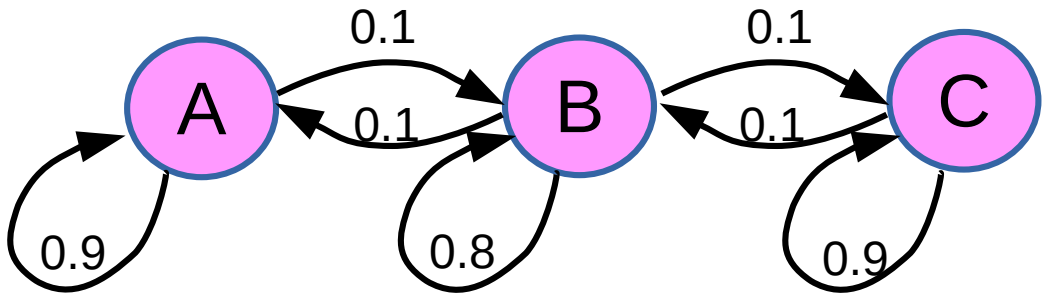
Exercises

1) What is the transition rate matrix for the following discrete time Markov Chain?



	A	B	C
A	.9	.1	0
B	.1	.8	.1
C	0	.1	.9

2) What is the probability of X_2=A, after two steps in a discrete time Markov Chain illustrated above, given that X_0=A? What is P(X_3=A|X_0=A)? 0.1*.1+.9*.9=0.82;

.1*.8*.1 + .1*.1*.9 + .9*.1*.1 + .9*.9*.9=0.755

ABA
AAA

ABBA
ABAA
AABA
AAAA

3) Why do we use a model with memory for nucleotide substitution? kappa and pi, ie different rates

Todays objectives

- Introduction to phylogenetics
- Phylogenetic tree reconstruction (UPGMA, NJ, Parsimony, Likelihood)
- Molecular clock and relative rates test
- Pruning algorithm, ancestral states
- MCMC

Phylogenetics

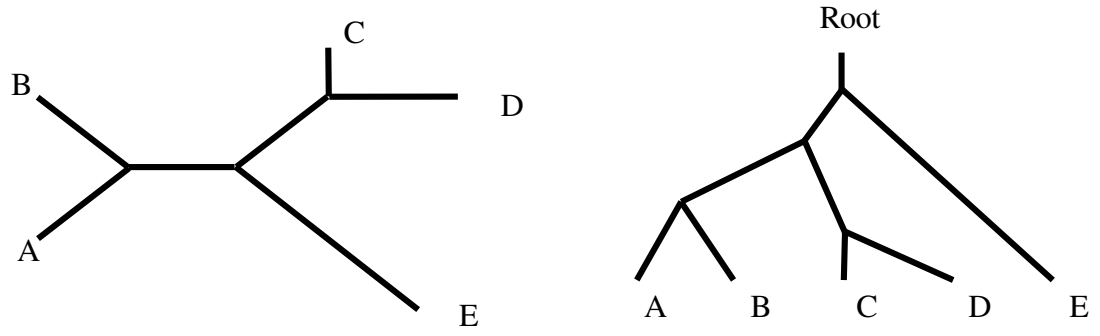
Phylogenetics - the reconstruction of the evolution history of genes or species.

Phylogenetic tree - a graphical representation of organisms evolutionary relationships.

Topology - the branching pattern in a phylogenetic tree.

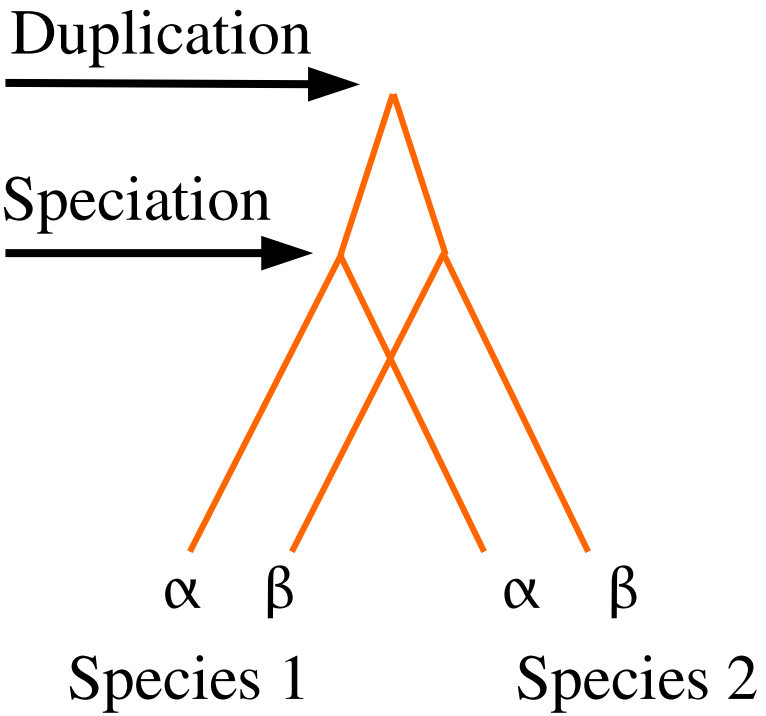
Root - a common ancestor to all taxa.

Rooted and unrooted phylogenetic trees



Ortholog, Paralog, Homolog

Orthologs are genes created by speciation events.
Paralogs are genes created by duplication events.
Homologs are genes that are similar because of shared ancestry.



Orthologues and paralogues can be distinguished by
i) synteny (gene order)
ii) phylogeny

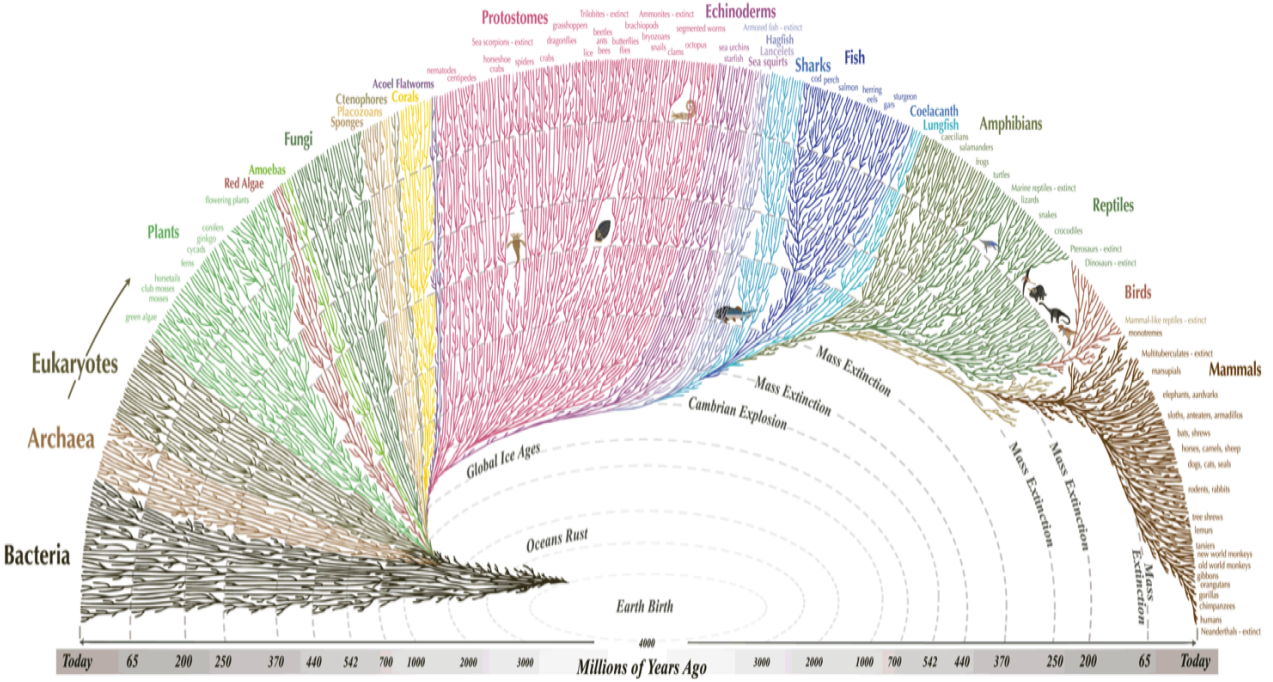
α-α orthologs
β-β orthologs
α-β paralogs

Homologs

The (exhaustive) tree problem

Table 1. Number of possible rooted and unrooted trees.

Number of sequences	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	954	105
10	34,459,425	2,027,025



Phylogenetics Methods

Three types of methods (debated which is best):

Distance: similarity based methods (also called phenetic methods), e.g. UPGMA and Neighbor-Joining. Generate a single tree, fast, heuristic

Parsimony: character state methods (also called cladistic methods). Scores trees, heuristic tree search

Maximum likelihood: parametric models of evolution. Scores trees, heuristic tree search

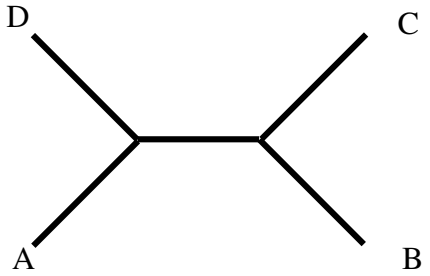
Parsimony and Maximum likelihood are both used when phylogeny is critical.

Distance method: UPGMA and Neighbor-Joining

Table 2. Distance matrix.

Sequence	A	B	C
A			
B	d(AB)		
C	d(AC)	d(BC)	
D	d(AD)	d(BD)	d(CD)

Each d is the distance (substitution rate) between pairs of sequences



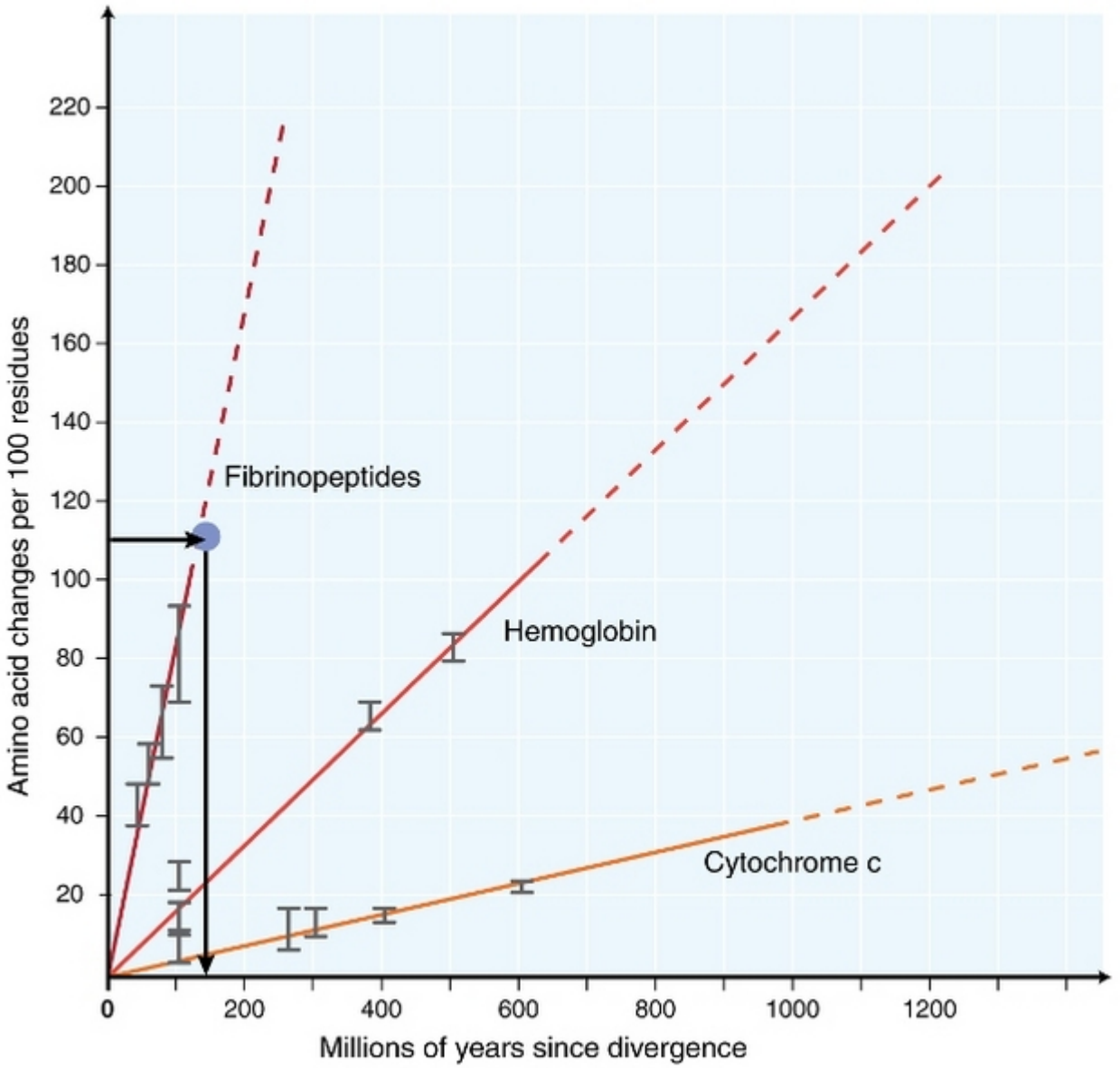
- Distance based methods use the distances between each pair of sequences to create a phylogenetic tree.
- Substitutions rates are the most commonly used measure for distance, but insertion/deletion rates or other characters can also be used.
- Methods are fast and generate a single tree

Distance Methods: UPGMA

UPGMA - unweighted pair-group method with arithmetic mean

- assumes a molecular clock, or constant-rate of evolution - that is the distances from the root to every branch tip are equal
- iteratively joins the two most similar sequences.
- a simple agglomerative (bottom-up) hierarchical clustering method, time complexity $O(n^3)$, memory $O(n^2)$
- if d(AB) is the smallest value in the matrix, the matrix is recalculated between AB, C and D. The distance between AB and C is the average between AC and BC.

