Tutorial on Phylogenetic Methods

- Understanding Trees
- Alignments
- Distances
- Clustering Methods
- Bootstrapping
- Likelihood Methods
- Parsimony

Recommended books:

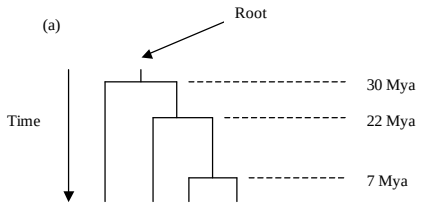
R. Page and A. Holmes – *Molecular evolution: a phylogenetic approach*

W. H. Li – *Molecular evolution*

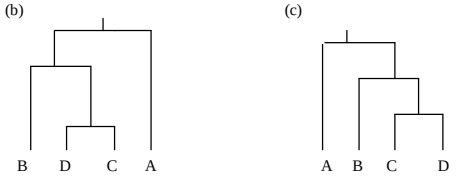
1

Understanding Trees #1

Rooted Trees with a Time Axis



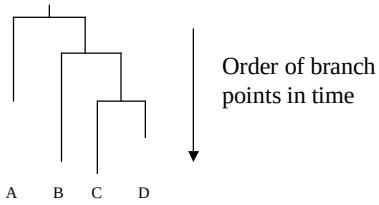
Trees are like mobiles …



2

Understanding Trees #2

A rooted tree with branches scaled according to the amount of evolutionary change.

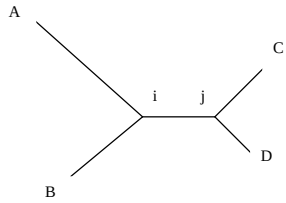


This is one of the most informative types of tree.

3

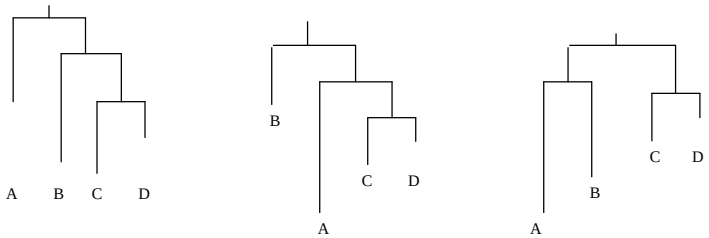
Understanding trees #3

An unrooted tree with branch lengths proportional to evolutionary distance



i and *j* are internal nodes at which the sequence is unknown

Three ways of rooting the above tree:



Most phylogenetic methods do not tell you where the root is.

Need to specify the outgroup (most distantly related species).

4

Part of sequence alignment of Mitochondrial Small Sub-Unit rRNA

Full gene is length ~950

11 Primate species with mouse as outgroup

```
                  *        20         *        40         *        60         *
Mouse       : CUCACCAUCUCUUGCUAAUUCAGCCUAUAUACCGCCAUCUUCAGCAAACCCUAAAAAGG-UAUUAAAGUAAGGAAAAGA : 78
Lemur       : CUCACCACUUCUUGCUAAUUCAAGUUAUAUACCGCCAUCCCAGCAAACCCUAUUAAGGCCC-CAAAGUAAGGAAAAAC : 78
Tarsier     : CUUACCACCUCUUGCUAAUUCAGUCUAUAUACCGCCAUCUUCAGCAAACCCUAAUAAAGGUUUUAAAGUAAGCAAGU : 79
SakiMonkey  : CUUACCACUCUUGGC-AU-CAGCCUGUAUACCGCCAUCUUCAGCAAACUCUA-UAAUGACAGUAAAGUAAGCAAGU : 76
Marmoset    : CUCACCACGUCUAGC-AU-CAGCCUGUAUACCGCCAUCUUCAGCAAACUCCU-UAAUGAUUGUAAAGUAAGCAGAAGU : 76
Baboon      : CCCACCCUCUCUUGCU----UAGUCUAUAUACCGCCAUCUUCAGCAAACCCUGAUGAAGGCUACGAAGUGAGCGCAAAU : 75
Gibbon      : CUCACCAUCUCUUGCU----CAGCCUAUAUACCGCCAUCUUCAGCAAACCCUGACAAAGGCUAUAAAGUAAGCACAAAC : 75
Orangutan   : CUCACCACCCCUUGCU----CAGCCUAUAUACCGCCAUCUUCAGCAAACCCUGAUGAAGGCCACGAAGUAAGCGCAAAC : 75
Gorilla     : CUCACCACCUCUUGCU----CAGCCUAUAUACCGCCAUCUUCAGCAAACCCUGACGAAGGCCACAAAGUAAGCACAAGU : 75
PygmyChimp  : CUCACCGCCUCUUGCU----CAGCCUAUAUACCGCCAUCUUCAGCAAACCCUGAUGAAGGUUACAAAGUAAGCGCAAGU : 75
Chimp       : CUCACCGCCUCUUGCU----CAGCCUAUAUACCGCCAUCUUCAGCAAACCCUGAUGAAGGUUACAAAGUAAGCGCAAGU : 75
Human       : CUCACCACCUCUUGCU----CAGCCUAUAUACCGCCAUCUUCAGCAAACCCUGAUGAAGGCUACAAAGUAAGGCCAAGU : 75
              CucACC  cuCUuGCu    cAgccUaUAUACCGCCAUCuuCAGCAAACCcu    A G     aAAGUaAGC  AA
```

