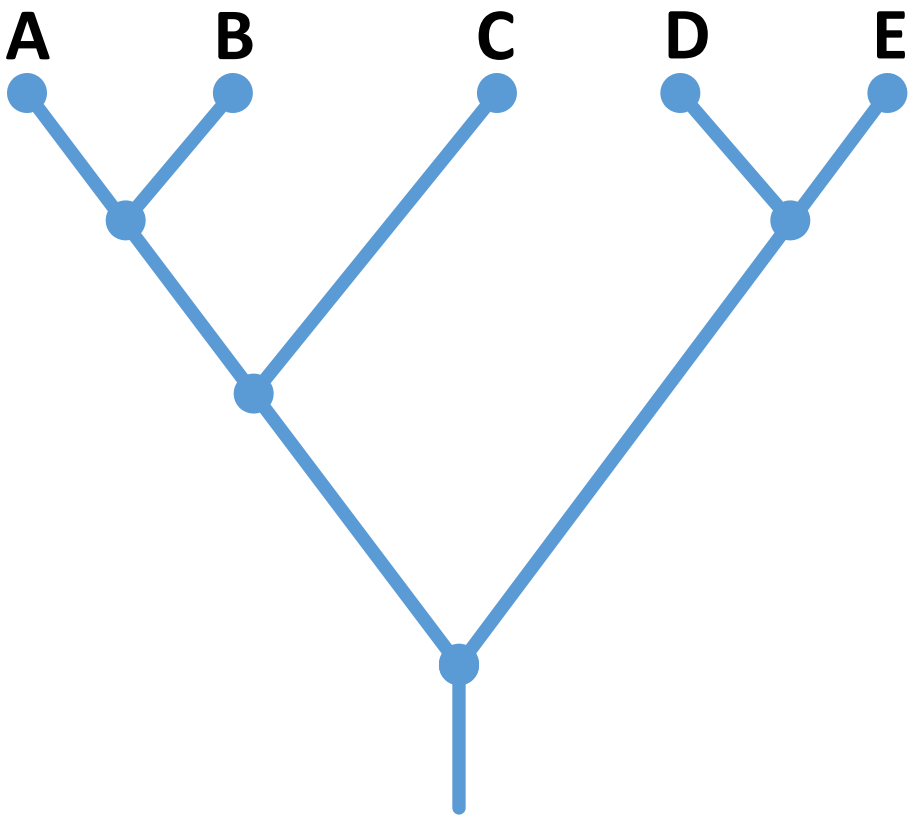
Among competing hypotheses, the one with the fewest assumptions should be selected.

Image sources: royalsociety.org, www.companhiadasletras.com.br

Assumptions for parsimony approach

1. Lineages on a phylogenetic tree are formed through reproduction process.
2. Branches on a phylogenetic tree represent diverging lineages.
3. Homologous characters are inherited through all generations.
4. Characters are independent.
5. Evolution is explained by **a minimal number of changes between character states**.

How to measure a particular phylogeny?

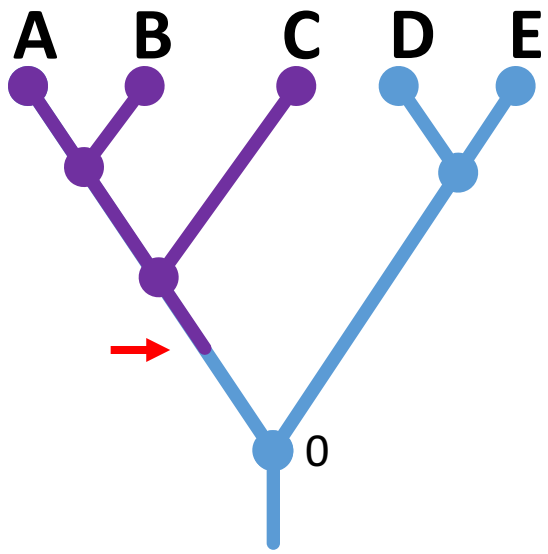


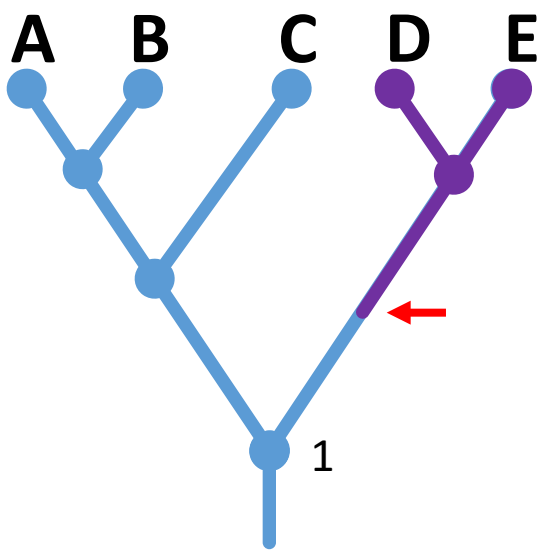
Character state matrix

	Characters					
Species	1	2	3	4	5	6
A	1	0	0	1	1	0
B	1	1	0	1	1	1
C	1	1	0	0	0	0
D	0	0	1	0	0	0
E	0	0	1	1	1	0

0 – absence
1 – presence

Possible scenarios of evolution for: character 1



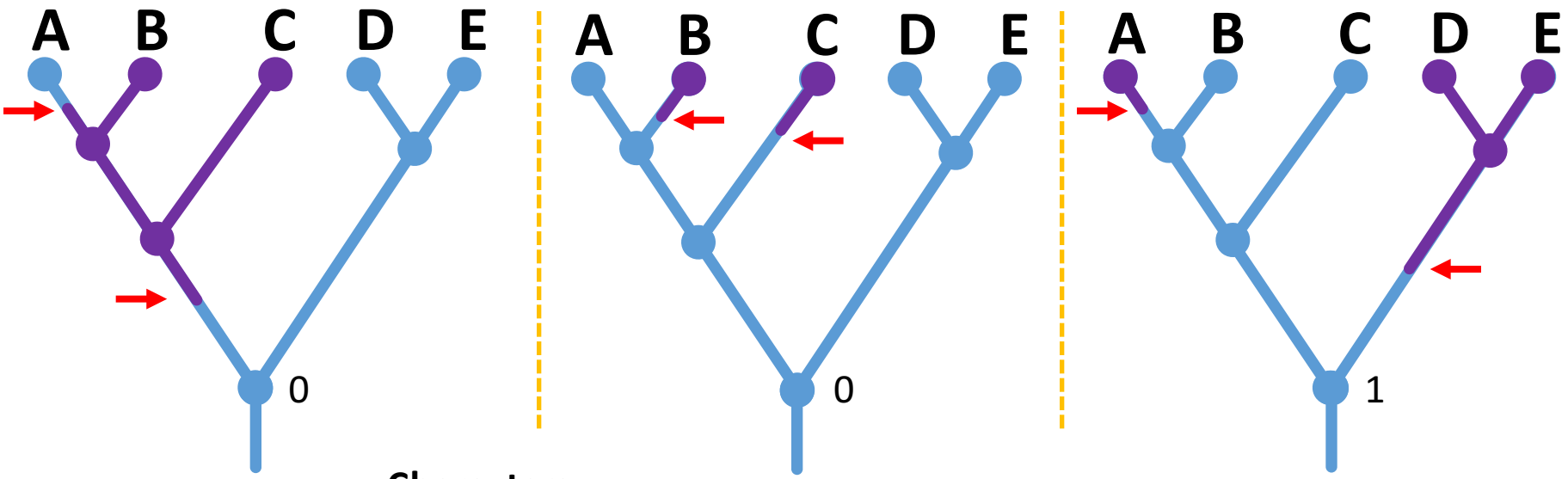


Characters

Species	1	2	3	4	5	6
A	1	0	0	1	1	0
B	1	1	0	1	1	1
C	1	1	0	0	0	0
D	0	0	1	0	0	0
E	0	0	1	1	1	0

One change of character states is required

Possible scenarios of evolution for: character 2



Characters

Species	1	2	3	4	5	6
A	1	0	0	1	1	0
B	1	1	0	1	1	1
C	1	1	0	0	0	0
D	0	0	1	0	0	0
E	0	0	1	1	1	0

Two changes of character states are required

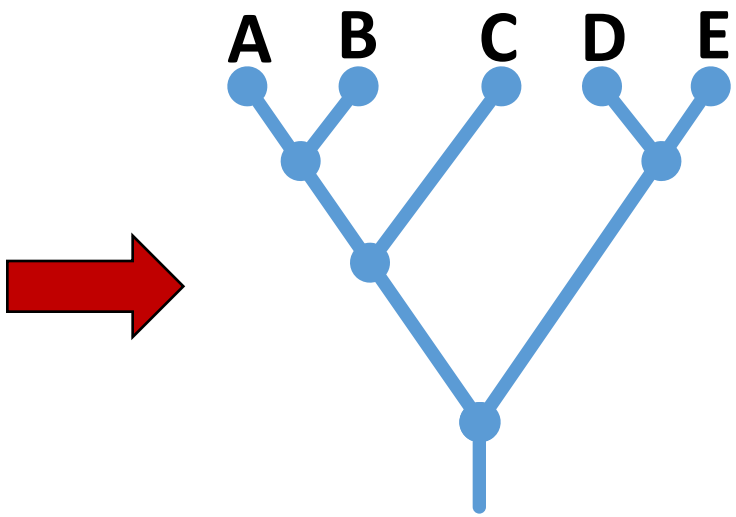
…Repeat procedure for all characters

How to measure a particular phylogeny?

Tree length = 1 + 2 + 1 +2 + 2 + 1 = **9**

the number of changes
for each character

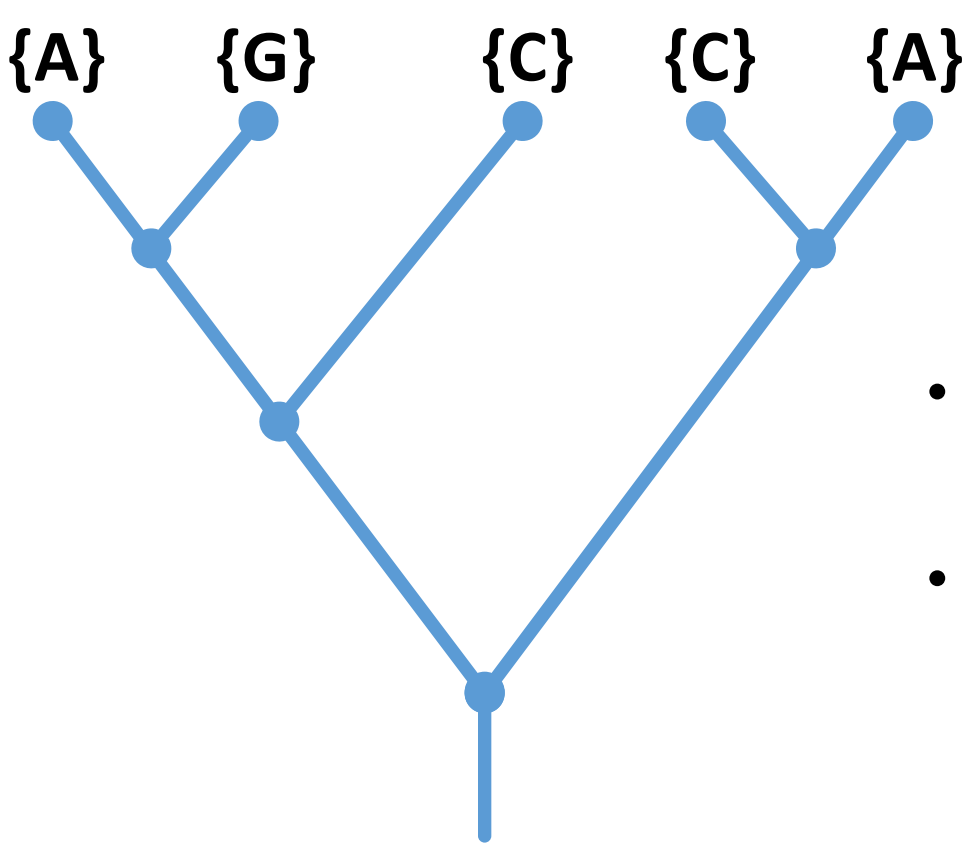
Species	Characters 1	2	3	4	5	6
A	1	0	0	1	1	0
B	1	1	0	1	1	1
C	1	1	0	0	0	0
D	0	0	1	0	0	0
E	0	0	1	1	1	0



Parsimony approach

- For each column in the multiple sequence alignment, the algorithm must compute all the changes in all the possible trees and reconstruct character states at interior nodes.
- Algorithm must handle more complex character states (e.g., nucleotides in DNA or amino acids in protein).