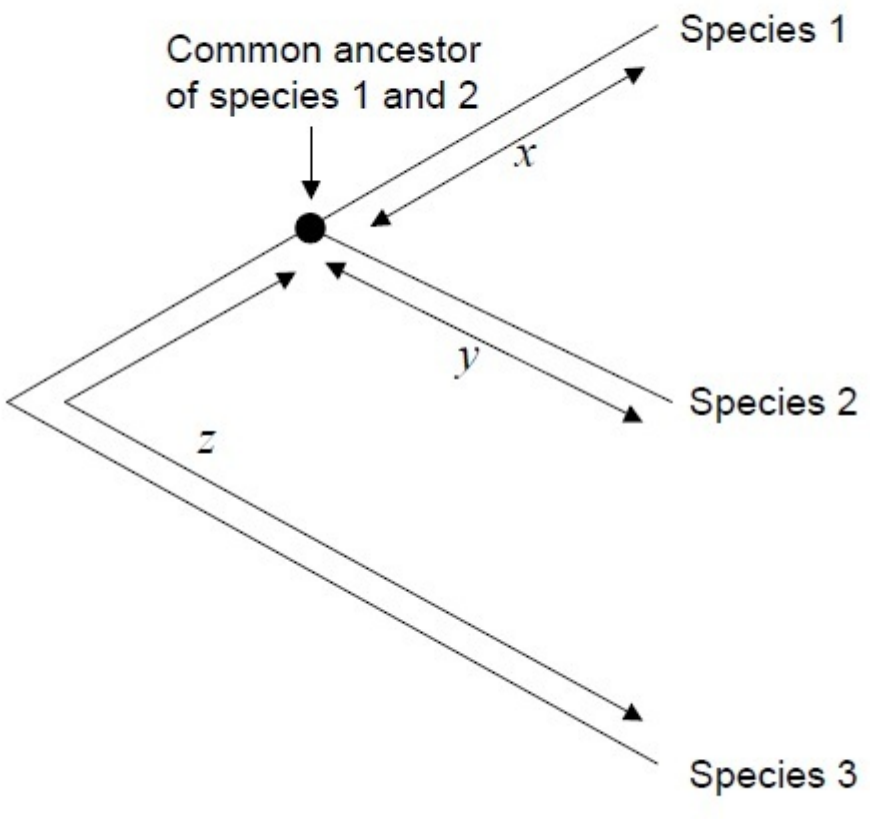
Molecular Clock



Zuckerkandl & Pauling (1965) likened the constant accumulation of amino acid substitutions over time to regular ‘ticks' of a clock. The term ‘molecular clock' was initially coined to describe changes in amino acids occurring in proportional to time since species divergence.

- Most proteins have a more or less constant rate
- The rate between proteins often differs

Relative Rates Test (testing a clock)



Clock hypothesis:

$x - y = 0$

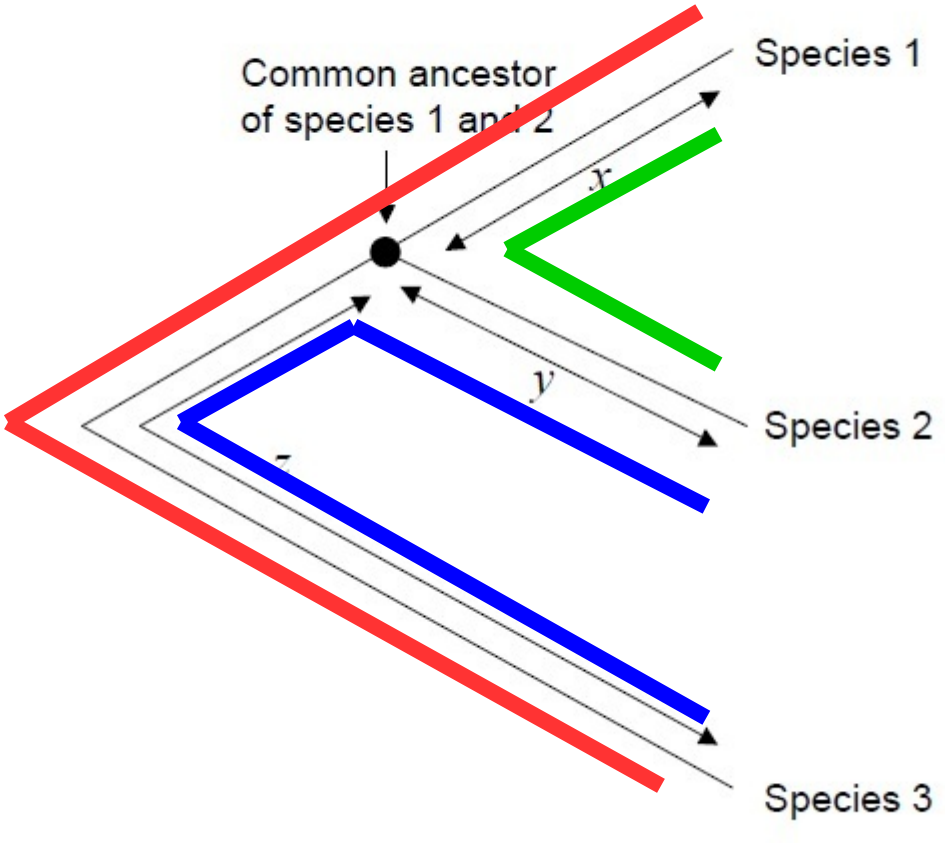
$x = (d13 + d12 - d23)/2$

$y = (d23 + d12 - d13)/2$

$z = d23 - y$

$z = d13 - x$

Relative Rates Test



Clock hypothesis:

$x - y = 0$

$x = (d13 + d12 - d23)/2$

$y = (d23 + d12 - d13)/2$

$z = d23 - y$

$z = d13 - x$

UPGMA

	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

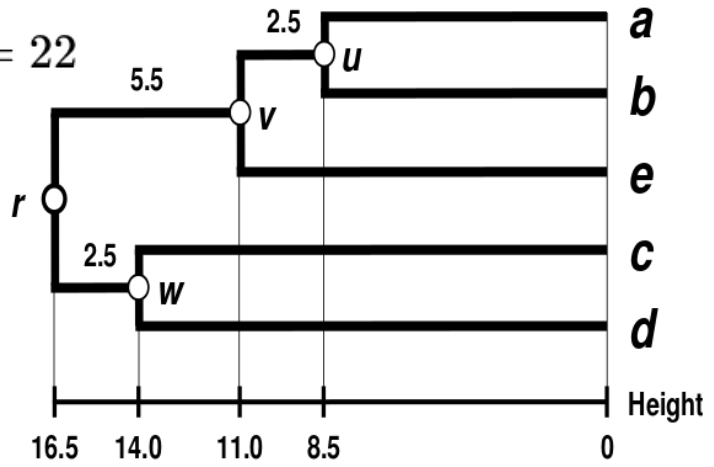
1) calculate distance matrix
2) join smallest value, calculate branch length as average distance
3) recalculate distance matrix
4) repeat

$$D_2((a,b),c) = (D_1(a,c) \times 1 + D_1(b,c) \times 1)/(1+1) = (21+30)/2 = 25.5$$

$$D_2((a,b),d) = (D_1(a,d) + D_1(b,d))/2 = (31+34)/2 = 32.5$$

$$D_2((a,b),e) = (D_1(a,e) + D_1(b,e))/2 = (23+21)/2 = 22$$

	(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



UPGMA

	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

1) calculate distance matrix
2) join smallest value, calculate branch length as average distance
3) recalculate distance matrix
4) repeat

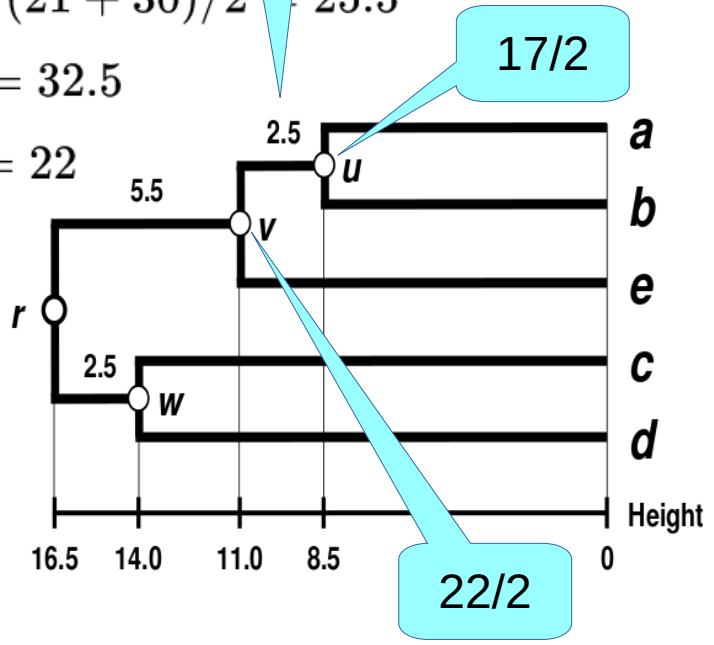
11-8.5

$$D_2((a,b),c) = (D_1(a,c) \times 1 + D_1(b,c) \times 1)/(1+1) = (21+30)/2 = 25.5$$

$$D_2((a,b),d) = (D_1(a,d) + D_1(b,d))/2 = (31+34)/2 = 32.5$$

$$D_2((a,b),e) = (D_1(a,e) + D_1(b,e))/2 = (23+21)/2 = 22$$

	(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



UPGMA

The distance between any two clusters A and B, with members x and y (respectively), is the average of all distances d(x,y) between pairs of objects x in A and y in B, that is, the mean distance between elements of each cluster:

$$d_{AB} = \frac{\sum_{A \in x} \sum_{B \in y} d(x, y)}{xy}$$

	(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

$$D_3(((a,b),e),c) = (D_2((a,b),c) \times 2 + D_2(e,c) \times 1)/(2+1) = (25.5 \times 2 + 39 \times 1)/3 = 30$$

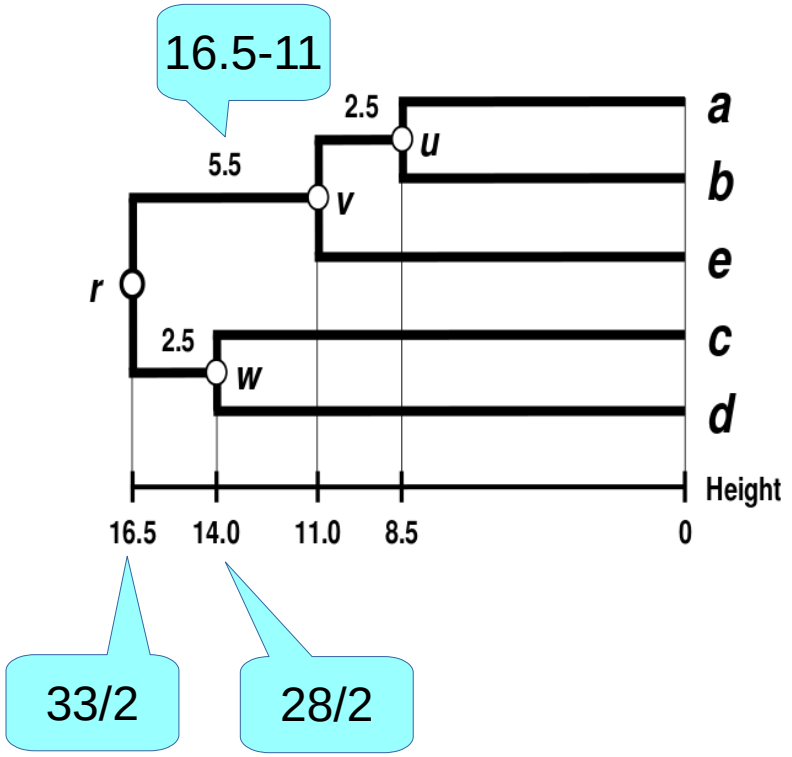
$$d_3 = (d_{AC} + d_{BC} + d_{EC})/3 = (21 + 30 + 39)/3 = 30$$

$$D_3(((a,b),e),d) = (D_2((a,b),d) \times 2 + D_2(e,d) \times 1)/(2+1) = (32.5 \times 2 + 43 \times 1)/3 = 36$$

UPGMA

	(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

	((a,b),e)	c	d
((a,b),e)	0	**30**	**36**
c	**30**	0	*28*
d	**36**	*28*	0



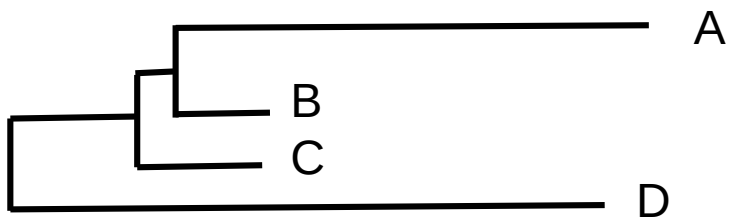
$$D_3(((a,b),e),c) = (D_2((a,b),c) \times 2 + D_2(e,c) \times 1)/(2+1) = (25.5 \times 2 + 39 \times 1)/3 = 30$$

$$D_3(((a,b),e),d) = (D_2((a,b),d) \times 2 + D_2(e,d) \times 1)/(2+1) = (32.5 \times 2 + 43 \times 1)/3 = 36$$

$$D_4((c,d),((a,b),e)) = (D_3(c,((a,b),e)) \times 1 + D_3(d,((a,b),e)) \times 1)/(1+1) = (30 \times 1 + 36 \times 1)/2 = 33$$

Problem: molecular clock is not always true

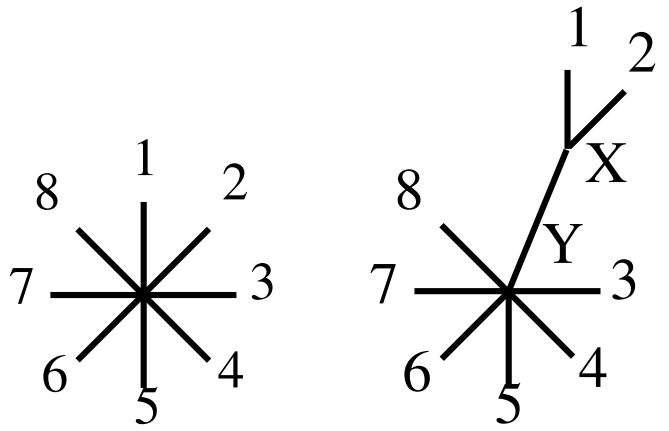
Which species would be joined first using UPGMA and the true tree below?



Distance between BC is the shortest.

Distance: Neighbor-joining

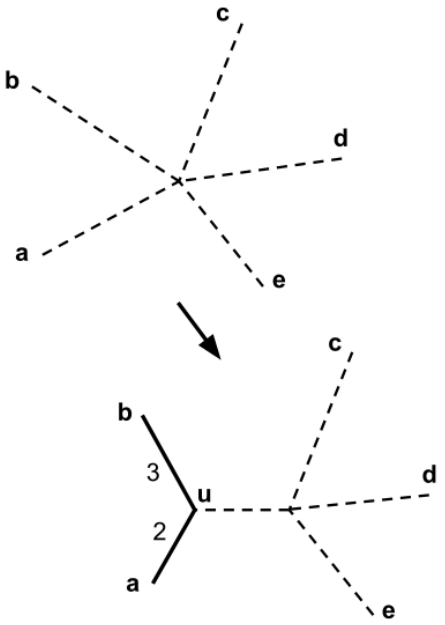
Neighbor-Joining - iteratively finds closest neighbors so as to minimize the total length of the tree (no molecular clock assumption). Starting with a star phylogeny, iteratively join the two sequences that minimize the sum of the branch lengths.



Tree on the right is smaller (total length of branches)

If 1 and 2 are closely related the sum of the branch lengths will be smaller than if they are distantly related.