5

What is this ?

Gorilla

(Gorilla gorilla)

and this ?

Baboon

(Papio hamadryas)

6

From alignment construct pairwise distances.

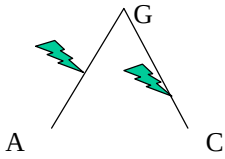
```
Species 1:  AAGTCTTAGCGCGAT
Species 2:  ACGTCGTATCGCGAT
              *    *   *
```

p = 3/15 = 0.2

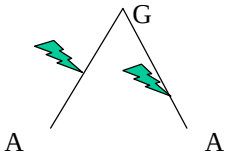
p = fraction of differences between sequences

BUT - p is not an additive distance,

p does not increase linearly with time.



2 substitutions happened - only 1 is visible



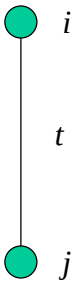
2 substitutions happened - nothing visible

7

Models of Sequence Evolution

r_{ij} is the rate of substitution from state i to state j

States label bases A,C,G & T



$P_{ij}(t)$ = probability of being in state j at time t given that ancestor was in state i at time 0.

$$\frac{dP_{ij}}{dt} = \sum_k P_{ik} r_{kj}$$

8

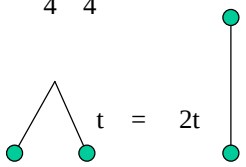
Jukes - Cantor Model

All substitution rates = α

All base frequencies are 1/4

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}\exp(-4\alpha t) \qquad P_{ij}(t) = \frac{1}{4} - \frac{1}{4}\exp(-4\alpha t)$$

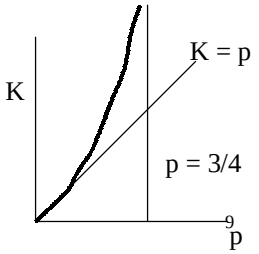
$$p = 1 - P_{ii}(2t) = \frac{3}{4}(1 - \exp(-8\alpha t))$$



Mean number of substitutions per site: $K = 6\alpha t$

$$K = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

K increases linearly with time



9

The HKY model has a more general substitution rate matrix

from \ to	*A*	*G*	*C*	*T*
A	*	$\kappa\alpha\pi_G$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\kappa\alpha\pi_A$	*	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	*	$\kappa\alpha\pi_T$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\kappa\alpha\pi_C$	*

The frequencies of the four bases are

κ is the transition-transversion rate parameter

* means minus the sum of elements on the row

10

What is this ?