Fitch algorithm
to count changes in character states

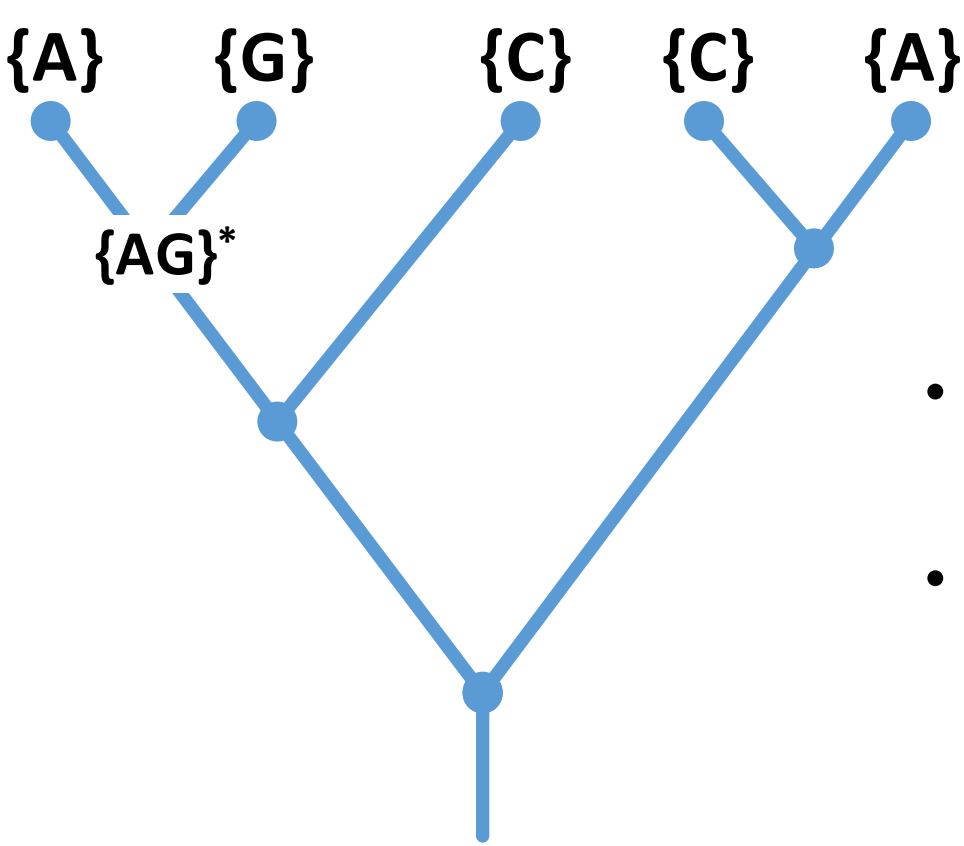


Nucleotide site in DNA sequence

- Keep only those states that overlap in the MRCA node.
- Merge states for ancestral node if they do not overlap in its descendants:

  *= change count

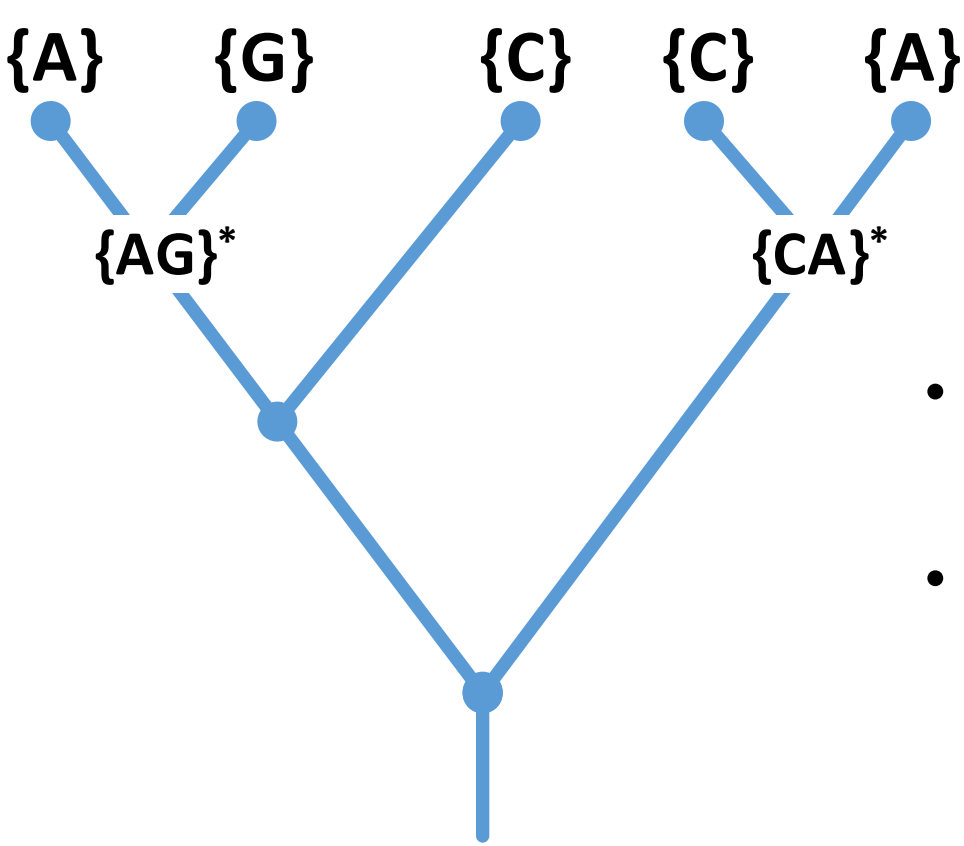
Fitch algorithm
to count changes in character states



- Keep only those states that overlap in the MRCA node.
- Merge states for ancestral node if they do not overlap in its descendants:

  *= change count

Fitch algorithm
to count changes in character states

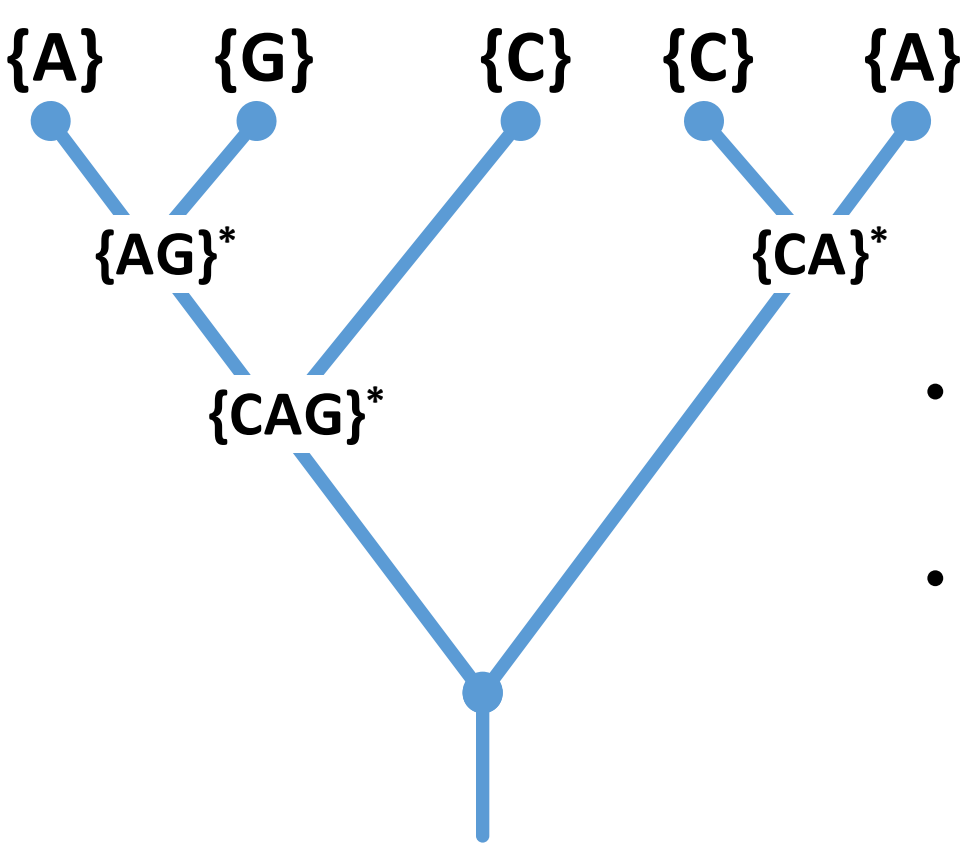


Nucleotide site in DNA sequence

- Keep only those states that overlap in the MRCA node.
- Merge states for ancestral node if they do not overlap in its descendants:

  *= change count

Fitch algorithm
to count changes in character states

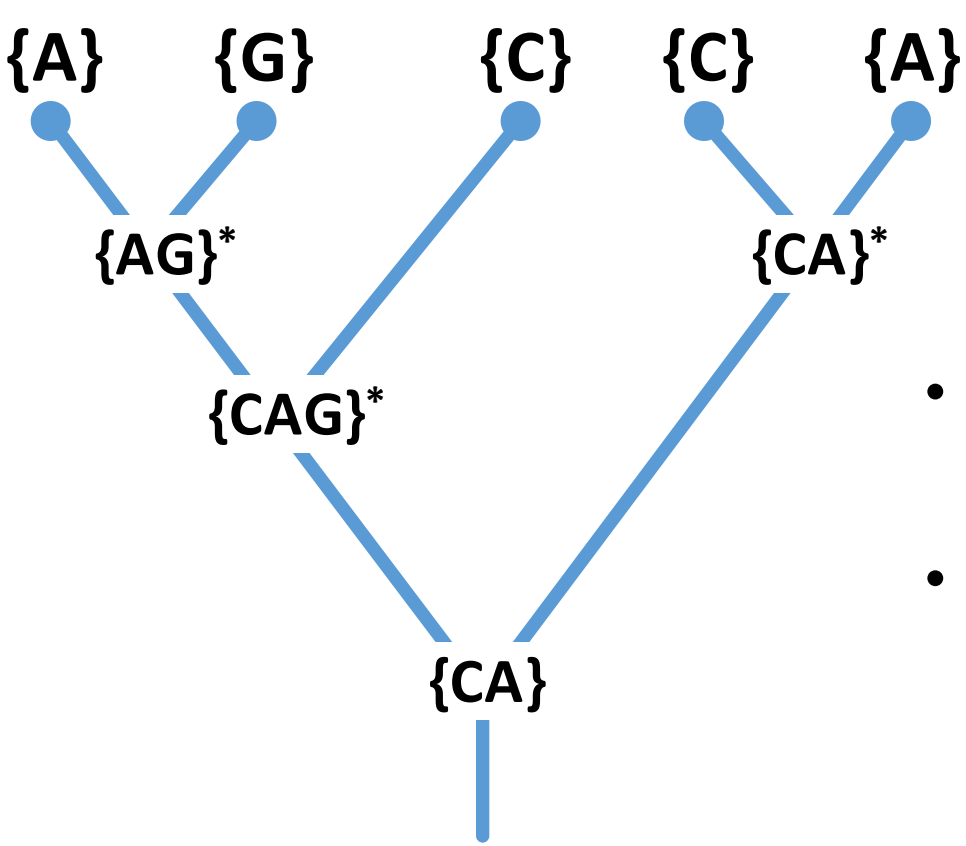


Nucleotide site in DNA sequence

- Keep only those states that overlap in the MRCA node.
- Merge states for ancestral node if they do not overlap in its descendants:

  *= change count

Fitch algorithm
to count changes in character states



Nucleotide site in DNA sequence

- Keep only those states that overlap in the MRCA node.
- Merge states for ancestral node if they do not overlap in its descendants:

  *= change count

Total = 3 changes (*'s) of character states

How many possible phylogenies are there?

$$\#\,\boldsymbol{rooted\ trees} = \frac{(2\boldsymbol{n}-3)!}{2^{\boldsymbol{n}-2}(\boldsymbol{n}-2)!}$$

$$\#\,\boldsymbol{unrooted\ trees} = \frac{(2\boldsymbol{n}-5)!}{2^{\boldsymbol{n}-3}(\boldsymbol{n}-3)!}$$

# OTUs	# of rooted trees
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,374
17	191,898,783,962,510,624
18	6,332,659,870,762,850,304
19	221,643,095,476,699,758,592
20	8,200,794,532,637,890,838,528

Tree search methods

- Exhaustive search (exact)
- Branch-and-bound search (exact)
- Heuristic search methods (approximate)

Exhaustive search

- Every possible phylogenetic tree is examined
- The best tree will be always found
- Computationally intensive (very slow)

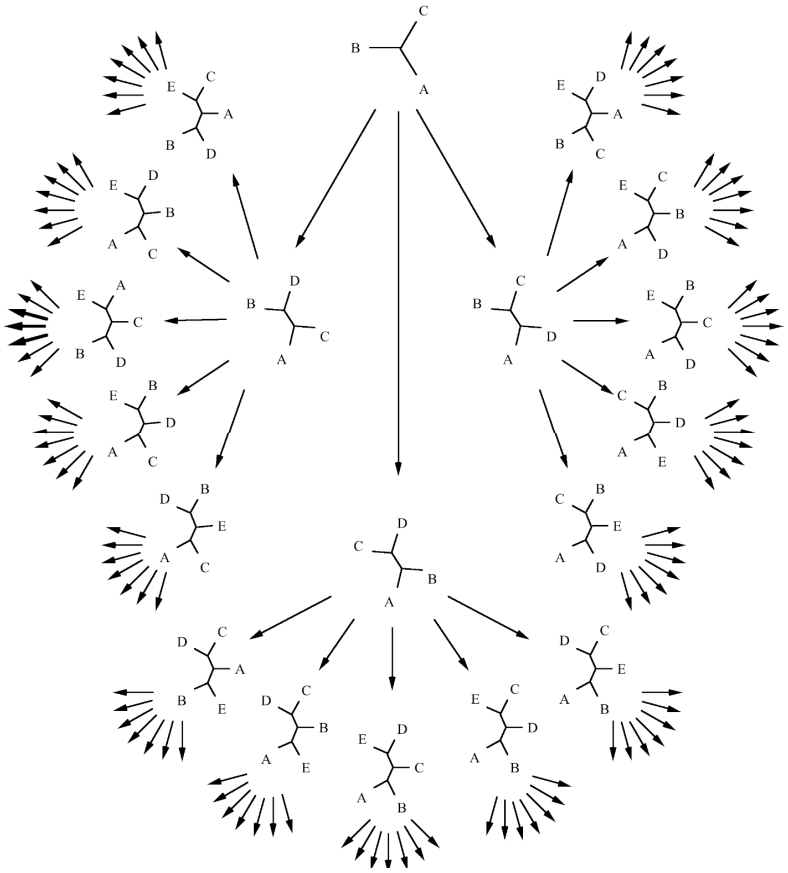
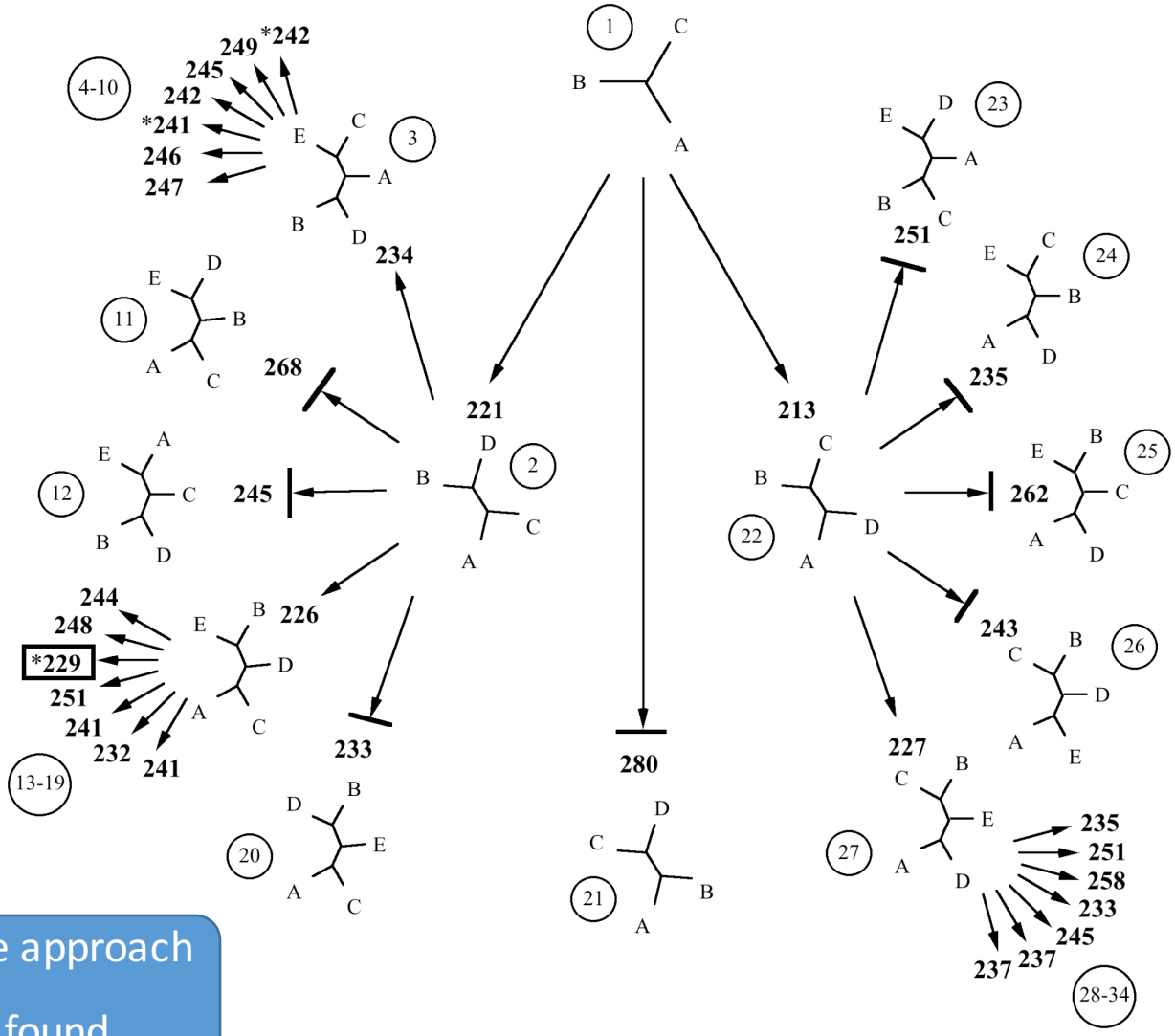


Image source: Swofford & Sullivan, 2009

Branch-and-bound search

Terminates search path if a tree has a score that exceeds the current lower bound

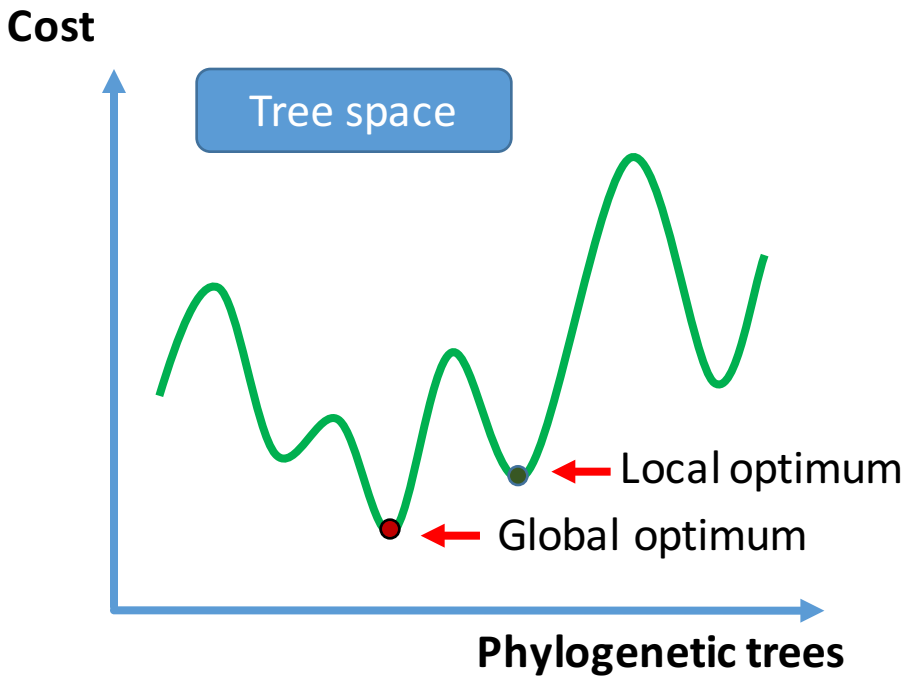


- Much faster than exhaustive approach
- The best tree will be always found

Image source: Swofford & Sullivan, 2009

Heuristic search methods

- Stepwise Addition (SA)
- Nearest-neighbor interchanges (NNI)
- Subtree pruning and regrafting (SPR)
- Tree bisection and reconnection (TBR)
- Star-decomposition methods (e.g., UPGMA and Neighbor-Joining)
- Simulated annealing (Metropolis-Hastings algorithm)



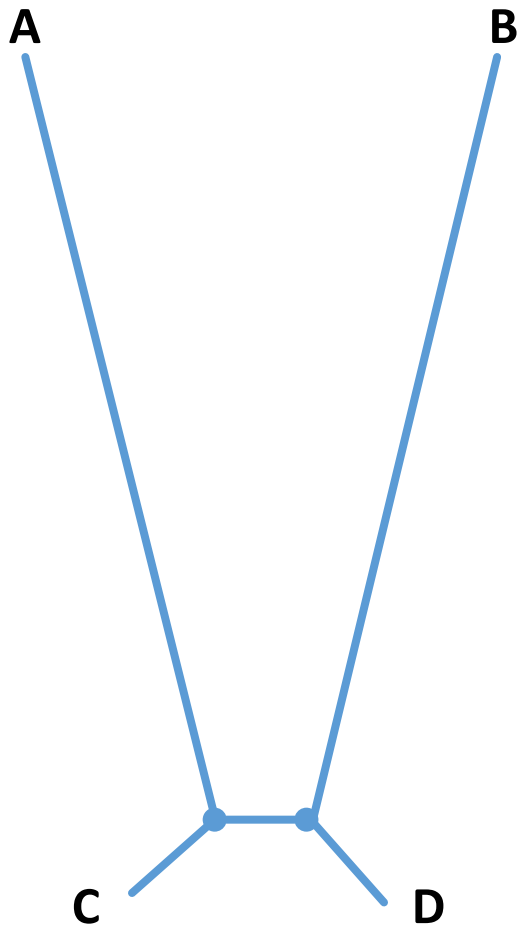
Do not always guarantee the best phylogenetic tree!

What if Maximum Parsimony finds multiple equally parsimonious (optimal) trees?