Neighbor-Joining



	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

1. Calculate distance matrix
2. Find Q that minimizes tree
3. Calculate branch length
4. Repeat with updated matrix

n is the # taxa
d(a,b) is distance between a and b

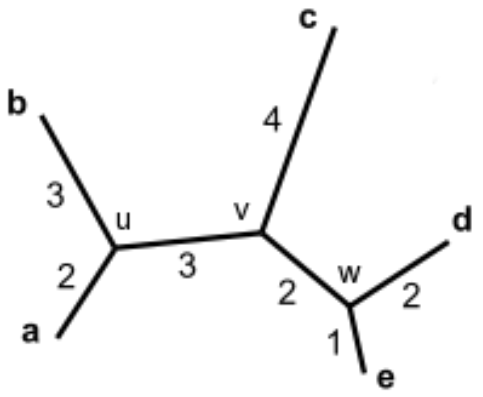
$$Q_1(a,b) = (n-2)d(a,b) - \sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)$$

$$= (5-2) \times 5 - (5+9+9+8) - (5+10+10+9) = 15 - 31 - 34 = -50$$

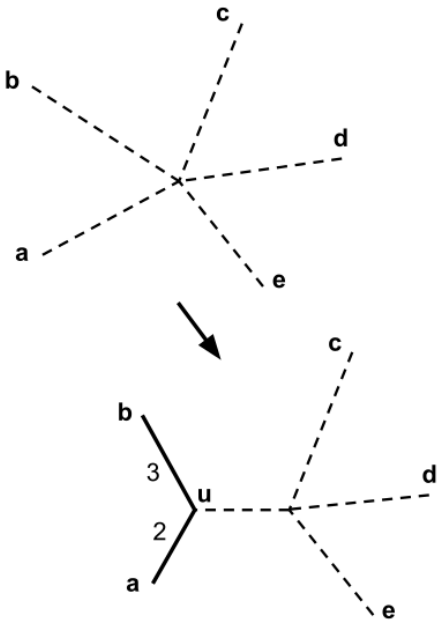
	u	c	d	e
u	0	**7**	**7**	**6**
c	**7**	0	*8*	*7*
d	**7**	*8*	0	*3*
e	**6**	*7*	*3*	0

	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	

Pick smallest value of Q



Neighbor-Joining



	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

1. Calculate distance matrix
2. Find Q that minimizes tree
3. Calculate branch length
4. Repeat with updated matrix

n is the # taxa
d(a,b) is distance between a and b

$$Q_1(a,b) = (n-2)d(a,b) - \sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)$$

$$= (5-2) \times 5 - (5+9+9+8) - (5+10+10+9) = 15 - 31 - 34 = -50$$

	a	b	c	d	e
a		−50	−38	−34	−34
b	−50		−38	−34	−34
c	−38	−38		−40	−40
d	−34	−34	−40		−48
e	−34	−34	−40	−48	

Pick smallest value of Q

	u	c	d	e
u	0	7	7	6
c	7	0	8	7
d	7	8	0	3
e	6	7	3	0

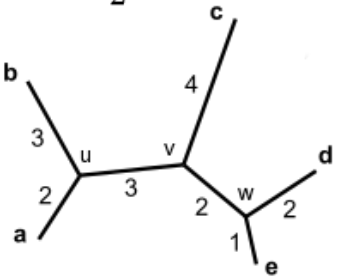
$$d(u,c) = \frac{1}{2}[d(a,c) + d(b,c) - d(a,b)] = \frac{9+10-5}{2} = 7$$

$$d(u,d) = \frac{1}{2}[d(a,d) + d(b,d) - d(a,b)] = \frac{9+10-5}{2} = 7$$

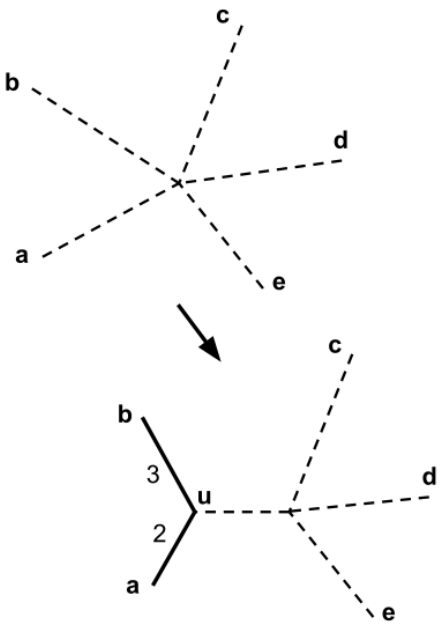
$$d(u,e) = \frac{1}{2}[d(a,e) + d(b,e) - d(a,b)] = \frac{8+9-5}{2} = 6$$

$$\delta(a,u) = \frac{1}{2}d(a,b) + \frac{1}{2(5-2)}\left[\sum_{k=1}^{5} d(a,k) - \sum_{k=1}^{5} d(b,k)\right] \quad = \frac{5}{2} + \frac{31-34}{6} = 2$$

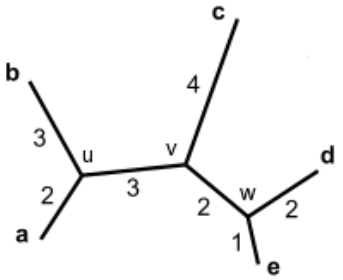
$$\delta(b,u) = d(a,b) - \delta(a,u) \quad = 5 - 2 = 3$$



Neighbor-Joining



1. Calculate distance matrix
2. Find Q that minimizes tree
3. Calculate branch length
4. Repeat with updated matrix



Neighbor-Joining:

- Generates a single tree, $O(n^3)$ time

- A greedy heuristic for minimum evolution

- Given an additive distance matrix as input, neighbor joining is guaranteed to find the tree whose distances between taxa agree with it. In practice, additivity is rare but not far from it.

How do we evaluate and compare different trees (hypotheses)?

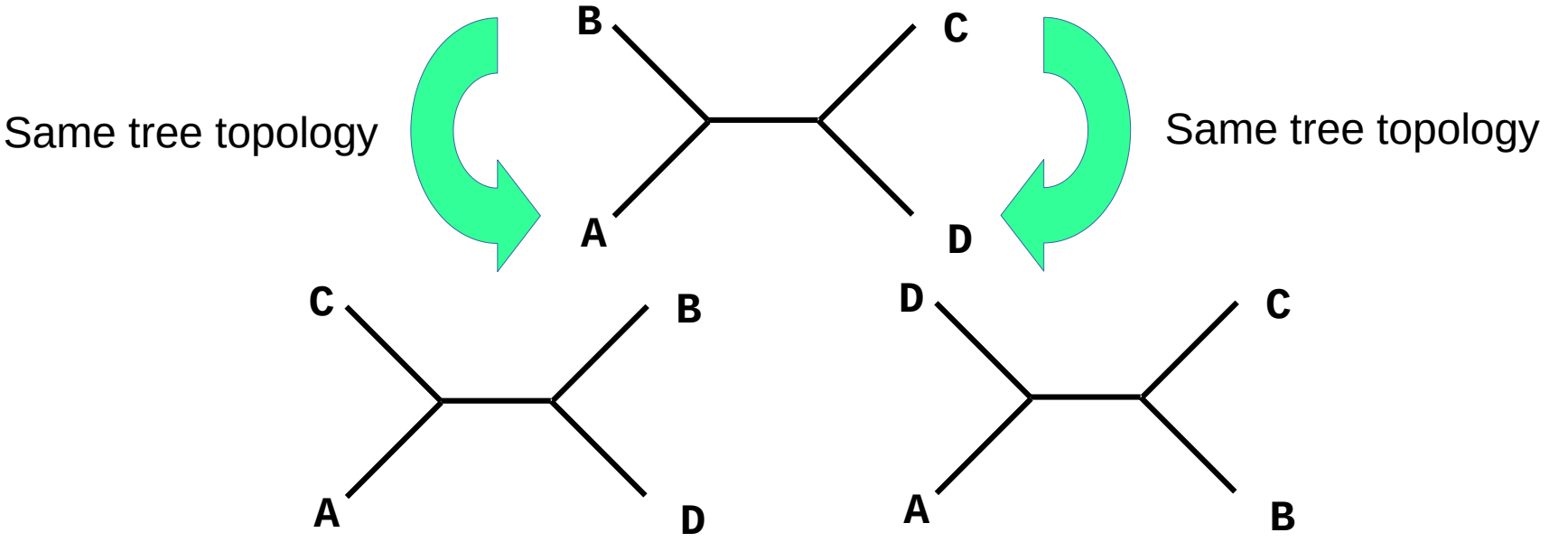
To compare trees we need a scoring method

Scoring methods:

- Maximum Parsimony
- Maximum Likelihood

Character state: Parsimony

Parsimony uses the minimum number of evolutionary changes to explain character states (nucleotides, peptides or morphological character states).

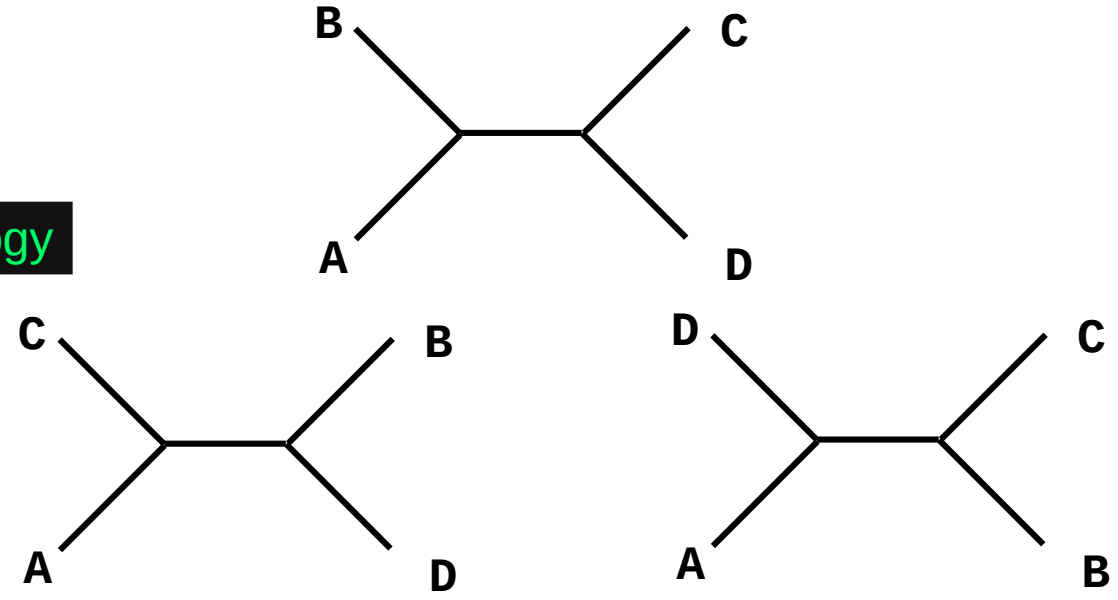


There are three unrooted tree topologies for four taxa.

Character state: Parsimony

Parsimony uses the minimum number of evolutionary changes to explain character states (nucleotides, peptides or morphological character states).

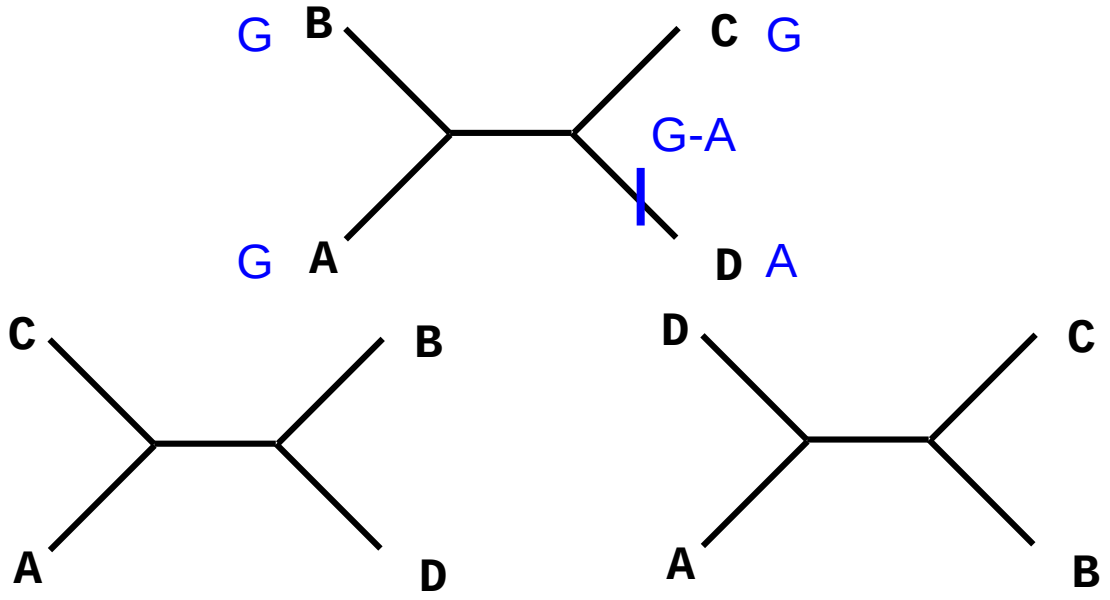
Different topology



There are three unrooted tree topologies for four taxa.

Character state: Parsimony

Parsimony uses the minimum number of evolutionary changes to explain character states (nucleotides, peptides or morphological character states).

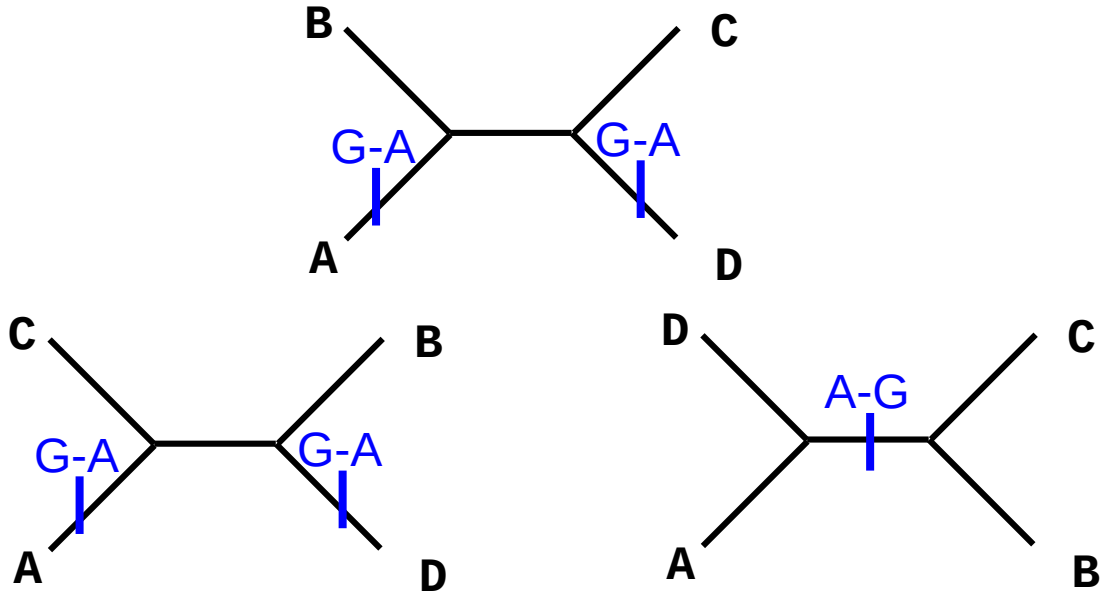


Sequence	Site 1	2	3	4
A	G	A	A	A
B	G	G	A	A
C	G	G	T	G
D	A	A	T	T

The maximum parsimony method searches for the best tree that can explain the data. Non-informative sites such as #1 are not used. This is different from distance based methods which use all the sites.