Ringtailed lemur
(Lemur catta)

Sifaka lemur
(Propithecus verreauxi)

11

Part of the Jukes-Cantor Distance Matrix for the Primates example

	Baboon	Gibbon	Orang	Gorilla	PygmyCh.	Chimp	Human
Baboon	0.00000	0.18463	**0.19997**	0.18485	0.17872	0.18213	0.17651
Gibbon	0.18463	0.00000	0.13232	0.11614	0.11901	0.11368	0.11478
Orang	0.19997	0.13232	0.00000	0.09647	0.09767	0.09974	0.09615
Gorilla	0.18485	0.11614	0.09647	0.00000	0.04124	0.04669	0.04111
PygmyChimp	0.17872	0.11901	0.09767	0.04124	0.00000	**0.01703**	0.03226
Chimp	0.18213	0.11368	0.09974	0.04669	0.01703	0.00000	**0.03545**
Human	0.17651	0.11478	0.09615	0.04111	0.03226	0.03545	0.00000

Mouse-Primates ~ 0.3

Use as input to clustering methods

12

Follow a clustering procedure:

1. Join closest 2 clusters
2. Recalculate distances between all the clusters
3. Repeat this until all species are connected in a single cluster.

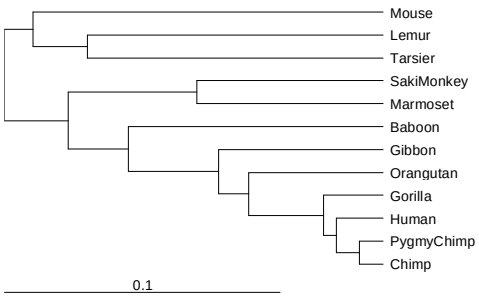
Initially each species is a cluster on its own. The clusters get bigger during the process, until they are all connected to a single tree.

There are many slightly different clustering methods. The two most important are:

- UPGMA (Unweighted Pair Group Method with Arithmetic Mean)
- Neighbour-Joining

13

UPGMA - The distance between cluster X and cluster Y is defined as the average of pairwise distances d_{AB} where A is in X and B is in Y.



• assumes a molecular clock

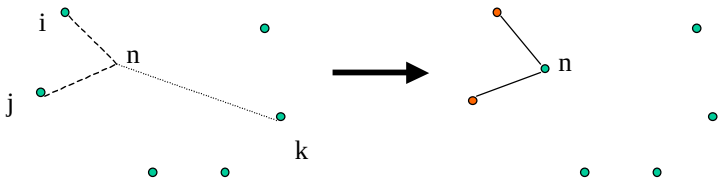
•distance = twice node height

• forces distances to be ultrametric (for any three species, the two largest distances are equal)

• produces rooted tree (in this case root is incorrect but topology is otherwise correct)

14

Neighbour-Joining method



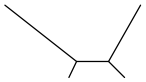
Take two neighbouring nodes *i* and *j* and replace by a single new node *n*.

$d_{in} + d_{nk} = d_{ik}$; $d_{jn} + d_{nk} = d_{jk}$; $d_{in} + d_{jn} = d_{ij}$;

therefore $d_{nk} = (d_{ik} + d_{jk} - d_{ij})/2$; applies for every *k*

define $r_i = \frac{1}{N-2}\sum_k d_{ik}$ $r_j = \frac{1}{N-2}\sum_k d_{jk}$

but ...



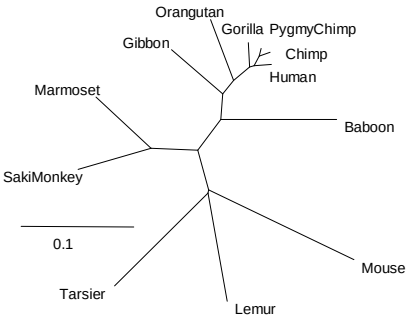
Let $d_{in} = (d_{ij} + r_i - r_j)/2$; $d_{jn} = (d_{ij} + r_j - r_i)/2$.

Rule: choose *i* and *j* for which D_{ij} is smallest, where $D_{ij} = d_{ij} - r_i - r_j$.