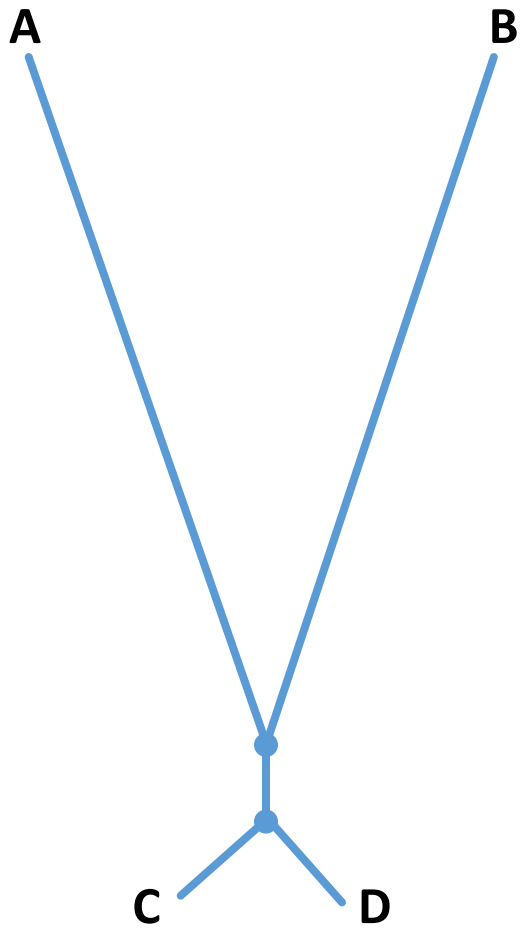
Types of consensus trees:

- Strict consensus tree (includes branches found in all trees)
- Majority-rule consensus tree (includes branches found in the majority of trees – in more than 50% of trees)
- Maximum clade credibility tree (bayesian trees)

Long branch attraction (Felsenstein, 1978)



TRUE PHYLOGENY



INFERRED PHYLOGENY

Maximum Parsimony: summary

Pros:

- No strict dependency on evolutionary model
- Easy to trace character evolution on a tree

Cons:

- Non-statistical approach
- Produces misleading results when using many homoplastic characters
- Highly sensitive to long branch attraction bias (Felsenstein, 1978)
- Assumptions are still questionable

Phenetics (Sokal & Sneath, 1962)

Algorithms for clustering organisms by similarity in their morphological characters (e.g., UPGMA)

left: Peter Sneath (1923-2011)
right: Robert Sokal (1926-2012)

Numerical Taxonomy

- First robust statistical approach
- Accounts phenotypic similarities rather than ancestry

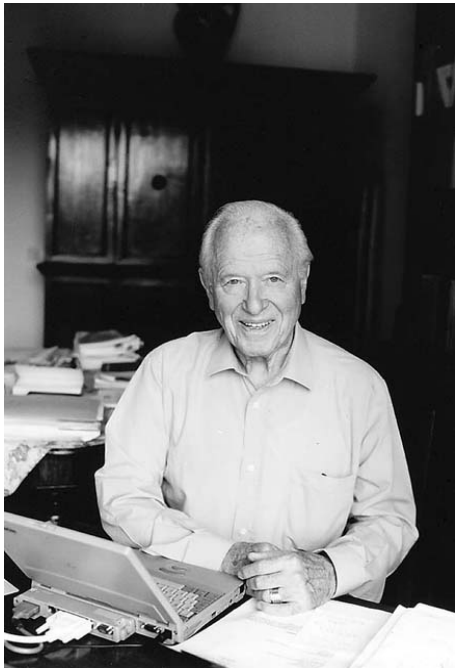
Photo courtesy: Daniel Barker

Distance-matrix methods

The goal was to measure a distance between each pair of species (OTUs)

Cavalli-Sforza & Edwards, 1966, 1967

Anthony Edwards
(b. 1935)



Luca Cavalli-Sforza
(b. 1922)

Image sources: royalsociety.org, www.companhiadasletras.com.br

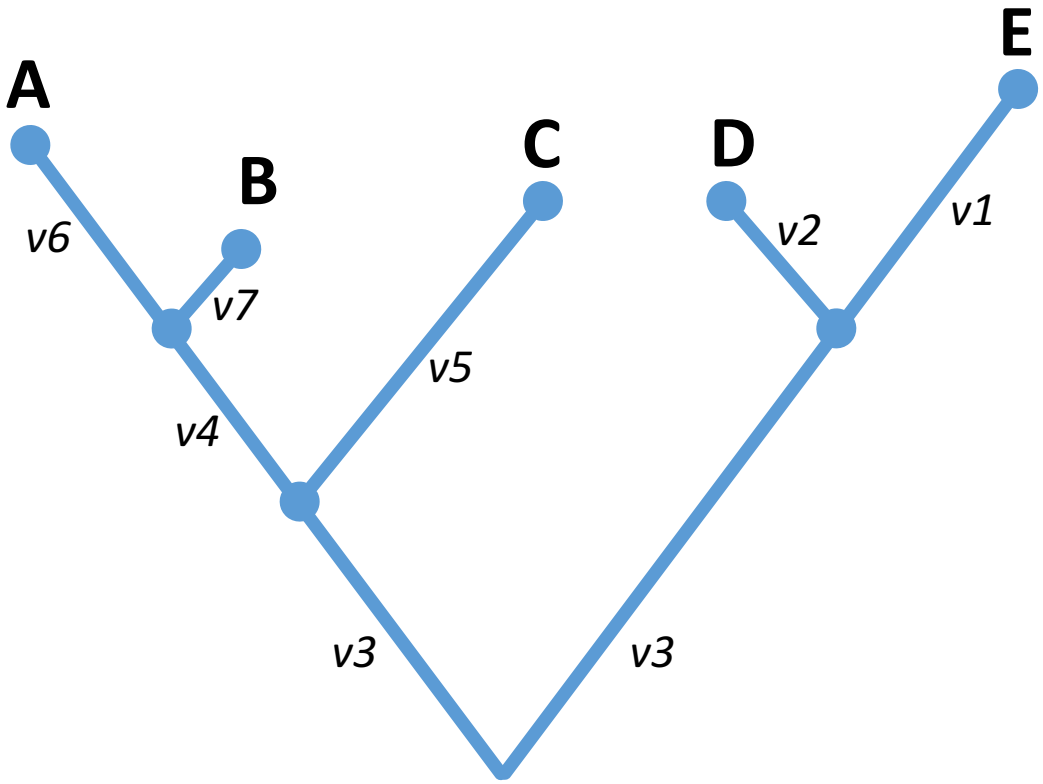
The least square distance method (LS)

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(D_{ij} - d_{ij})^2; \quad i \neq j;\ w_{ij} = {}^{1}\!/_{D_{ij}^2}$$

(Fitch & Margoliash, 1967)

- D_{ij} and d_{ij} are ***observed*** and ***expected*** distances between taxa *i* and *j*
- Algorithm searches for the best tree (topology and branch lengths) that minimizes Q (finds global optimum)
- All distances are independent (no covariance)
- Observed distances are normally distributed: $D_{ij} \sim N(d_{ij}, {}^{K}\!/_{w_{ij}})$

How to measure a distance?



Additive distances:

$D_{AB} = v6 + v7$

$D_{AC} = v4 + v5 + v6$

$D_{AD} = v2 + v3 + v4 + v6$

$D_{AE} = v1 + v3 + v4 + v6$

$D_{BC} = v4 + v5 + v7$

$D_{BD} = v2 + v3 + v4 + v7$

$D_{BE} = v1 + v3 + v4 + v7$

$D_{CD} = v2 + v3 + v5$

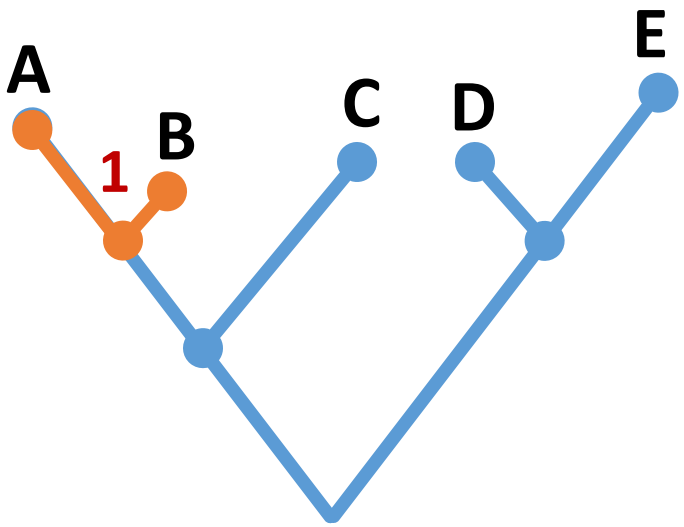
$D_{CE} = v1 + v3 + v5$

$D_{DE} = v1 + v2$

$$d_{ij} = \sum_k x_{ij,k}\, v_k \qquad v_k = rate * time = \mu_i t_i$$

How to measure a distance?

	123456789
Taxon A	AA**G**TGATTC
Taxon B	AA**C**TGATTC
Taxon C	AAGCGTTTG
Taxon D	ATGTGACAG
Taxon E	ATCCGATAG

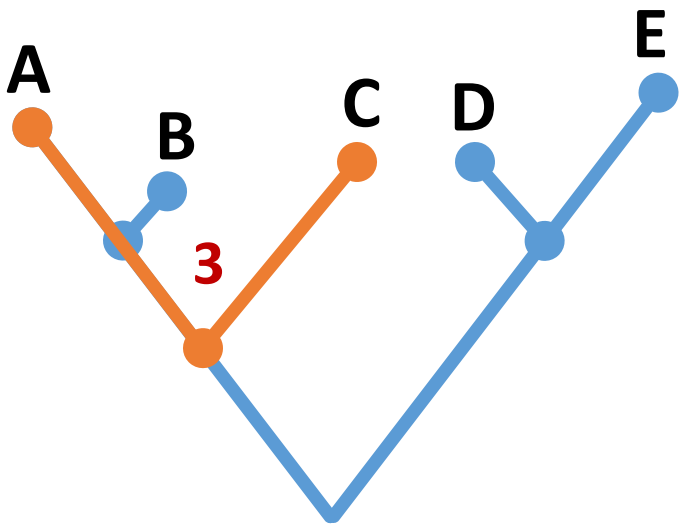


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

	123456789
Taxon A	AAG**T**G**A**TT**C**
Taxon B	AACTGATTC
Taxon C	AAG**C**G**T**TT**G**
Taxon D	ATGTGACAG
Taxon E	ATCCGATAG

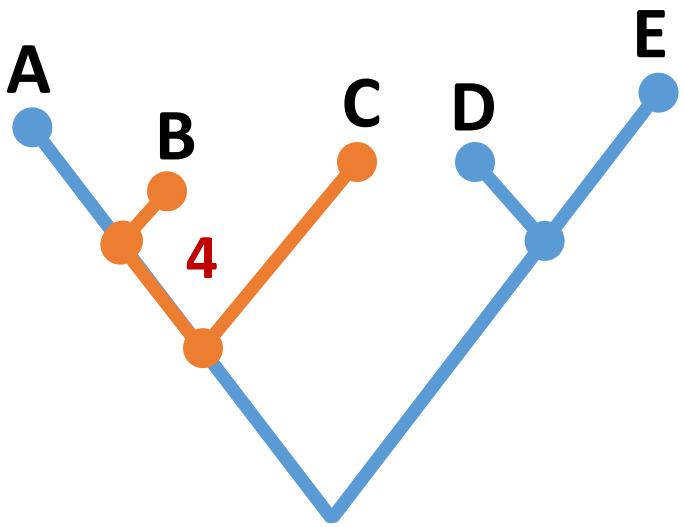


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

	123456789
Taxon A	AAGTGATTC
Taxon B	AA**CT**G**A**TT**C**
Taxon C	AA**GC**G**T**TT**G**
Taxon D	ATGTGACAG
Taxon E	ATCCGATAG