Character state: Parsimony

Parsimony uses the minimum number of evolutionary changes to explain character states (nucleotides, peptides or morphological character states).



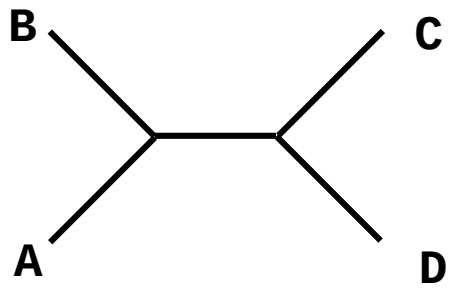
Sequence	Site 1	2	3	4
A	G	A	A	A
B	G	G	A	A
C	G	G	T	G
D	A	A	T	T

The maximum parsimony method searches for the best tree that can explain the data. Non-informative sites such as #1 are not used. This is different from distance based methods which use all the sites.

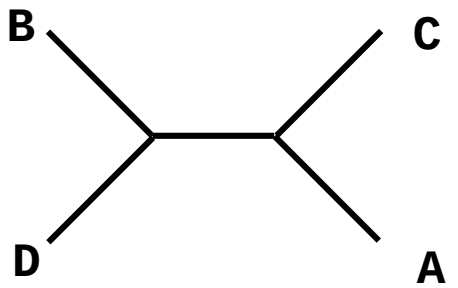
Exercises

1) Which trees are redundant, ie the same tree?

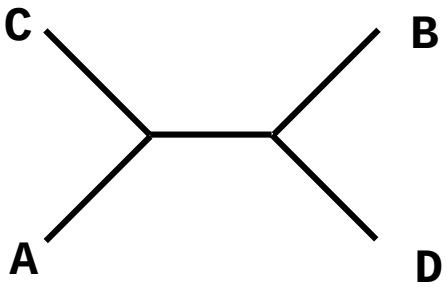
1)



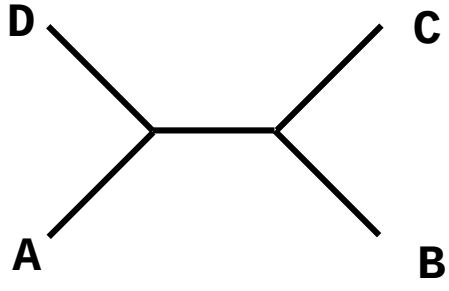
2)



3)



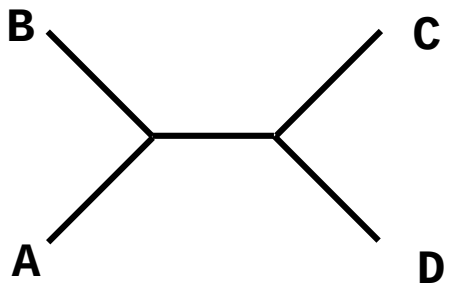
4)



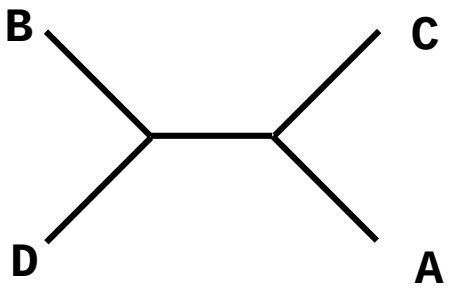
Exercises

1) Which trees are redundant, ie the same tree? 2 and 3

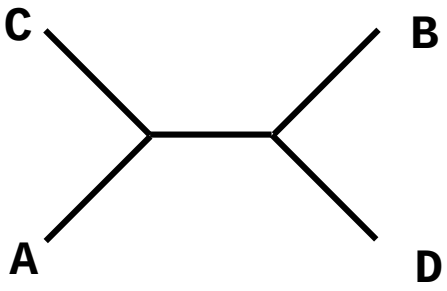
1)



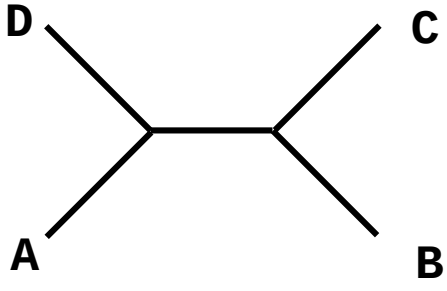
2)



3)



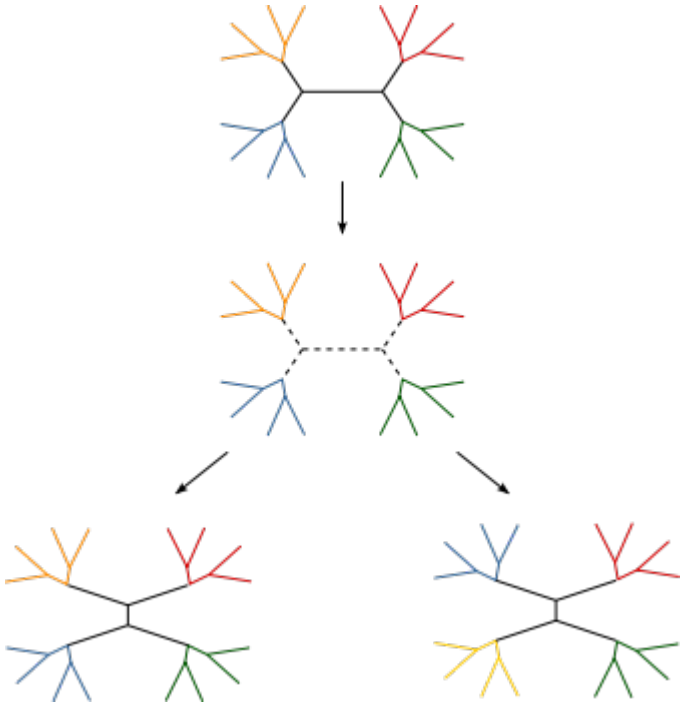
4)



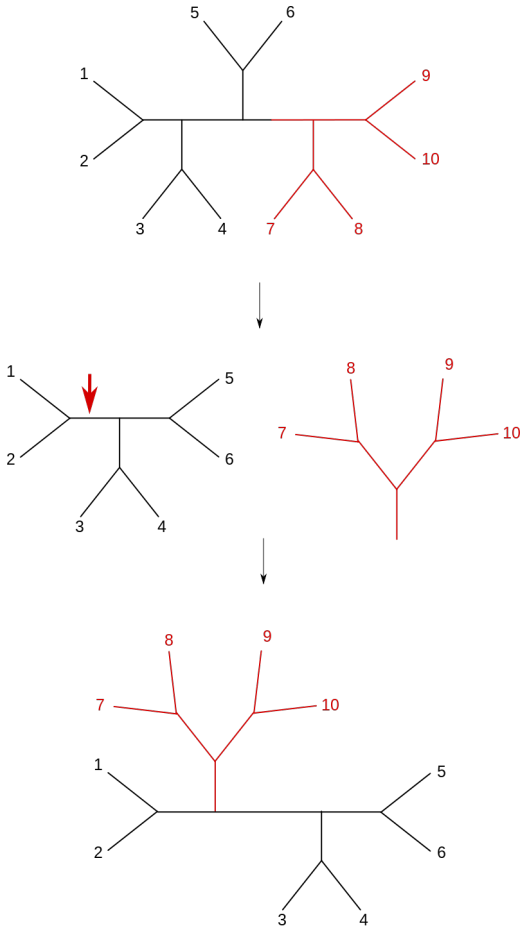
Maximum parsimony

- Maximum parsimony is an optimality criterion under which the phylogenetic tree is chosen that minimizes the total number of character-state changes
- Trees are scored (evaluated) by how many "steps" (evolutionary transitions) are required to explain the distribution of each character
- Because there are many trees, a number of algorithms are used to search among the possible trees, which are then evaluated by maximum parsimony
- Homoplasy (parallel, convergent evolution) are minimized

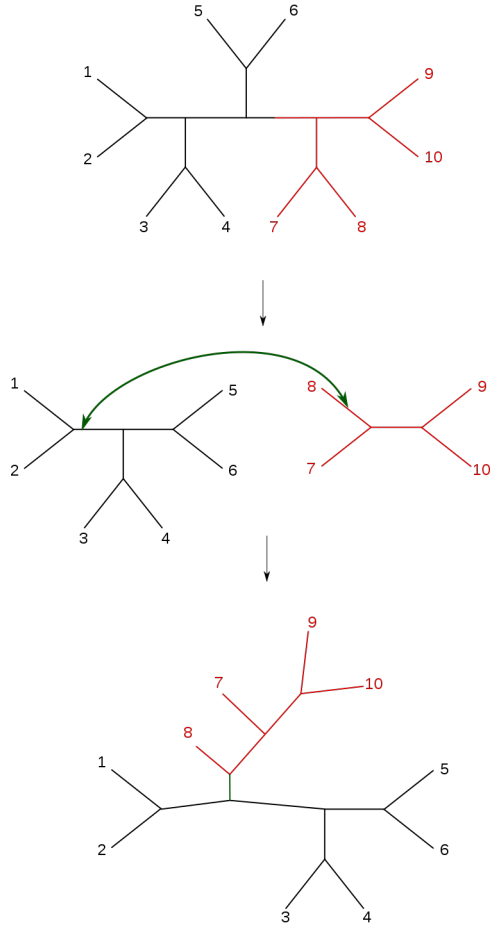
Tree search heuristics



Nearest Neighbor Interchange

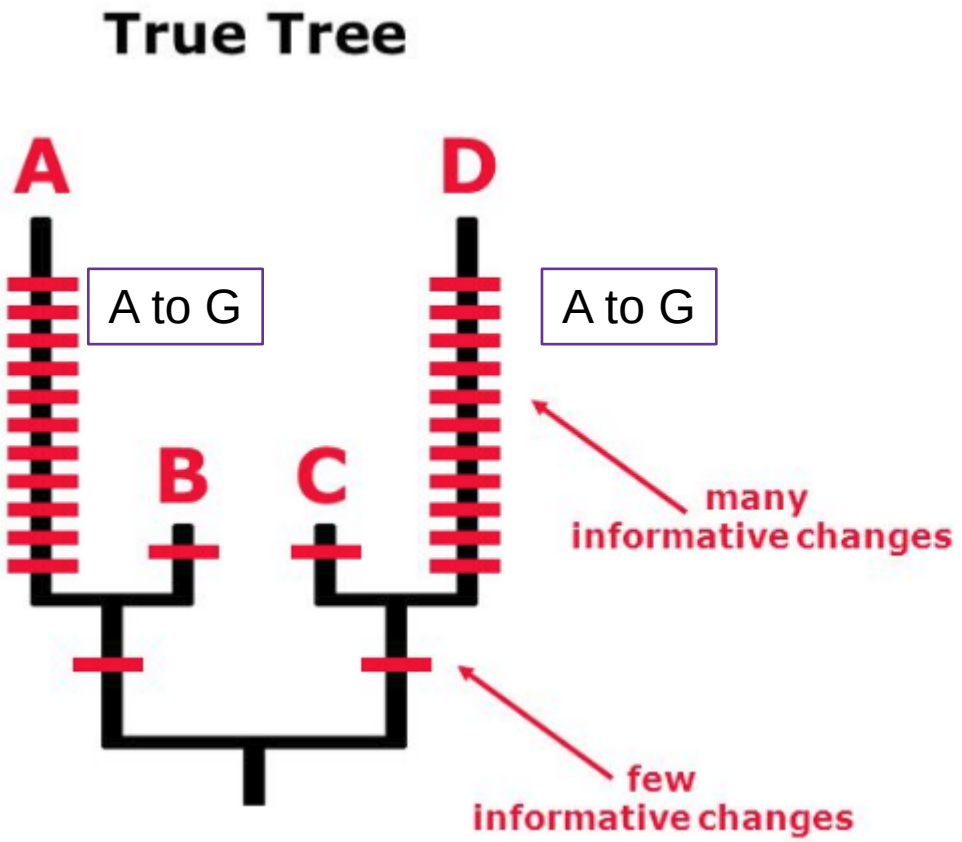


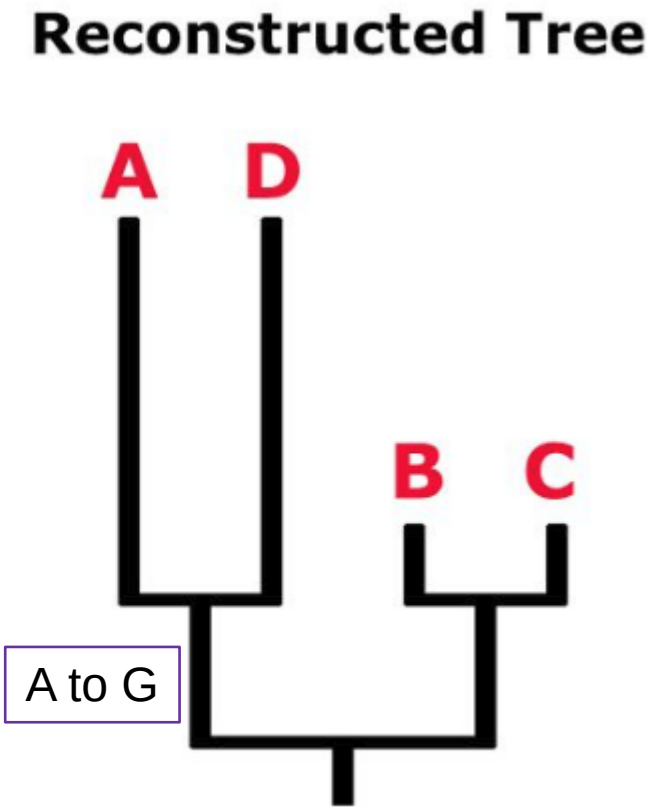
Subtree pruning and regrafting



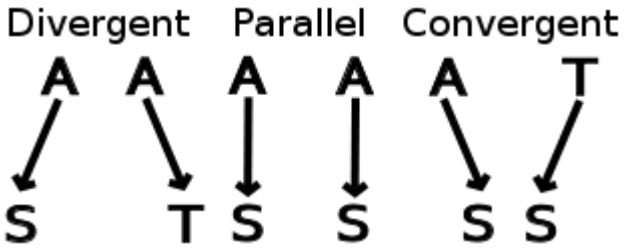
Tree bisect and reconnect

Parsimony Problem: Long branch attraction





Homoplasy is when a trait has been gained or lost independently in separate lineages over the course of evolution. Leads to similarity due multiple events rather than shared ancestry (homology).