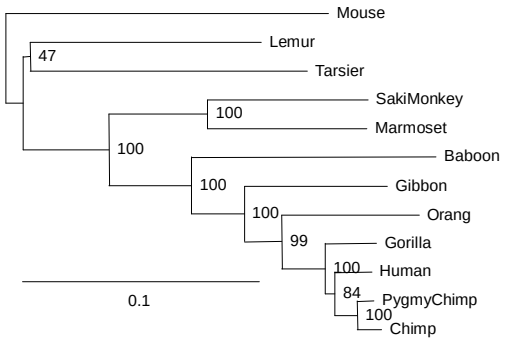
15

NJ method produces an Unrooted, Additive Tree

Additive means distance between species = distance summed along internal branches



The tree has been rooted using the Mouse as outgroup



16

Bootstrapping

real sequences:

```
Human        CAACAGAGGC TTACGACCCC TTATTTACC
Chimp        .......... .C........ .........
Gorilla      T......... .C..A..... .........
Orang utan   T..T..G.C. CC..A..... .........
Gibbon       ...T...... .CGAA...T. ..GC.....
```

resample columns at random:

```
             Human        C  C  G  etc.
             Chimp        .  .  .
             Gorilla      .  .  A
             Orang utan   .  T  A
             Gibbon       .  T  A
```

Generate many random samples. Calculate tree for each one. These trees will be slightly different because the input sequences were slightly different.

For each clade, the bootstrap value is the percentage of the trees for which this group of species falls together in the tree. High bootstrap values (>70%) indicate reliable trees. Lower percentages indicate that there is insufficient information in the sequences to be sure about the resulting tree.

17

What is this ?

Pygmy marmoset

(Callithrix pygmaea)

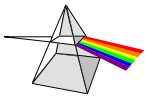
Common marmoset

(Callithrix jacchus)

18

Why are trees never exact?

Ideal World :



Take a known tree. Simulate evolution along that tree according to a known evolutionary model. Generate a set of sequences for the tips of the tree.

Use correct evolutionary model to generate distance matrix. Regenerate Tree.

You will not get exactly back to the original tree because substitution events are a stochastic process (finite size effects).

Real World:

You don't know the correct model of evolution.

There may be variability of rates between sites in the sequence, between species in the tree, or between times during evolution.

Saturation of mutations will occur for widely separated species.

Evolutionary radiations may have occurred in short periods of time.

Recombination or horizontal transfer may have occurred.

19

Criteria for Judging Additive Trees

There are N(N-1)/2 pairwise distances between N species.

There are only 2N-3 branch lengths on an unrooted tree, therefore in general the data do not fit exactly to an additive tree (cf. N-1 node heights in an ultrametric tree).

NJ:

• Produces an additive tree that approximates to the data matrix, but there is no criterion for defining the best tree.

• Algorithm for tree construction is well-defined and very rapid.

Fitch-Margoliash method:

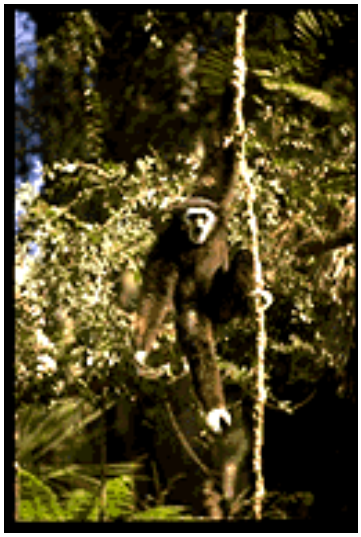
• Define measure of non-additivity $E = \sum_{ij} \frac{\left(d_{ij} - d_{ij}^{tree}\right)^2}{d_{ij}^2}$

• Choose tree topology and branch lengths that minimise *E*.

• Algorithm for tree construction is not defined. Requires a slow heuristic tree search method.

20

What is this ?



Gibbon

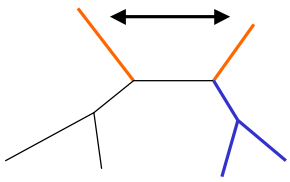
(Hylobates lar)

21

Searching Tree Space

Require a way of generating trees to be tested by Additivity criterion, Maximum Likelihood, or Parsimony.