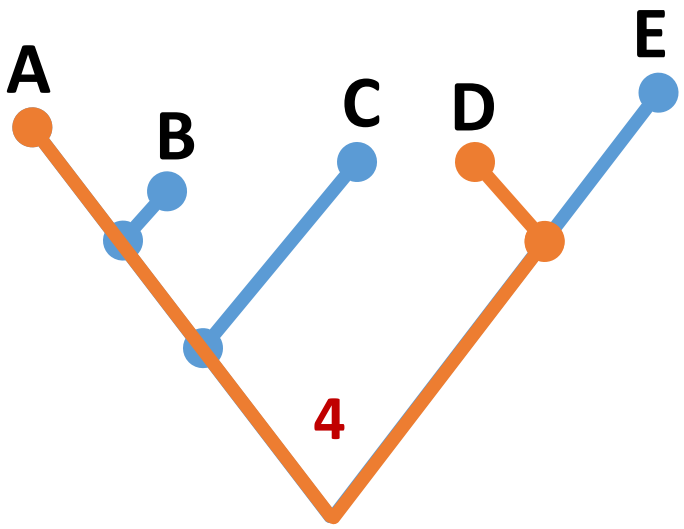


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

	123456789
Taxon A	A**A**GTGA**TTC**
Taxon B	AACTGATTC
Taxon C	AAGCGTTTG
Taxon D	A**T**GTGA**CAG**
Taxon E	ATCCGATAG

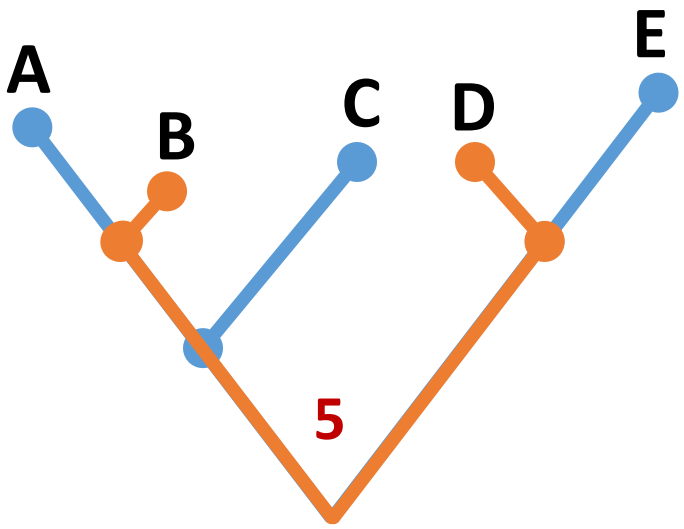


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

	123456789
Taxon A	AAGTGATTC
Taxon B	A**AC**TGA**TTC**
Taxon C	AAGCGTTTG
Taxon D	A**TG**TGA**CAG**
Taxon E	ATCCGATAG



Number of substitutions between pairs of taxa

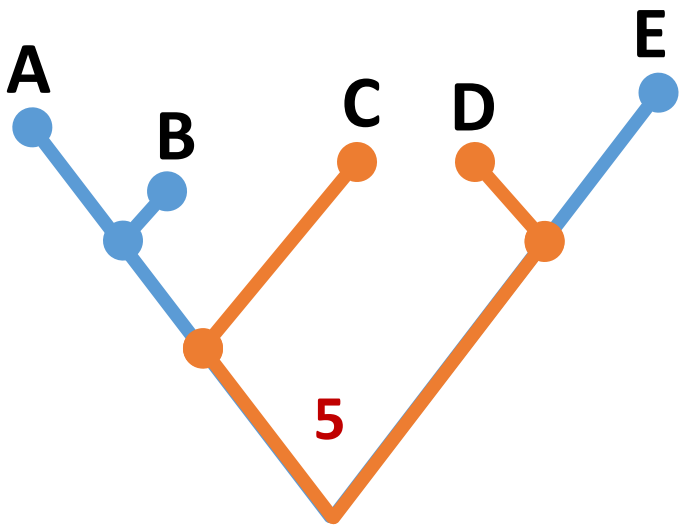
	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

Number of substitutions between pairs of taxa

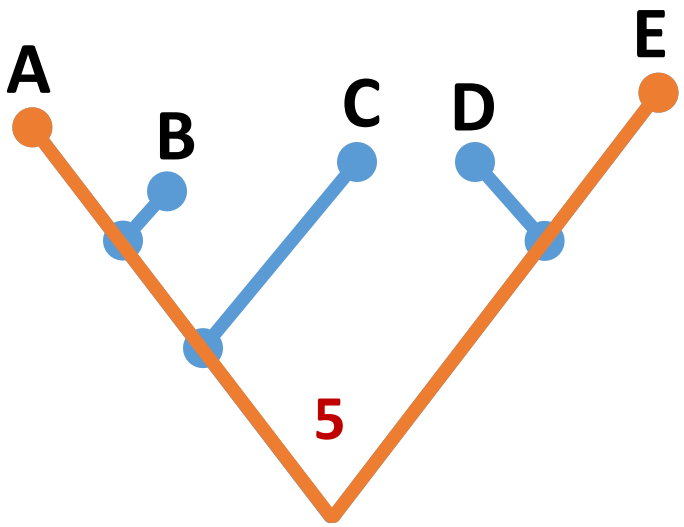
	123456789
Taxon A	AAGTGATTC
Taxon B	AACTGATTC
Taxon C	AAGCGTTTG
Taxon D	ATGTGACAG
Taxon E	ATCCGATAG

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0



How to measure a distance?

	123456789
Taxon A	A**AGT**GAT**TC**
Taxon B	AACTGATTC
Taxon C	AAGCGTTTG
Taxon D	ATGTGACAG
Taxon E	A**TCC**GAT**AG**

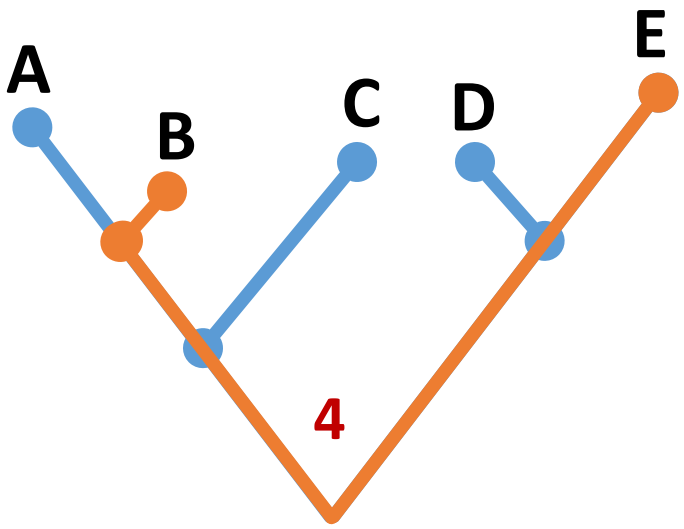


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

	123456789
Taxon A	AAGTGATTC
Taxon B	A**A**C**T**GAT**TC**
Taxon C	AAGCGTTTG
Taxon D	ATGTGACAG
Taxon E	A**T**C**C**GAT**AG**

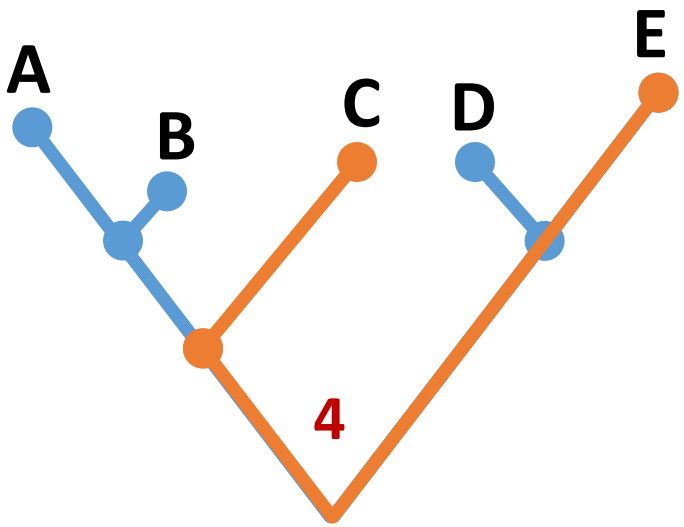


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

	123456789
Taxon A	AAGTGATTC
Taxon B	AACTGATTC
Taxon C	A**AG**CG**T**T**T**G
Taxon D	ATGTGACAG
Taxon E	A**TC**CG**A**T**A**G

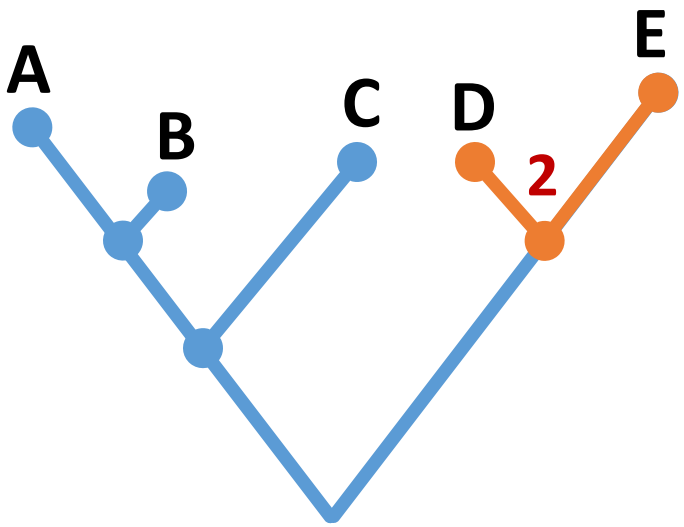


Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0

How to measure a distance?

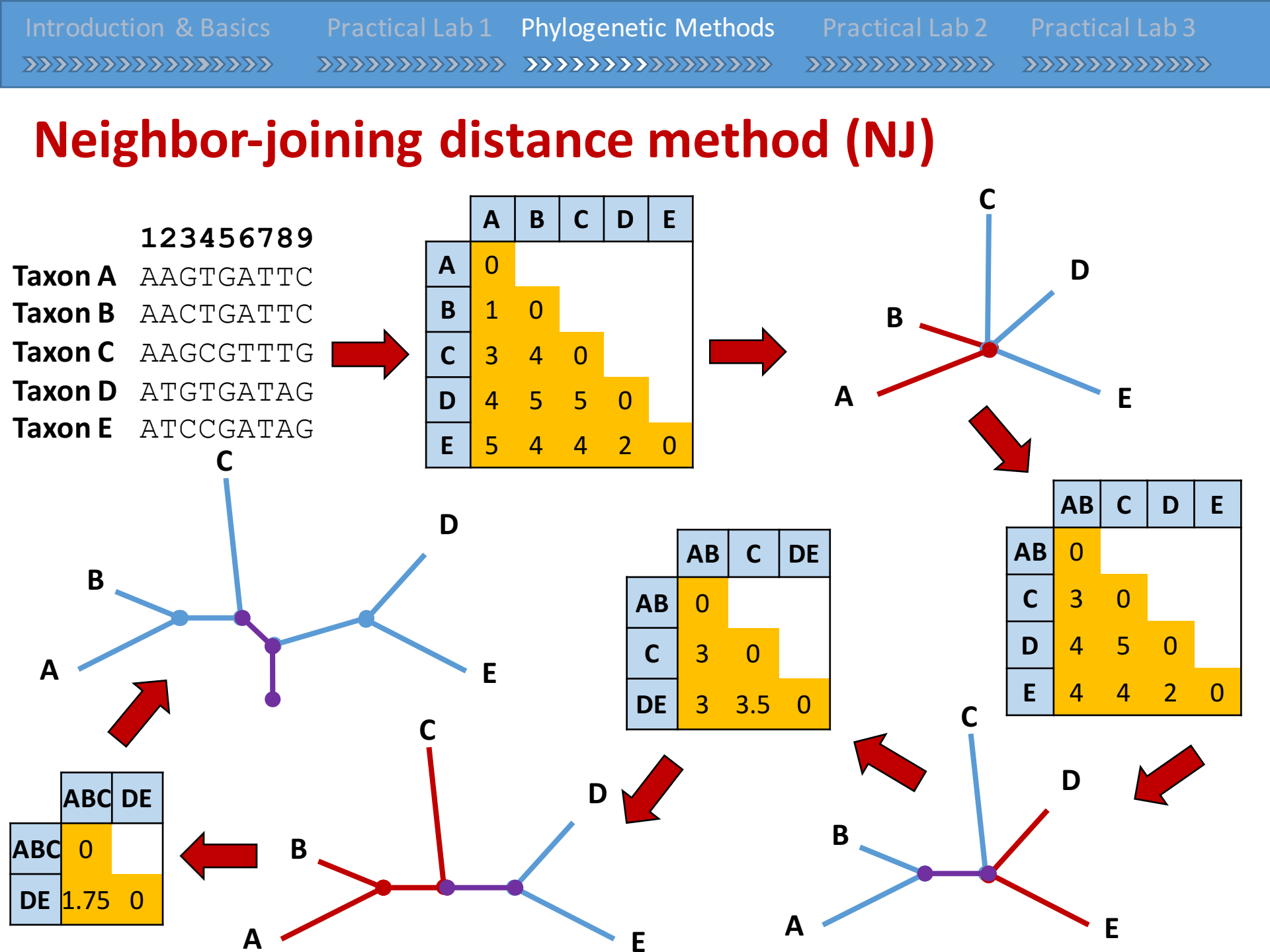
	123456789
Taxon A	AAGTGATTC
Taxon B	AACTGATTC
Taxon C	AAGCGTTTG
Taxon D	AT**GT**GATAG
Taxon E	AT**CC**GATAG