Maximum Likelihood

Maximum Likelihood: Statistical inference of probabilities of the data given a phylogenetic tree and a nucleotide (or amino acid) substitution model.

- Homoplasy is accounted for using branch lengths

τ = tree topology
ν = branch lengths
Q = substitution model

$$L(\Theta) \propto P(X|\Theta)$$

$$L(\tau, \nu, Q) \propto P(X|\tau, \nu, Q)$$

OPTIMIZE to find maximum likelihood

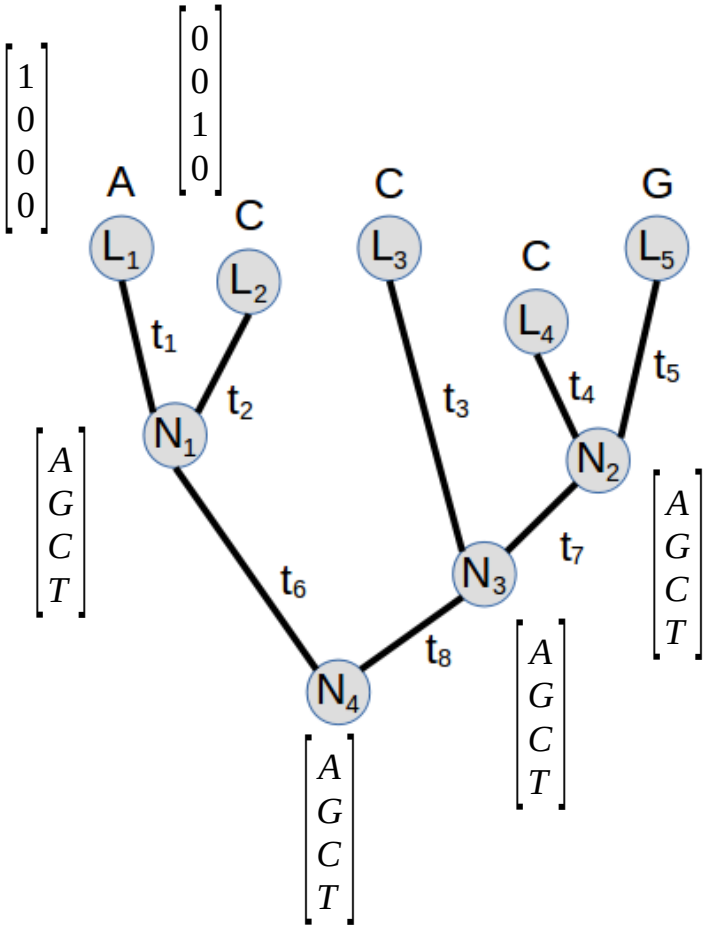
Used to:

1) Infer trees by estimation of the probability of the data given a tree
2) Estimates rates and other parameters of sequence evolution given a tree

Likelihood of tree

$$P\left(L_1 \ldots L_5 \middle| \tau, \nu, Q\right) =$$

$$\sum_{N_4} \sum_{N_3} \sum_{N_2} \sum_{N_1} P\left(N_4\right) P_{N_4 N_1}\left(t_6\right)$$
$$P_{N_1 L_1}\left(t_1\right) P_{N_1 L_2}\left(t_2\right)$$
$$P_{N_4 N_3}\left(t_8\right) P_{N_3 L_3}\left(t_3\right) P_{N_3 N_2}\left(t_7\right)$$
$$P_{N_2 L_4}\left(t_4\right) P_{N_2 L_5}\left(t_5\right)$$



Sum is over all possible ways of generating leaves given the model. Models can then be compared.

Felsenstein's Pruning Algorithm

Branches are independent so sums can be broken up into parts and solved by recursion

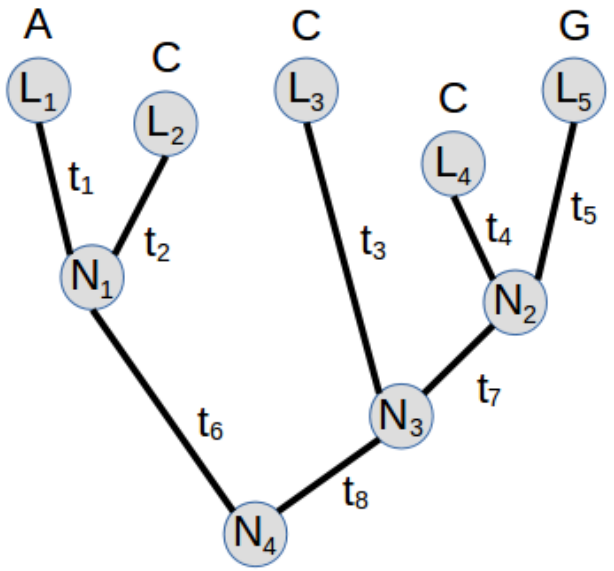
$$P\left(L_1 \ldots L_5, \middle| \tau, \nu, Q\right) =$$

$$\sum_{N_4} P\left(N_4\right) P\left(N_1 \middle| N_{4,} t_6\right) P\left(N_3 \middle| N_4, t_8\right)$$

$$\sum_{N_1} P\left(L_1 \middle| N_{1,} t_1\right) P\left(L_2 \middle| N_{1,} t_2\right)$$

$$\sum_{N_3} P\left(L_3 \middle| N_3, t_3\right) P\left(N_2 \middle| N_3, t_7\right)$$

$$\sum_{N_2} P\left(L_4 \middle| N_2, t_4\right) P\left(L_5 \middle| N_2, t_5\right)$$



The likelihood of state *s* at node *k* can thus be written as a recursion:

$$L_k(s) = \left(\sum_x P\left(x \middle| s, t_l\right) L_l(x)\right)\left(\sum_y P\left(y \middle| s, t_m\right) L_m(y)\right)$$

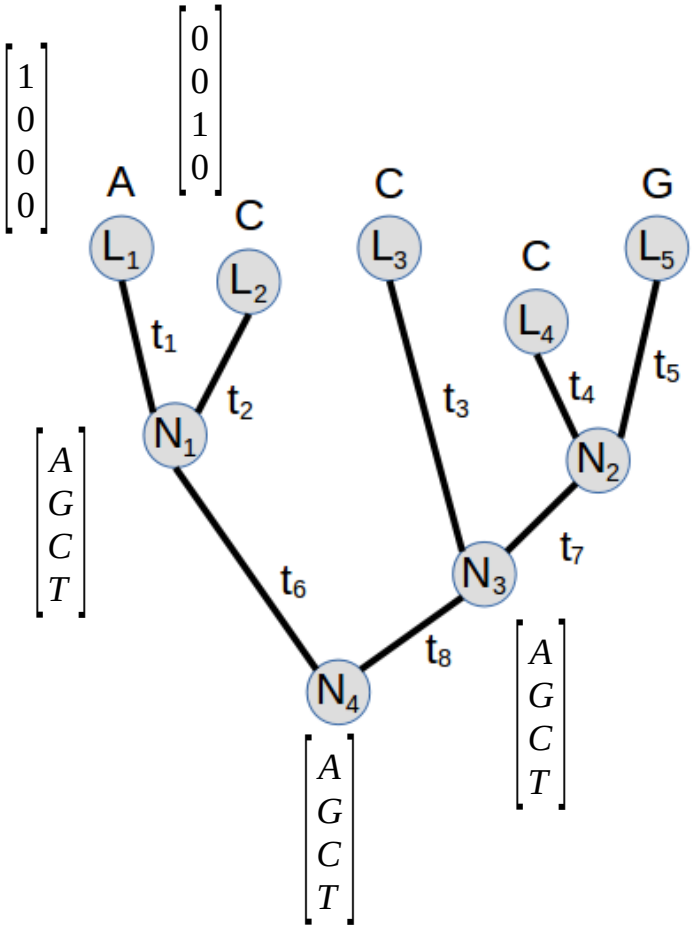
is the product of the likelihoods of state *x* at node *l* and state *y* at node *m* summed over all possible states.

Calculating the likelihood

For subtree N_1:

$$P(N_1 = A) = P_{AA}(t_1) P_{CA}(t_2)$$

$$P(N_1 = \begin{bmatrix} A \\ G \\ C \\ T \end{bmatrix}) = \begin{bmatrix} P_{AA}(t_1) P_{CA}(t_2) \\ P_{AG}(t_1) P_{CG}(t_2) \\ P_{AC}(t_1) P_{CC}(t_2) \\ P_{AT}(t_1) P_{CT}(t_2) \end{bmatrix}$$



What is the probability of N_4 = A?

$$P(N_4 = A) = L(\text{left}) L(\text{right})$$

$$L(\text{left}) = \sum_X P_{XA}(t_6) L_1(X)$$

$$L(\text{left}) = P_{AA}(t_6) L_1(A) + P_{GA}(t_6) L_1(G) + P_{CA}(t_6) L_1(C) + P_{TA}(t_6) L_1(T)$$

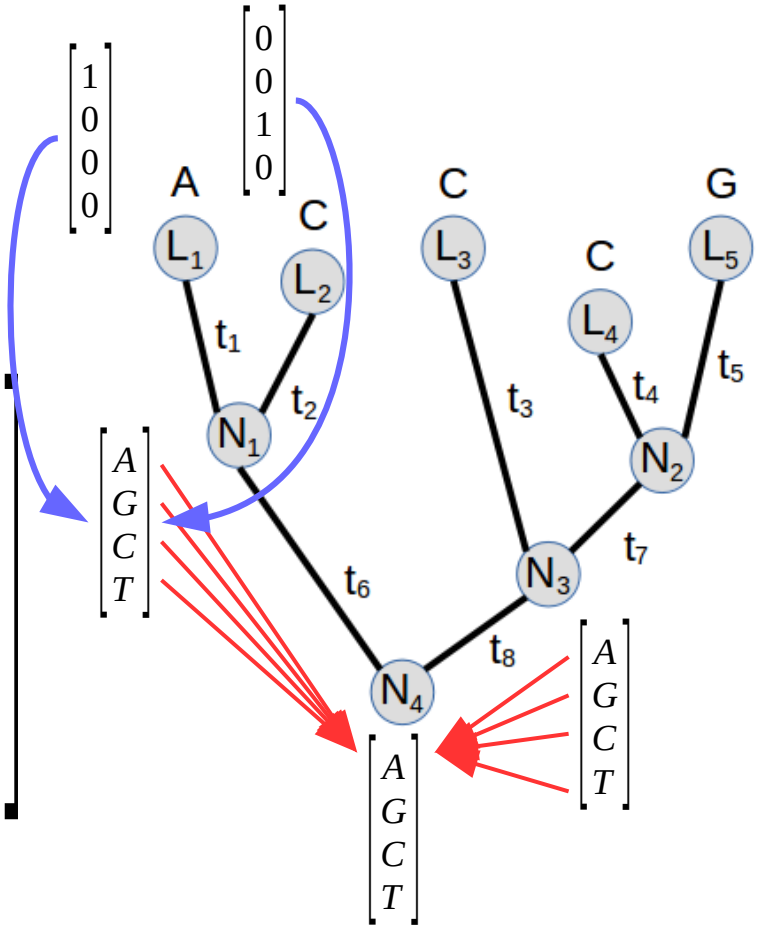
Calculating the likelihood

For subtree N_1:

$$P(N_1 = A) = P_{AA}(t_1) P_{CA}(t_2)$$

$$P(N_1 = \begin{bmatrix} A \\ G \\ C \\ T \end{bmatrix}) = \begin{bmatrix} P_{AA}(t_1) P_{CA}(t_2) \\ P_{AG}(t_1) P_{CG}(t_2) \\ P_{AC}(t_1) P_{CC}(t_2) \\ P_{AT}(t_1) P_{CT}(t_2) \end{bmatrix}$$

What is the probability of N_4 = A?



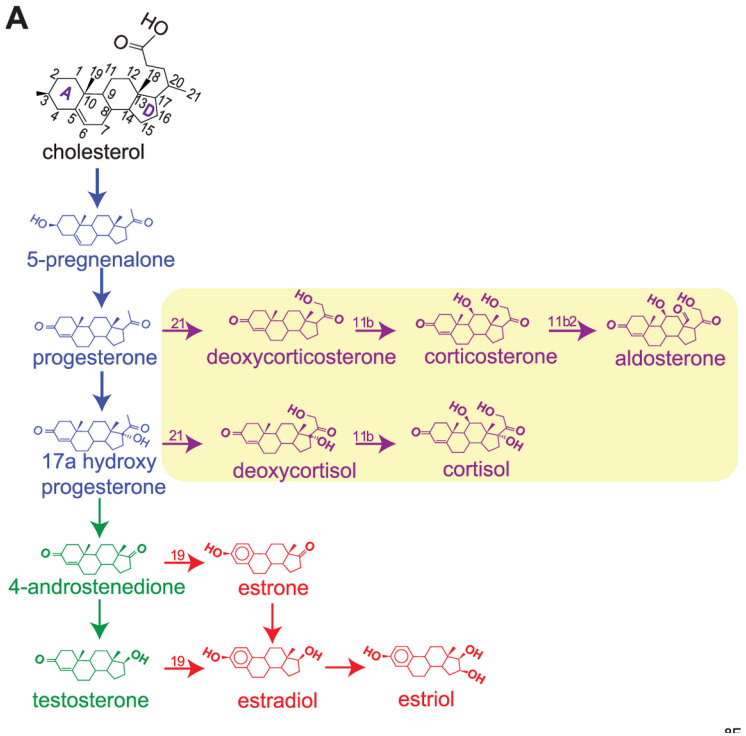
$$L = P(D|\tau, M) = \prod_{i=1}^{m} P(D_i|\tau, M)$$

Likelihood over m sites is the product of each likelihood

Finding the best tree

- Tree topology search
  - Exhaustive (too slow)
  - Local rearrangement of tree topology (greedy)
  - Sampling methods (MCMC)
  - Accept or reject based on the likelihood
    - Requires maximization of branch lengths
- Branch lengths (greedy, no local maximum)
  - Optimized by expectation maximization (EM)
  - Newton-Ralphson method (iterative derivatives)

Ancestral states
Resurrecting ancient steroid receptors



Pathway for synthesis of vertebrate steroid hormones

androgen receptors (AR): testosterone (muscle, bone, hair)