Nearest neighbour interchange



Subtree pruning and regrafting

No. of distinct tree topologies

N	Unrooted (U_N)	Rooted (R_N)
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395

$U_N = (2N-5)U_{N-1}$ $R_N = (2N-3)R_{N-1}$

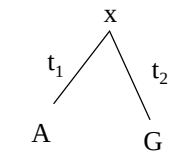
Conclusion: there are huge numbers of trees, even for relatively small numbers of species. Therefore you cannot look at them all.

22

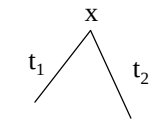
The Maximum Likelihood Criterion

Calculate the likelihood of observing the data on a given the tree.

Choose tree for which the likelihood is the highest.



$$L(x) = P_{xA}(t_1)P_{xG}(t_2)$$



$$L(x) = \sum_y \sum_z L(y)P_{xy}(t_1)L(z)P_{xz}(t_2)$$

Can calculate total likelihood for the site recursively.

Likelihood is a function of tree topology, branch lengths, and parameters in the substitution rate matrix. All of these can be optimized.

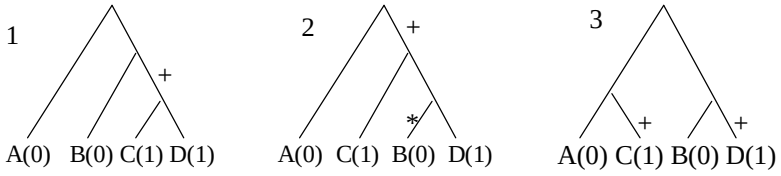
23

The Parsimony Criterion

Try to explain the data in the simplest possible way with the fewest arbitrary assumptions

Used initially with morphological characters.

Suppose C and D possess a character (1) that is absent in A and B (0)



In 1, the character evolves only once (+)

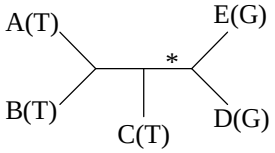
In 2, the character evolves once (+) and is lost once (*)

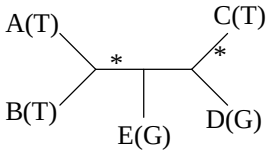
In 3, the character evolves twice independently

The first is the simplest explanation, therefore tree 1 is to be preferred by the parsimony criterion.

24

Parsimony with molecular data

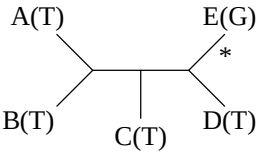


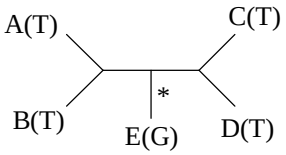


1. Requires one mutation

2. Requires two mutations

By parsimony, 1 is to be preferred to 2.





This site is *non-informative*. Whatever the arrangement of species, only one mutation is required. To be informative, a site must have at least two bases present at least twice.

The best tree is the one that minimizes the overall number of mutations at all sites.

25

Parsimony and Maximum Likelihood

- There is an efficient algorithm to calculate the parsimony score for a given topology, therefore parsimony is faster than ML.
- Parsimony is an approximation to ML when mutations are rare events.
- Weighted parsimony schemes can be used to treat most of the different evolutionary models used with ML.
- Parsimony throws away information from non-informative sites that *is* informative in ML and distance matrix methods.
- Parsimony gives little information about branch lengths.
- Parsimony is inconsistent in certain cases (Felsenstein zone), and suffers badly from long branch attraction.

Personal conclusion : if you've got a decent computer, you're better off with ML.

26

What is this ?

White-faced saki monkey

(Pithecia pithecia)

and this ?

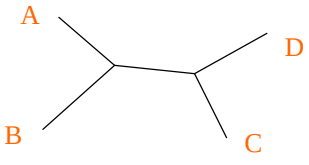
Pygmy Chimpanzee or

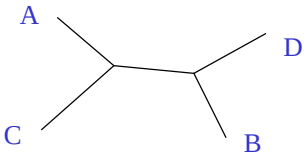
Bonobo *(Pan paniscus)*

27

Is the best tree significantly better than alternative trees?