Number of substitutions between pairs of taxa

	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	2	0

Neighbor-joining distance method (NJ)



Minimum Evolution distance method (ME)

- Uses the optimally criterion based on the total branch length of the reconstructed tree

$$Q_s = \sum_{k}^{T} |v_k|$$

- Branch lengths are determined using unweighted LS method
- Among all topologies, chooses one with the lowest total branch length

Distance-matrix methods: summary

Pros:

- Consistent if the appropriate evolutionary model is chosen
- Very fast computationally – easily handles thousands of OTUs and molecular markers

Cons:

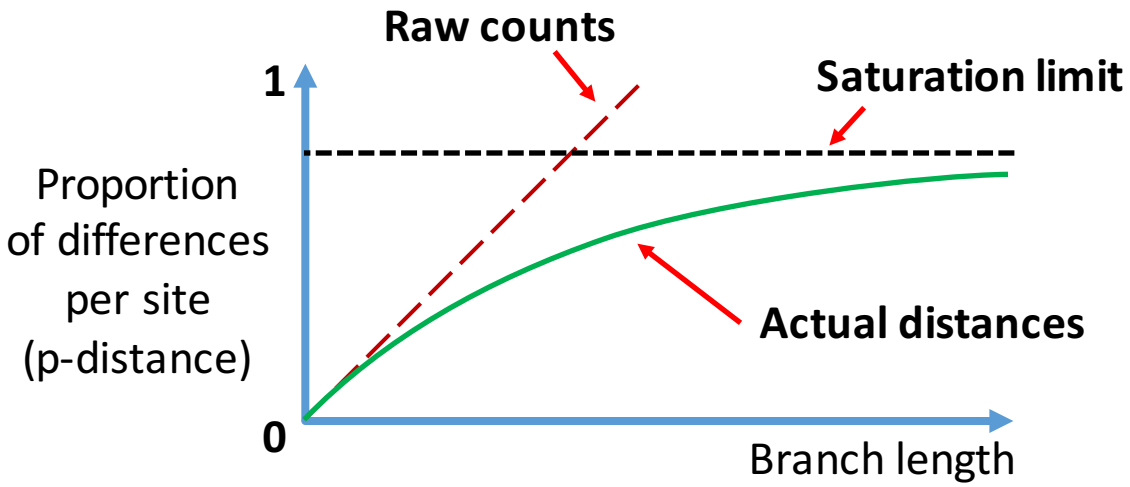
- Operates with distances rather than with actual characters
- Much information is lost because it uses only pairwise distances between tips and not the internal nodes
- Can produce misleading and inconsistent results
- Many assumptions are unrealistic (e.g., all nucleotide substitutions are independent, constant rate over time, etc.)

How to measure a distance?

Number of substitutions between pairs of taxa

	123456789
Taxon A	AAGTGATTC
Taxon B	AACTGATTC
Taxon C	AAGCGTTTG
Taxon D	ATGTGACAG
Taxon E	ATCCGATAG

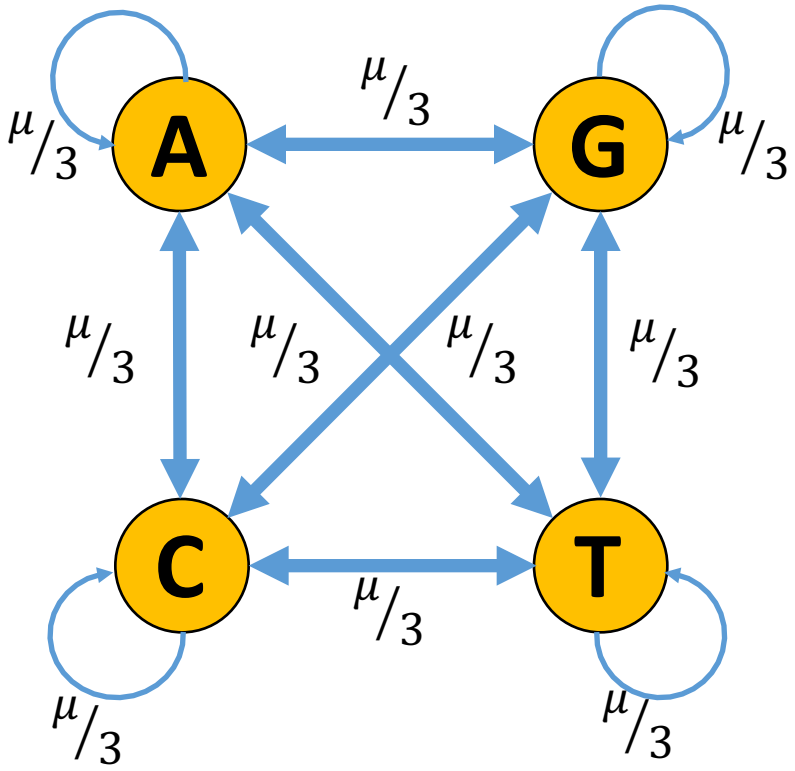
	A	B	C	D	E
A	0				
B	1	0			
C	3	4	0		
D	4	5	5	0	
E	5	4	4	3	0



Raw counts underestimate large distances!

Evolutionary models are needed to correct distances

Jukes-Cantor Model (JC69)



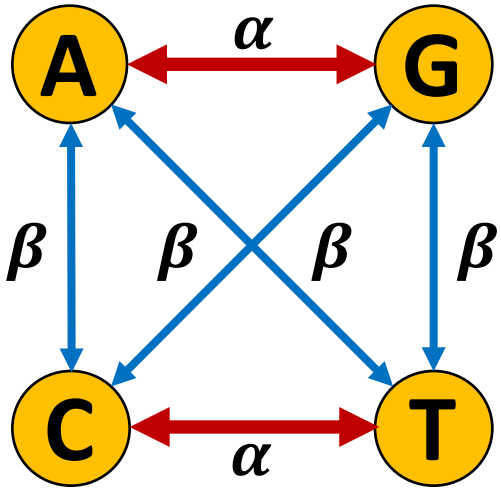
- Gives the same probability for any substitution at each locus, including change to itself: $\frac{4}{3}\mu$ (μ = mutation rate)
- Probability of at least one substitution per unit of time: $1 - e^{-\frac{4}{3}\mu t}$
- Probability of substitution to any other nucleotide (expected distance):

$$D_{obs} = \frac{3}{4}(1 - e^{-\frac{4}{3}\mu t})$$

- Therefore, the corrected distance:

$$D_{true} = -\frac{3}{4} ln(1 - \frac{4}{3} D_{obs})$$

Evolutionary models: parameters

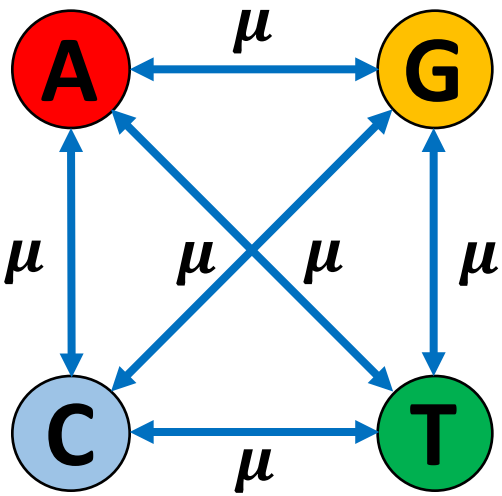


Kimura Model (K80)

α - transition rate

β - transversion rate

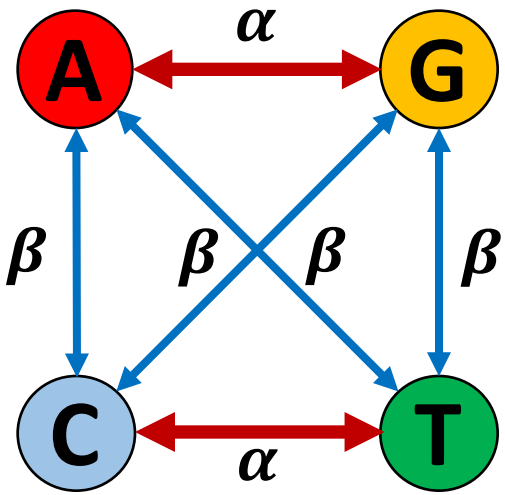
$$K = \frac{\alpha}{\beta}$$



Felsenstein Model (F81)

Base frequencies:

$\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$



Hasegawa-Kishino-Yano Model (HKY85)

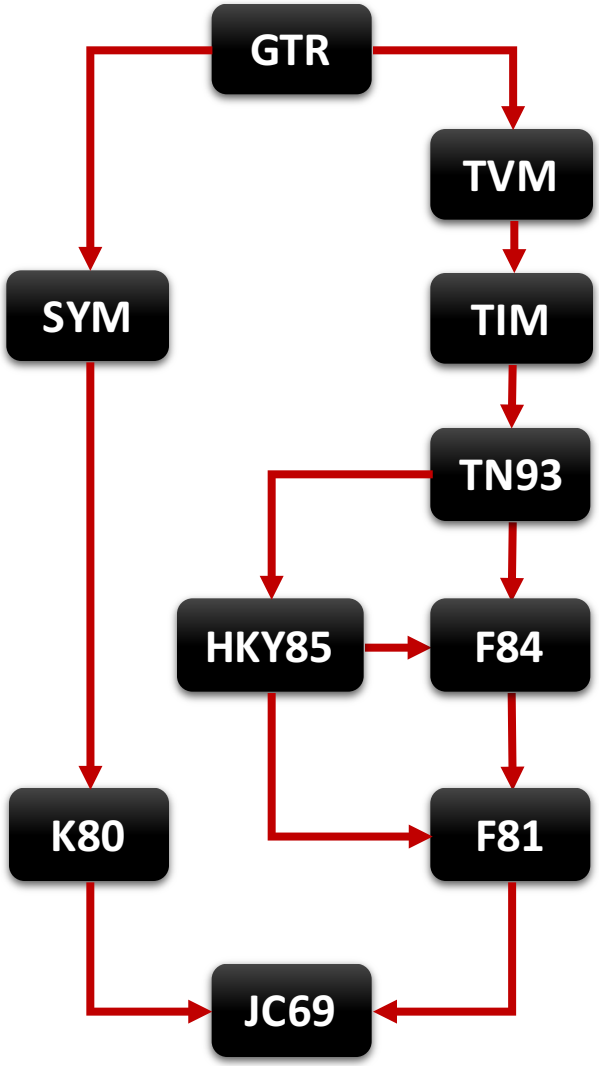
Base frequencies:

$\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$

α - transition rate

β - transversion rate

Evolutionary (substitution) models: hierarchy



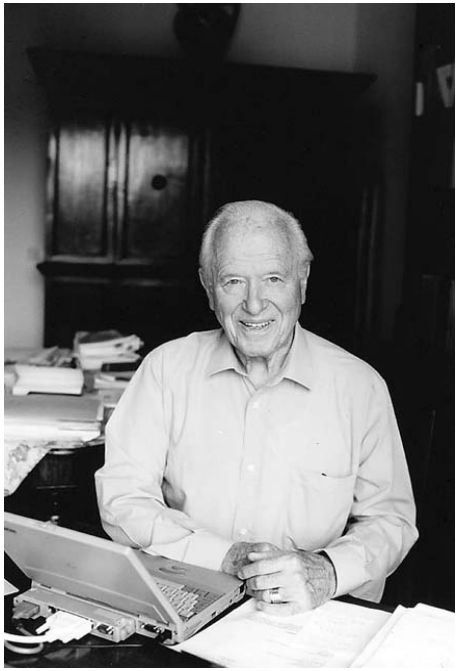
Model	Rates (tr + tv)	Base Frequencies	# of free parameters
JC69	equal	equal	0
K80	1+1	equal	1
SYM	variable	equal	5
F81	equal	variable	3
F84	2+1	$\pi_A + \pi_G$ $\pi_C + \pi_T$	4
HKY85	1+1	variable	4
TN93	2+1	variable	5
TIM	2+2	variable	6
TVM	1+4	variable	7
GTR	variable	variable	8

Maximum likelihood methods (ML)

Use statistical approach to estimate a tree Likelihood

Edwards & Cavalli-Sforza, 1964

Anthony Edwards
(b. 1935)



Luca Cavalli-Sforza
(b. 1922)

$$P(DATA \mid HYPOTHESIS)$$

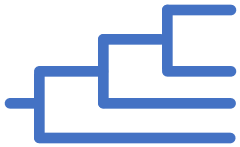
MAXIMIZATION

Image sources: royalsociety.org, www.companhiadasletras.com.br

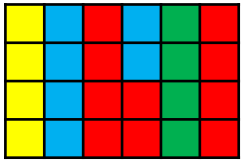
Model-based approach

θ

Evolutionary Model Parameters
(e.g., GTR+G)



Tree Topology & Branch Lengths



Multiple Sequence Alignment

Model-based approach:

Nucleotide substitution matrix (Q)

General Time-Reversible model (GTR)

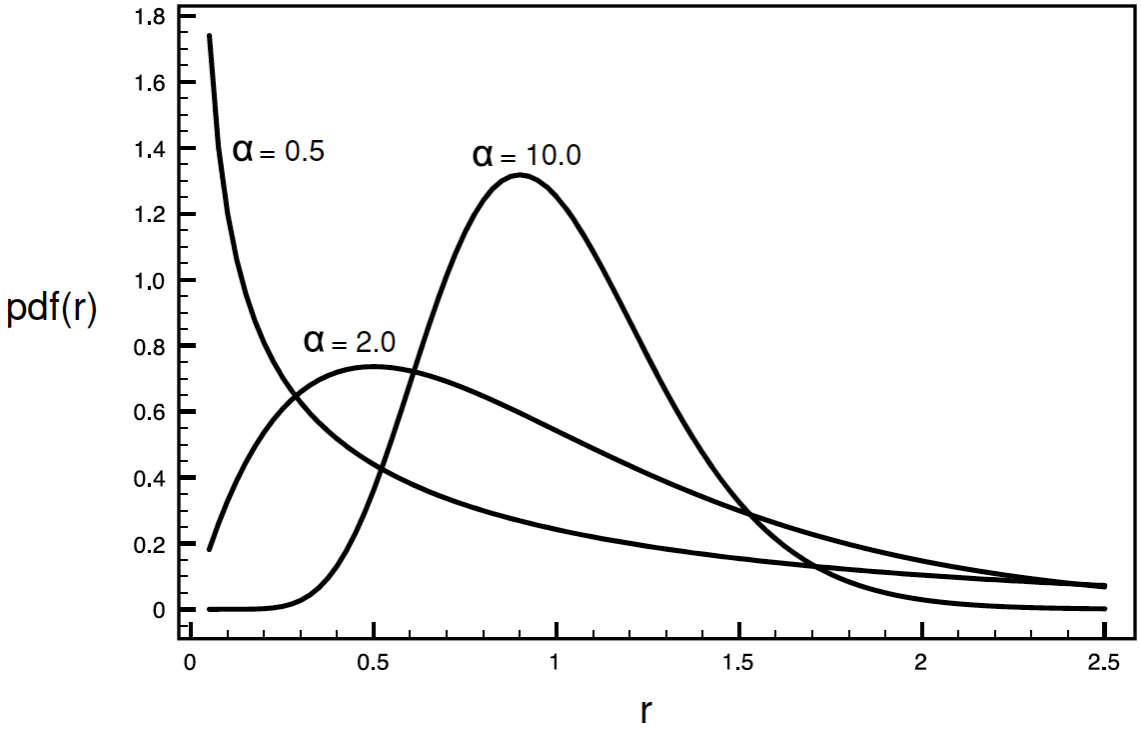
from\to	A	G	C	T
A	$-(\alpha\pi_G + \beta\pi_C + \gamma\pi_T)$	$\alpha\pi_G$	$\beta\pi_C$	$\gamma\pi_T$
G	$\alpha\pi_A$	$-(\alpha\pi_A + \delta\pi_C + \varepsilon\pi_T)$	$\delta\pi_C$	$\varepsilon\pi_T$
C	$\beta\pi_A$	$\delta\pi_G$	$-(\beta\pi_A + \delta\pi_G + \nu\pi_T)$	$\eta\pi_T$
T	$\gamma\pi_A$	$\varepsilon\pi_G$	$\eta\pi_C$	$-(\gamma\pi_A + \varepsilon\pi_G + \nu\pi_C)$

Optimal branch length estimate: $\boldsymbol{P(\nu) = e^{Q\nu}}$

Model-based approach:

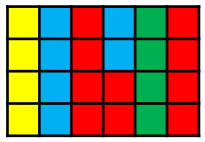
Rate heterogeneity among sites (gamma-distribution)

Γ- distributions with different α **parameter**



Rates of nucleotide substitutions vary among sites (e.g., 3rd codon position)

The Likelihood Function

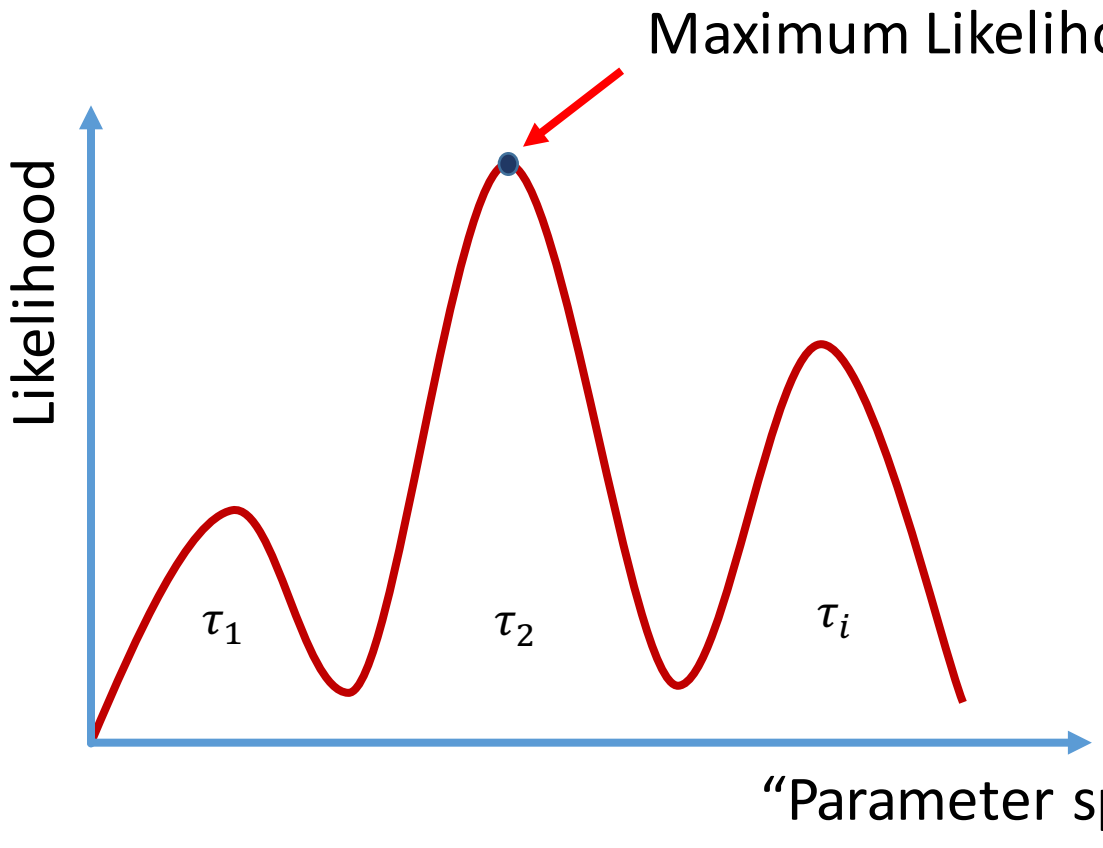
$P($  $|\theta,$  $) =$

The probability of data given model parameters and a tree

Assumptions:

- Evolution in different sites is independent
- Evolution in different lineages is independent

Maximum Likelihood scores are relative!



τ_i – tree topology

v_{ik} – branch length

Q_{M_i} – Q matrix for evolutionary model *M* for each tree *i*

Maximum Likelihood for Model Selection

Nucleotides in a single site of sequence alignment:

AAAAATTCCC

Model 1 (JC69): $P_{(A,C,G,T)} = \frac{1}{4}$

Model 2: $P_{(A,C,G,T)} = \frac{1}{2}; \frac{1}{5}; \frac{1}{5}; \frac{1}{10}$

$$L_{M1} = \left(\frac{1}{4}\right)^5 \left(\frac{1}{4}\right)^3 \left(\frac{1}{4}\right)^0 \left(\frac{1}{4}\right)^2 = \left(\frac{1}{4}\right)^{10} = 9.54\times10^{-7}$$

$$L_{M2} = \left(\frac{1}{2}\right)^5 \left(\frac{1}{5}\right)^3 \left(\frac{1}{5}\right)^0 \left(\frac{1}{10}\right)^2 = 2\times10^{-5}$$

$$\boldsymbol{L_{M2} > L_{M1}}$$

Criteria for Model Selection

- Likelihood Ratio Test: $LRT = 2(l_1 - l_0)$
- Akaike Information Criterion: $AIC = -2l + 2K$
- Corrected AIC: $cAIC = AIC + \frac{2K(K+1)}{n-K-1}$
- Bayesian Information Criterion: $BIC = -2l + K \log n$
- Decision Theoretical Framework (*DT*)

Maximum Likelihood: summary

Pros:

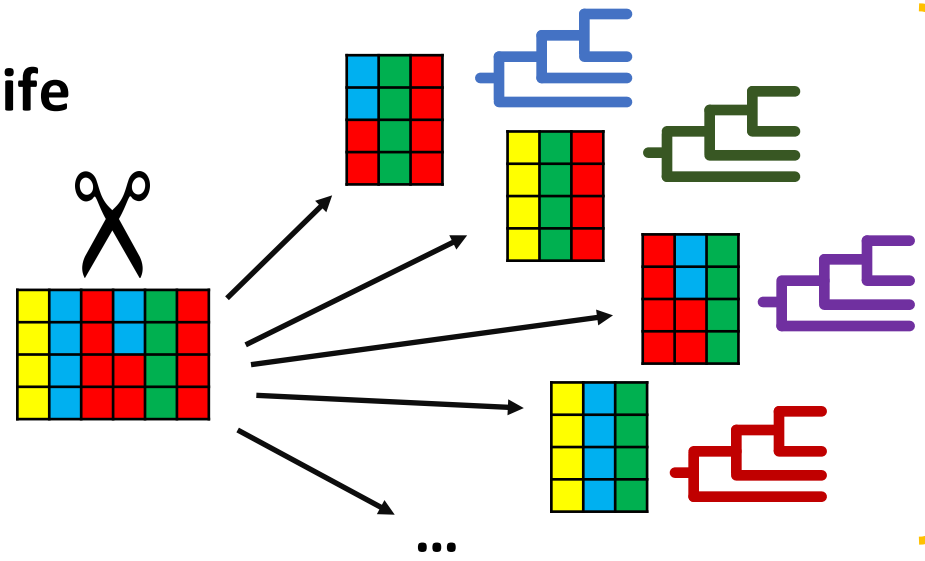
- Statistical approach: each site has likelihood
- Very accurate on branch lengths and tree topology
- Accommodates large number of evolutionary parameters

Cons:

- Provides a single point estimate (Maximum Likelihood)
- The correct model must be estimated
- Also sensitive to long-branches
- Very slow: impossible to implement if # of OTUs >> 10 and therefore often comes in combination with other methods (e.g., MP)

How to estimate tree confidence?