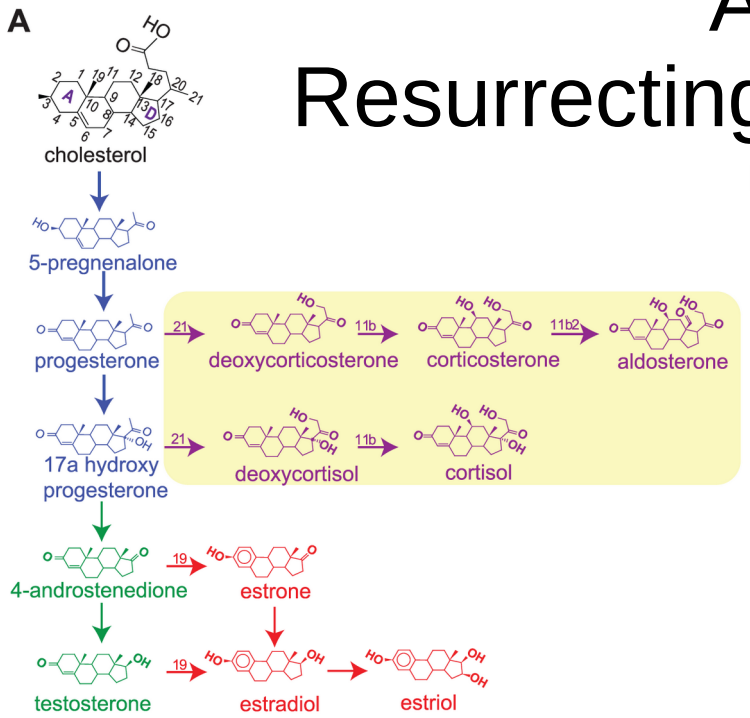
progesterone receptors (PR): progesterone (menstrual cycle, pregnancy)

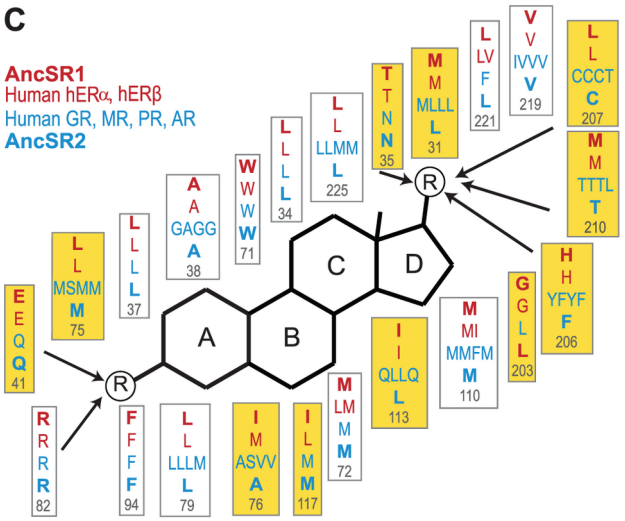
estrogen receptors (ER): estrogen (reproductive cycle, fat, bone)

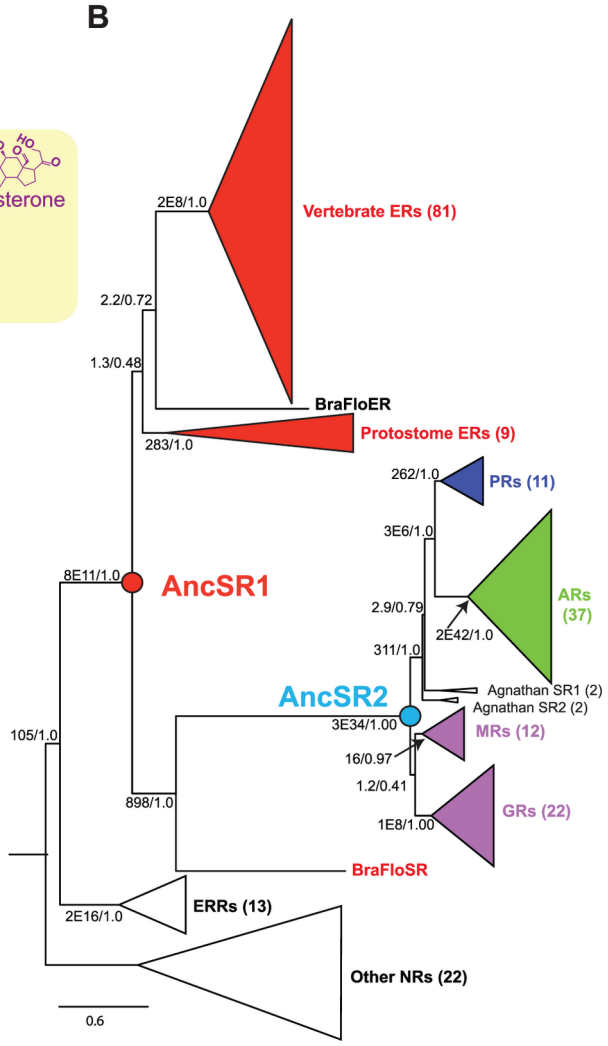
glucocorticoid receptors (GR): aldosterone (sweat, blood pressure), cortisol (metabolic, immune, sleep, memory, etc)

- Make a tree
- reconstruction ancestral states
- measure their receptor activity to each compound
- infer the evolution of steroid receptors

Ancestral states
Resurrecting ancient steroid receptors







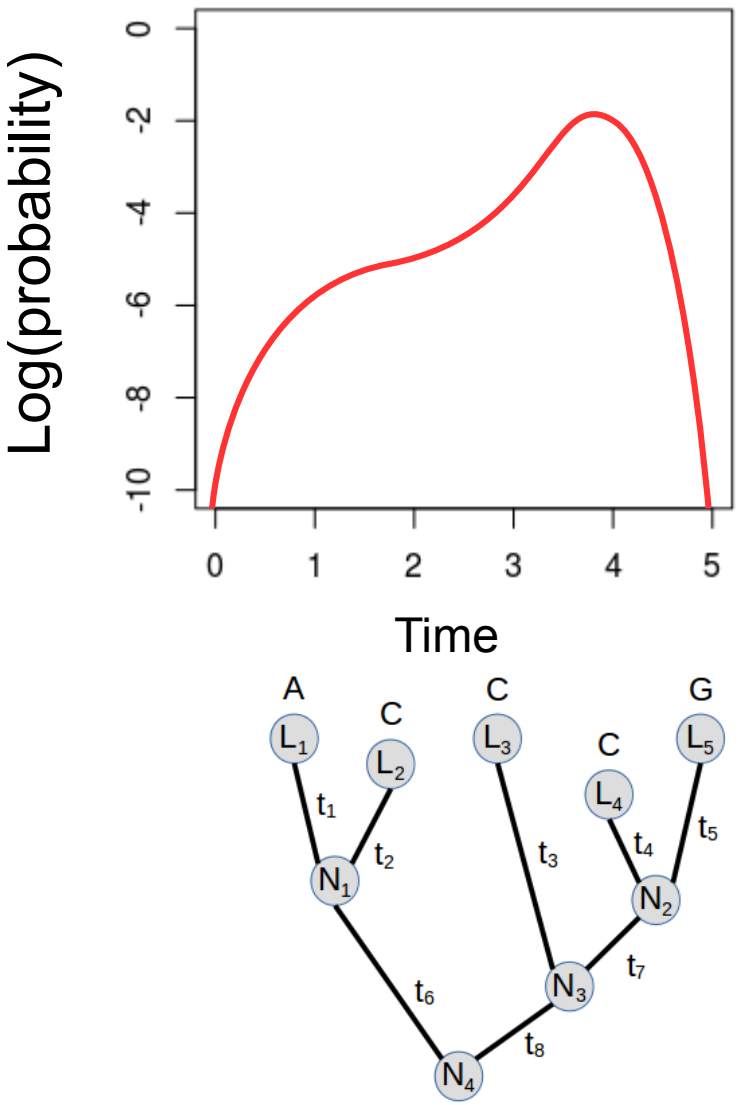
B) Phylogeny of the steroid hormone receptors (SR) gene family.

C) Maximum likelihood reconstruction of ligand-contacting amino acids in AncSR1 and AncSR2

Evolved according to a principle of minimal specificity-at each point in time, receptors evolved ligand recognition that were just specific enough to parse the set of endogenous substances to which they were exposed

Problem: multiple unknown parameters
Solution: Markov Chain Monte Carlo

$$P(t) = \exp(Qt)$$



Maximum likelihood

- For simple functions, solve by setting the derivative to zero (Bernoulli)
- For simple computations, evaluate numerically (branch length)
- What about when there are many unknown parameters?

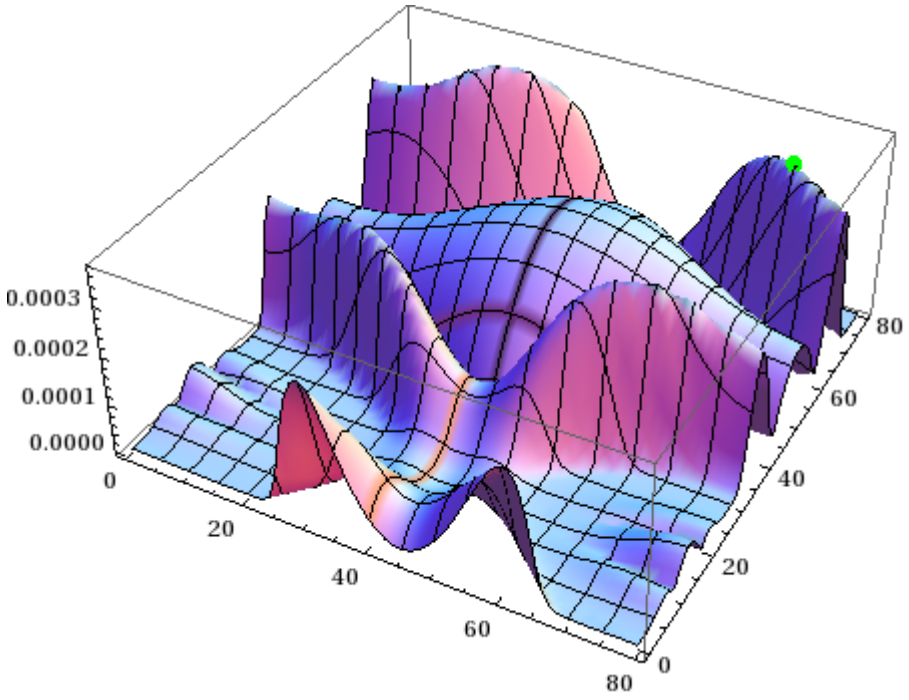
Tree parameters

- t_1-t_8 (each on positive number line)
- N_1-N_4 (each with 4 possible states: Felsenstein's pruning algorithm solves this problem)

Problem: multiple unknown parameters
Solution: Markov Chain Monte Carlo

$$P(t) = \exp(Qt)$$

As number of parameters increases so does the possibility of multiple local maxima



Maximum likelihood

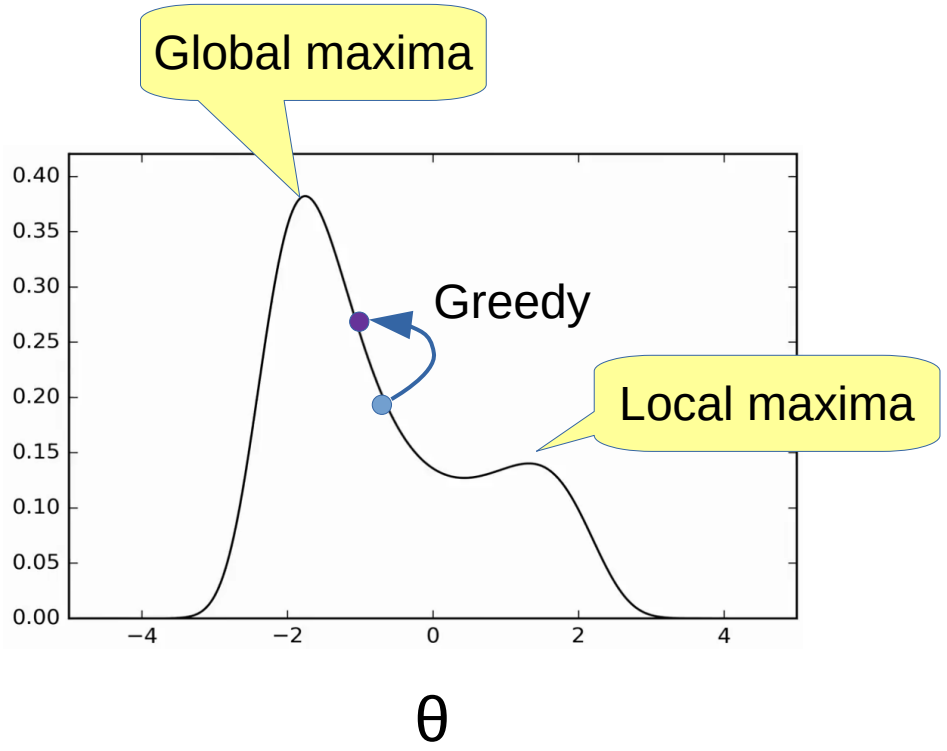
- For simple functions, solve by setting the derivative to zero (Bernoulli)
- For simple computations, evaluate numerically (branch length)
- What about when there are many unknown parameters?

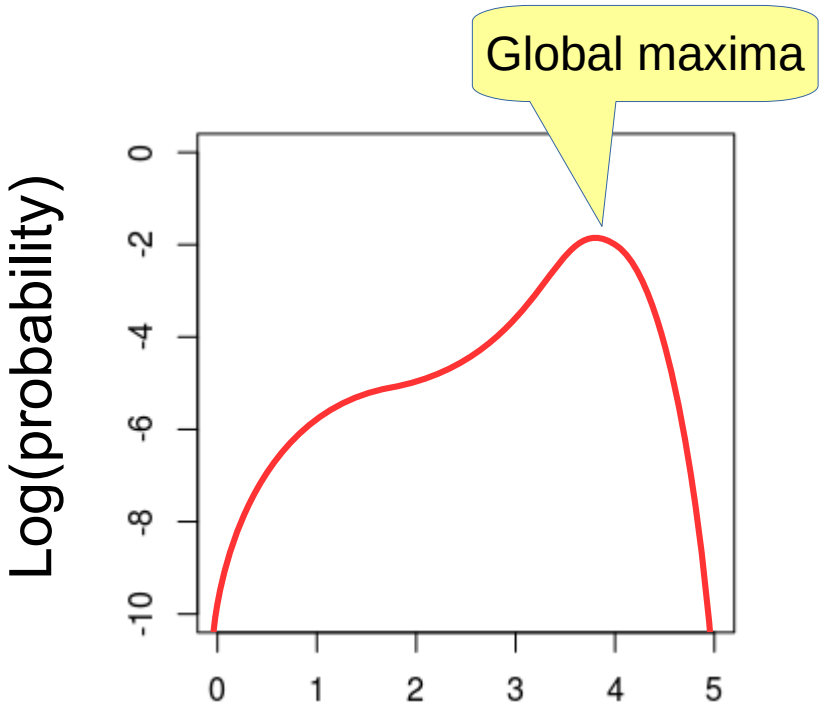
Tree parameters

- t_1-t_8 (each on positive number line)
- N_1-N_4 (each with 4 possible states: Felsenstein's pruning algorithm solves this problem)

Problem #2: multiple local maxima

Greedy algorithms can rapidly find local maxima: evaluate likelihood to the left and right of current position and move uphill





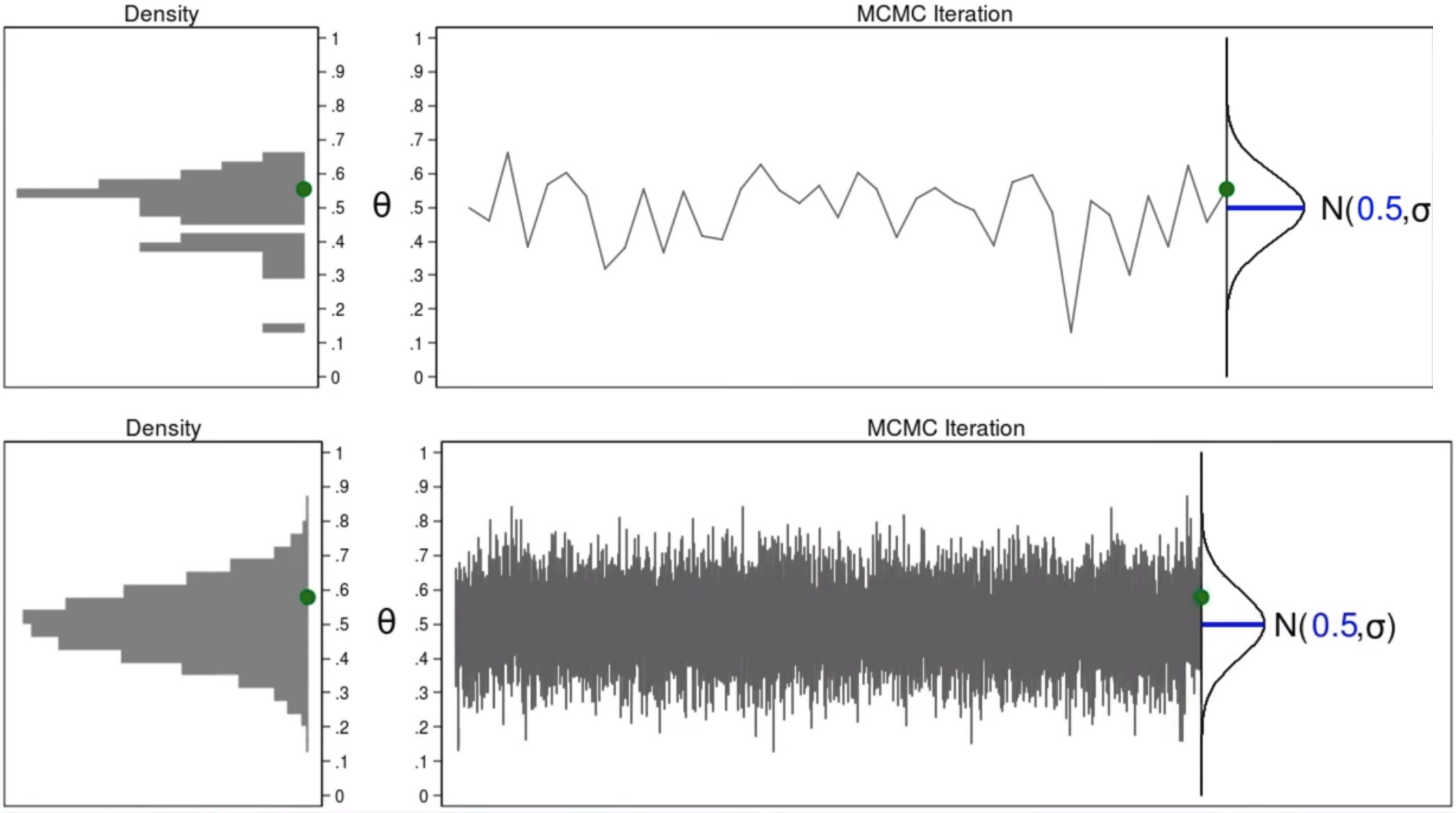
Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC)

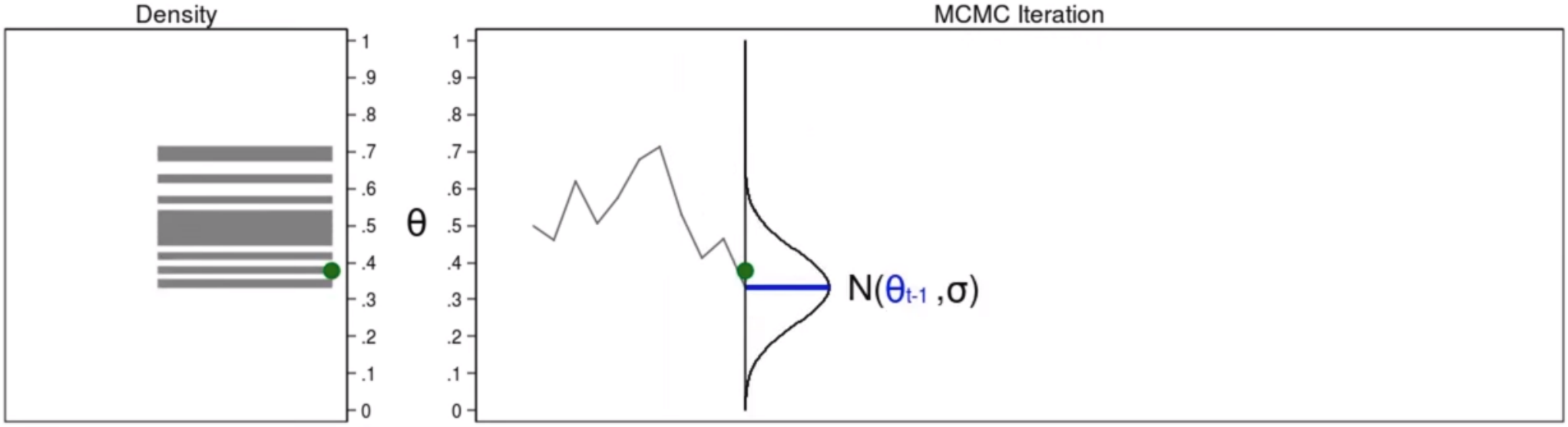
- a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution.
- The state of the chain after a number of steps is then used as a sample of the desired distribution.
- The quality of the sample improves as a function of the number of steps.
- MCMC methods
  - Metropolis–Hastings algorithm: random walk
  - Gibbs sampling: requires conditional distributions

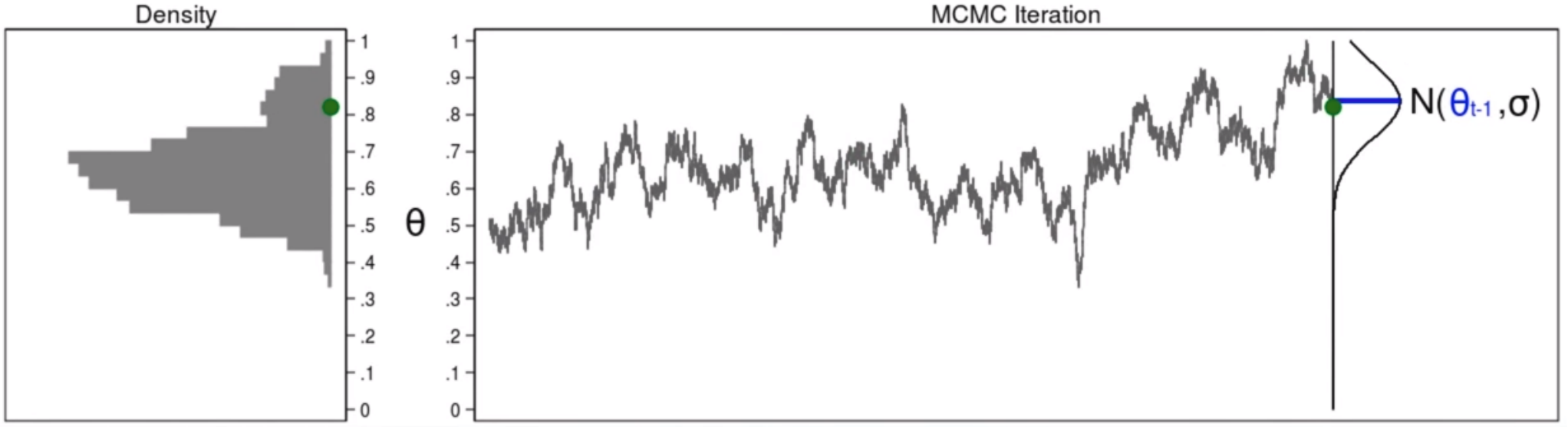
Used to estimate parameters of a model, among other things. For example, MLE when complex likelihood function.

Monte Carlo: algorithm that uses repeated random sampling to obtain numerical results



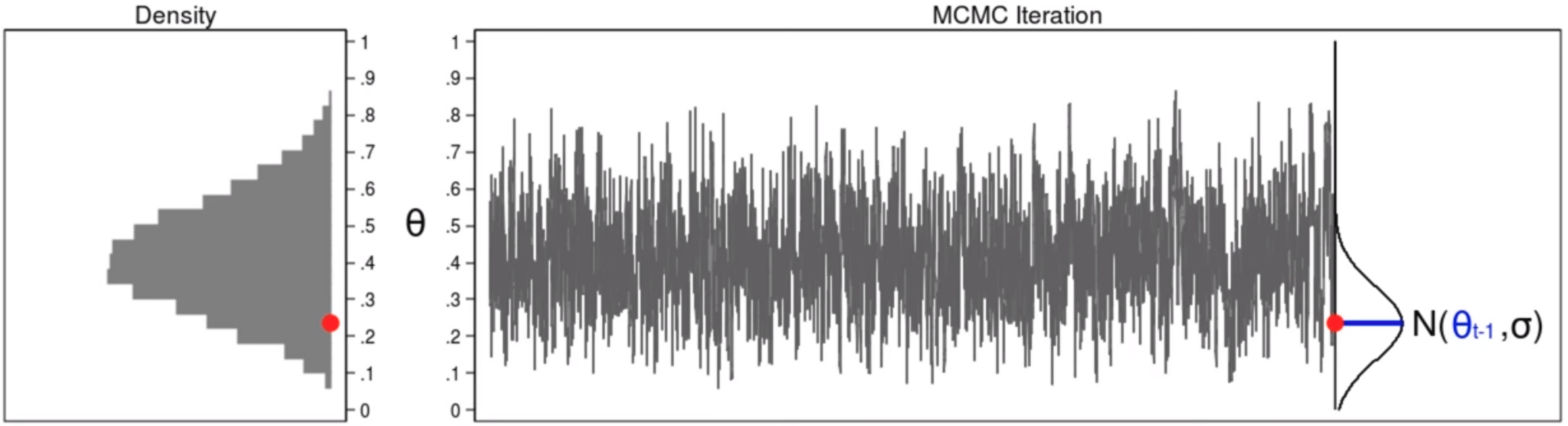
Random Walk: a path of successive random steps following a stochastic process (Markov Chain)





Metropolis-Hastings

- Construct a Markov chain such that its equilibrium probability distribution is our target distribution (e.g. likelihood)
- Iterate the Markov chain many times
- Approximate the quantities of interest (e.g. max) using the draws from the chain



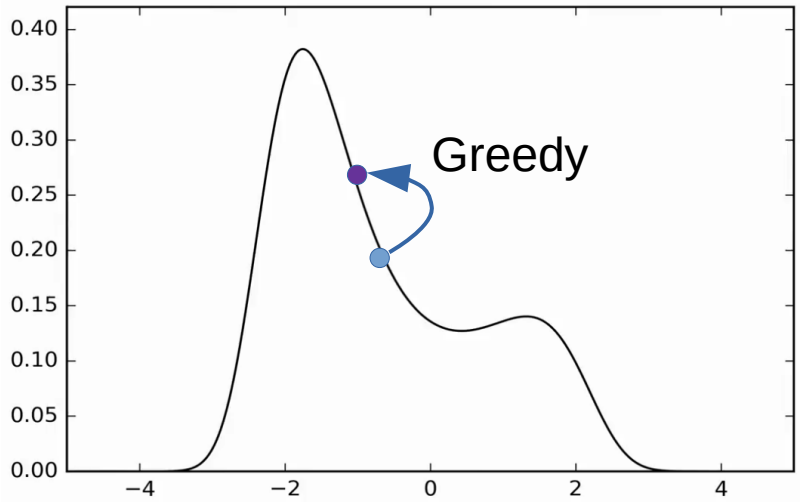
Example using a biased coin with parameter Theta

Metropolis-Hastings equations

$$g(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

$$\rho = \frac{p(X|\theta_p) \cdot p(\theta_p)}{p(X|\theta) \cdot p(\theta)}$$

proposal



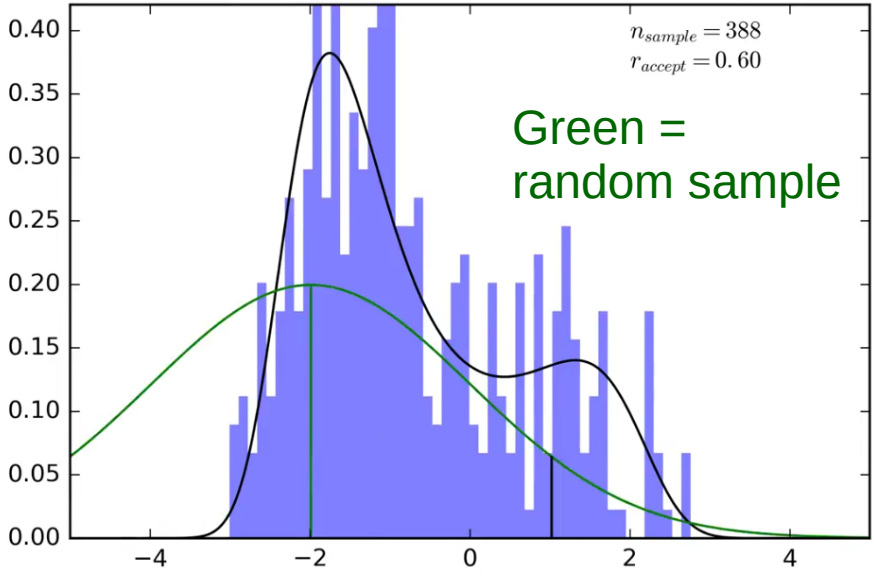
Numerical integration vs MCMC

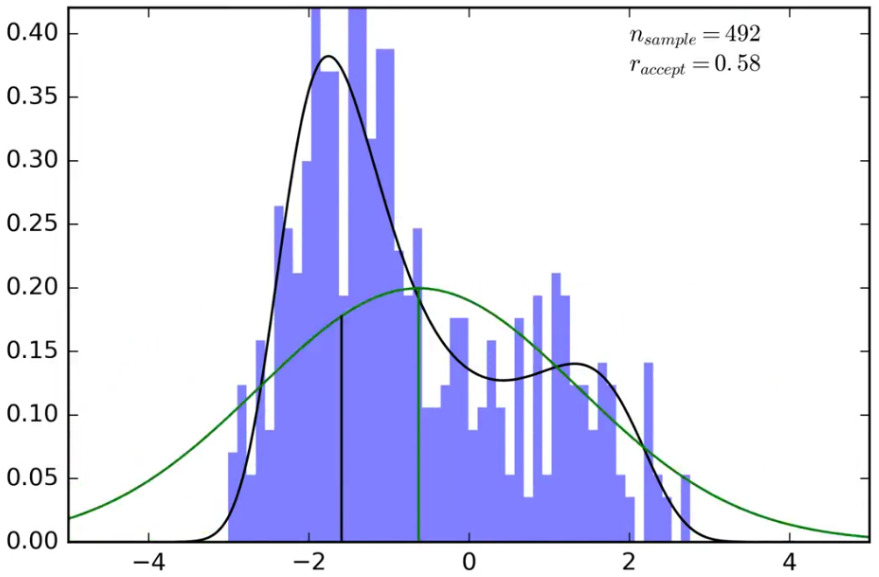
$$p(X) = \int d\theta^* \, p(X|\theta^*) \, p(\theta^*)$$

What if many parameters?
How do we avoid local maxima?
Answer: tune acceptance ratio using the variance of the proposal distribution.