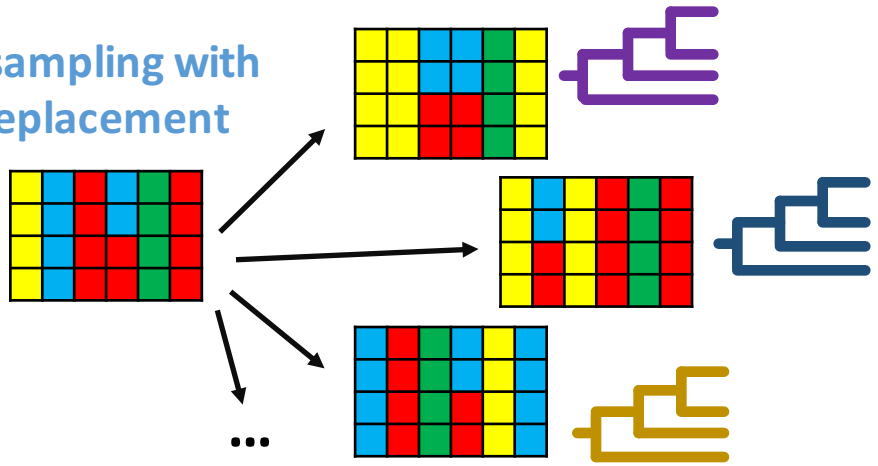
- **Jackknife**



MP, NJ or ML tree

- **Bootstrap (non-parametric)**

resampling with replacement



Count occurrences of nodes/branches on the consensus or on the original tree

Support:

>90% – high

75-90% – medium

<75% – low

Bayesian Approach for Phylogenetic Inference

What if we want to estimate phylogenetic hypothesis given the data?

Fredrik Ronquist

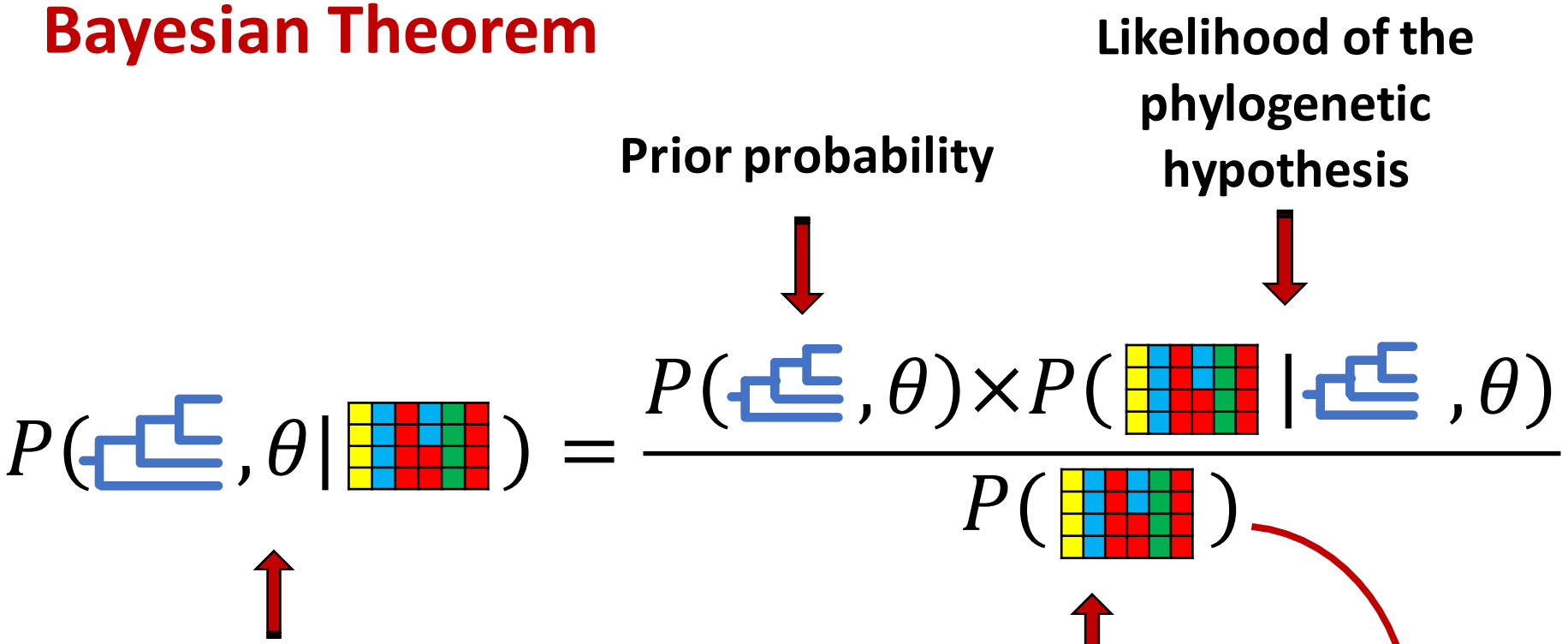
John Huelsenbeck

$$P(\text{tree} \mid \text{data}) =$$

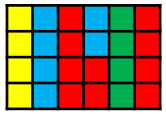
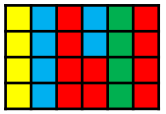
Posterior probability:
the probability of a particular tree conditional on a sequence alignment

Image source: royalsociety.org

Bayesian Theorem



Marginal probability approximation

$$P(\text{  }) = \int P(\text{  }, \boldsymbol{\theta}) * P(\text{  } \mid \text{  }, \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

- Thus, the integration over all tree topologies and model parameters is required
- **Markov Chain Monte Carlo (MCMC)** approach is used to search over the parameter space with stationary state distribution

Markov Chain Monte Carlo

1. Start at an arbitrary point
2. Make a small random move
3. Calculate the posterior density ratio (r) of new state to old state:
   - accept the new state if r > 1
   - stay in the old state if r < 1
4. Go to step 2

REPEAT millions of times

When calculating the posterior density ratio, the marginal probability of data cancels!!!

☺

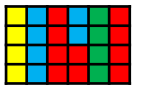
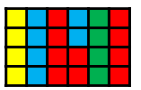
Metropolis Coupling (MCMCMC or MC^3)

implements search with multiple Markov chains of different 'temperatures':

- sampling of trees and posterior probabilities for all its parameters is done only by **a cold chain**;
- Some portion of the first sampled trees should be dismissed (**burn-in**);
- **heated chains** raise the ratio of posterior densities to $\mathbf{1} - \boldsymbol{temp}$ when deciding whether to accept a move (more flexible);
- when heated chain occasionally finds better place in the parameter space, it swaps with the cold chain allowing the latter to search in that region.

Prior probability

- is the probability of phylogenetic hypothesis before observing the actual data;
- Subjective but flexible;
- Always try to compare the results obtained with different settings for priors :

$$Bayes\ Factor = \frac{P(\text{  } \mid \text{  }, \theta_1)}{P(\text{  } \mid \text{  }, \theta_2)}$$

Ratio of marginal likelihoods

Bayesian approach: summary

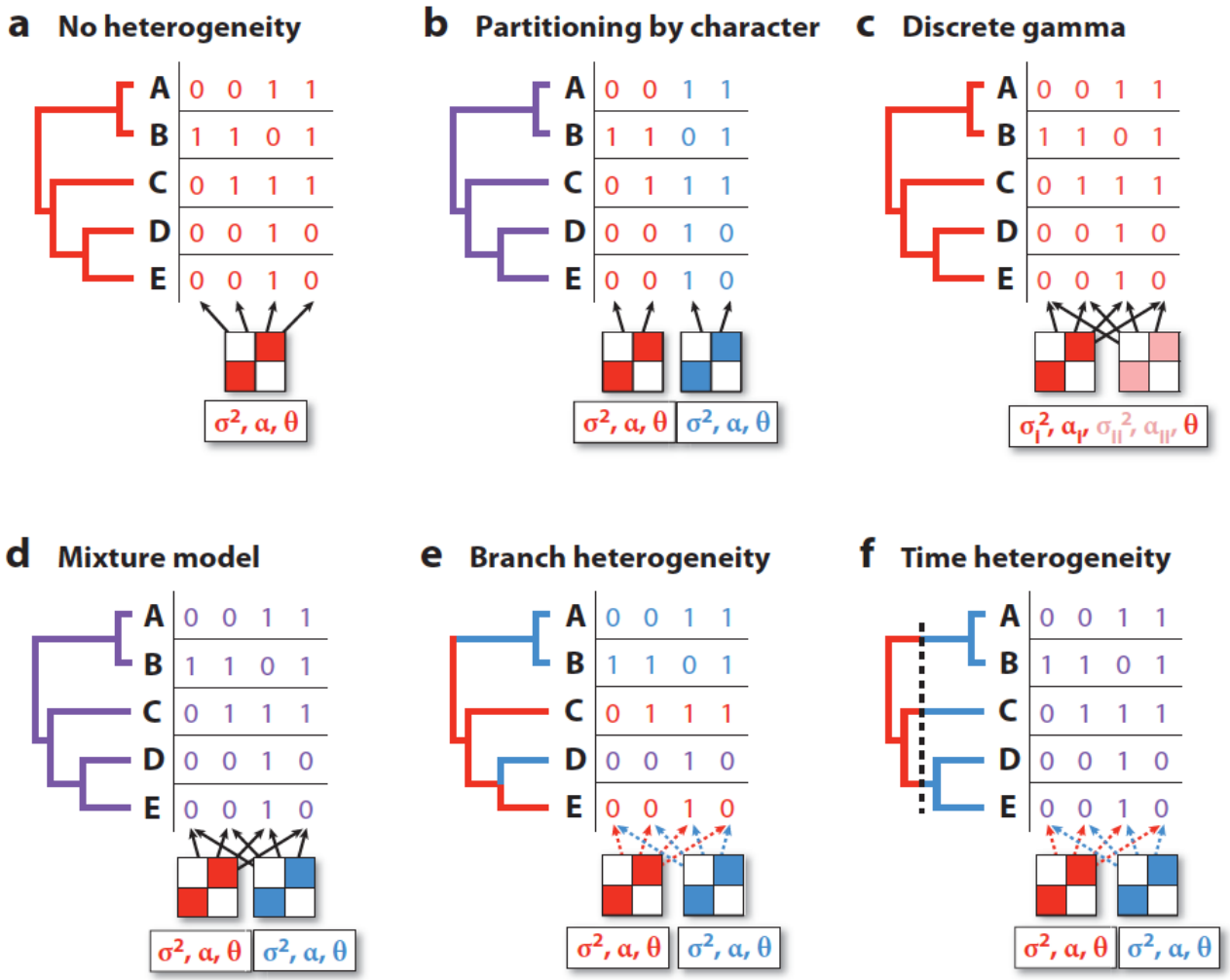
Pros:

- Statistical approach
- Fast!
- Estimates many different parameters
- Provides posterior probability distributions
- Very accurate on branch lengths and tree topology
- Provides its own measure of confidence instead of bootstrap

Cons:

- Very sensitive to your choice of prior distributions
- Sometimes may not converge well

Dealing with tree heterogeneity:



O'Meara, 2012