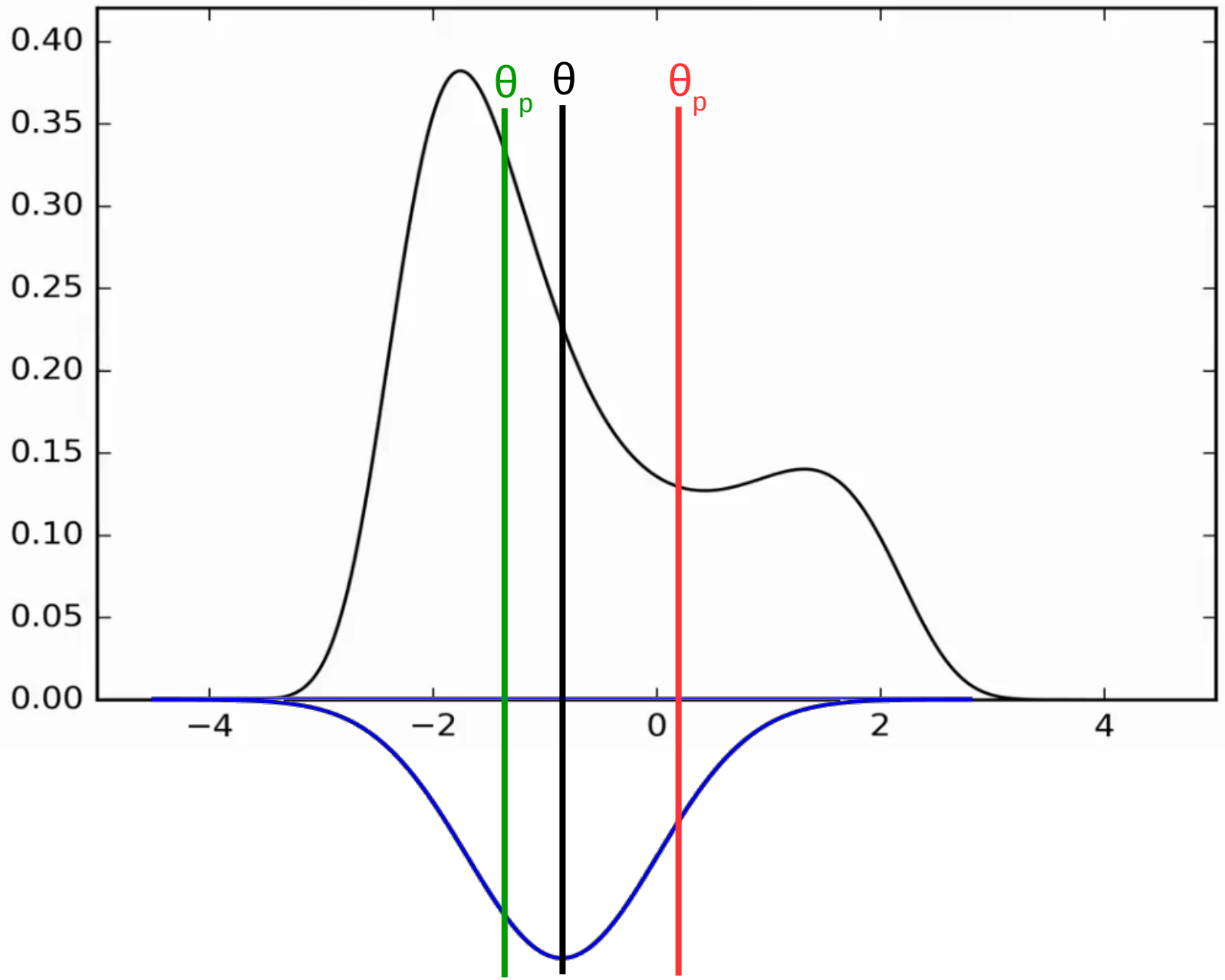
Metropolis-Hastings Algorithm

- Choose a new proposed value (θ_p) such that $\theta_p = \theta + \Delta\theta$ where $\Delta\theta \sim N(0, \sigma^2)$, i.e. a Gaussian distribution with mean of 0 and variance σ^2.
- Calculate the ratio: $\rho = \frac{g(\theta_p|X)}{g(\theta|X)}$ where $g(\theta|X)$ is the posteriori probability.
- If the proposal distribution is not symmetrical, we need to weight the acceptance probability to maintain detailed balance (*reversibility*) of the stationary distribution: $\rho = \frac{g(\theta_p|X) \cdot p(\theta|\theta_p)}{g(\theta|X) \cdot p(\theta_p|\theta)}$
- If $\rho \geq 1$, then set $\theta = \theta_p$, if $\rho < 1$ set $\theta = \theta_p$ with probability ρ, and otherwise set $\theta = \theta$.
- Repeat

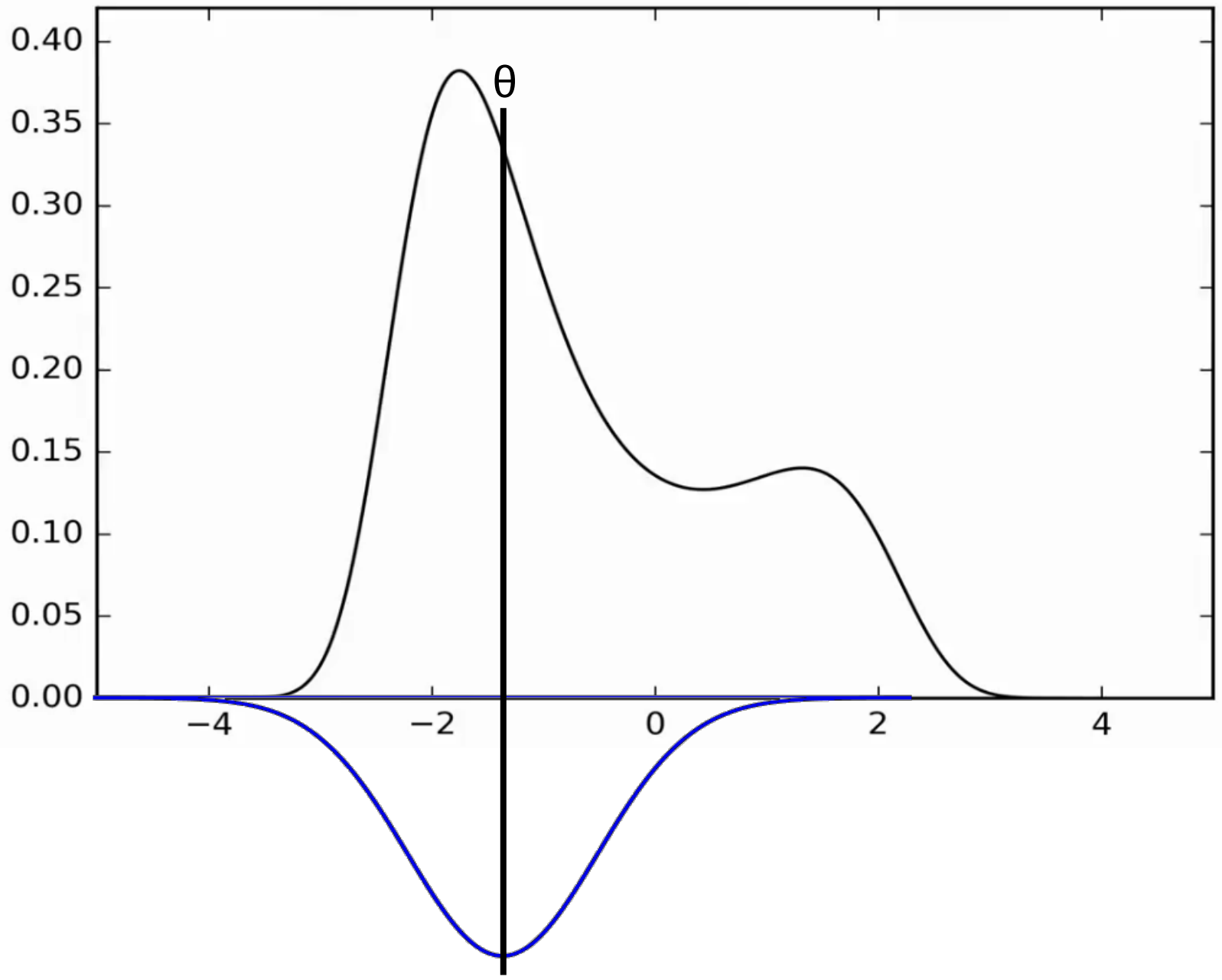




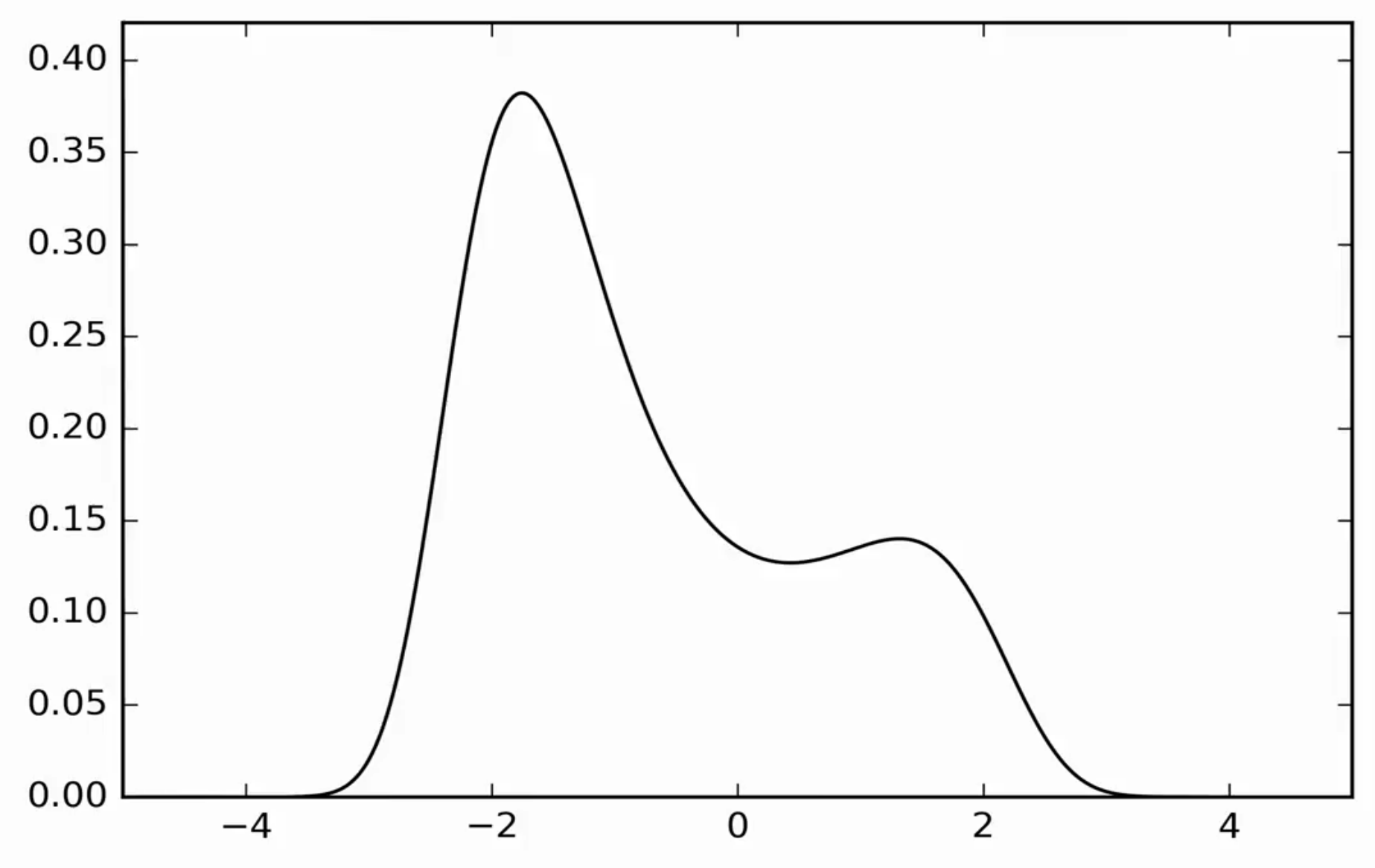
MCMC



MCMC



Metropolis-Hastings



Exercises

1) Which methods assume a molecular clock and which methods evaluate multiple trees: UPGMA, neighbor-joining, parsimony, maximum likelihood

2) Calculate branch lengths and tree (UPGMA) given the distance matrix:

	a	b	c
a	0		
b	13	0	
c	4	7	0

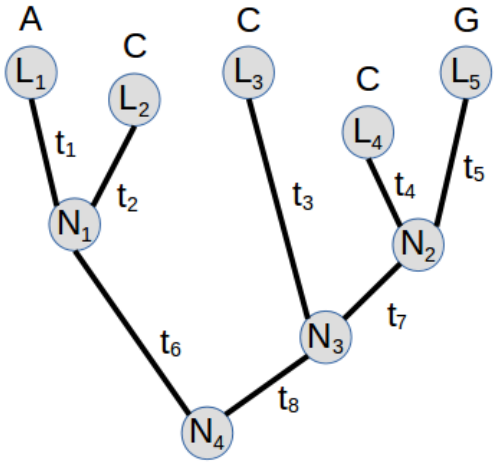
Exercises

3) What causes long branch attraction in parsimony tree reconstruction?

4) What is the most parsimonious unrooted tree? What is the minimum number of mutations?

Sequence	Site 1	2	3	4
A	T	A	A	A
B	G	G	G	A
C	G	A	A	G
D	A	G	T	T

5) What is the probability of N_3 = A, written in the form of $P_{ij}(t)$?



Exercises

6) Which algorithm would you use to find the maximum likelihood of the functions below: greedy or Monte Carlo Markov Chain