The Kishino-Hasegawa test.

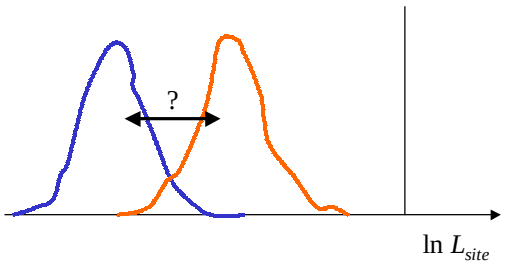




L_1 is slightly higher than L_2

$\ln L = \sum \ln L_{site}$

Calculate mean and variance of $\ln L_{site}$. Use standard statistical test to say whether means are different.



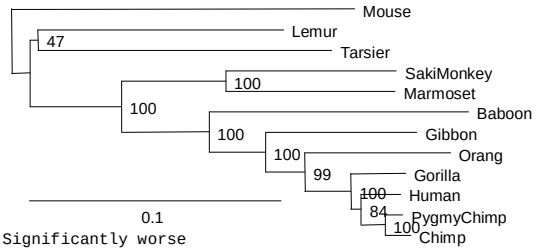
28

Using ML to rank the alternative trees for the primates example

The NJ Tree has:

(Gorilla,(Human,(Chimp,PygmyCh)))

((Lemur,Tarsier),Other Primates))



Tree	log L	difference	S.E.	Significantly worse	
1	-5783.03	0.39	7.33	no	NJ tree
2	-5782.64	0.00	<----------best tree		(Human,(Gorilla,(Chimp,PygmyCh)))
3	-5787.99	5.35	5.74	no	((Human,Gorilla),(Chimp,PygmyCh))
4	-5783.80	1.16	9.68	no	(Lemur,(Tarsier,Other Primates))
5	-5784.76	2.12	9.75	no	(Tarsier,(Lemur,Other Primates))
	??				
6	-5812.96	30.32	15.22	yes	(Gorilla,(Chimp,(Human,PygmyCh)))
7	-5805.21	22.56	13.08	no	Orang and Gorilla form clade
8	-5804.98	22.34	13.15	no	Gorilla branches earlier than Orang
9	-5812.97	30.33	14.17	yes	Gibbon and Baboon form a clade

29

Using Parsimony to rank the alternative trees for the primates example.

Tree	Steps	Diff Steps	Its S.D.	Significantly worse?	
1	1098.0	1.0	5.3879	No	NJ tree
2	1101.0	4.0	6.4840	No	(Human,(Gorilla,(Chimp,PygmyCh)))
3	1104.0	7.0	6.0858	No	((Human,Gorilla),(Chimp,PygmyCh))
4	1097.0	<------ best			(Lemur,(Tarsier,Other Primates))
5	1099.0	2.0	5.2942	No	(Tarsier,(Lemur,Other Primates))
	??				
6	1106.0	9.0	6.2481	No	(Gorilla,(Chimp,(Human,PygmyCh)))
7	1110.0	13.0	6.8591	No	Orang and Gorilla form clade
8	1110.0	13.0	6.8591	No	Gorilla branches earlier than Orang
9	1116.0	19.0	7.8141	Yes	Gibbon and Baboon form a clade

The best tree according to parsimony is different from that according to ML and NJ

Again, the NJ tree is second best

30

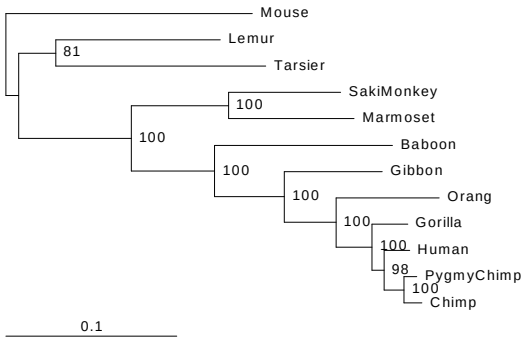
Quartet Puzzling

Calculate the ML tree topology for each subset of four species that can be chosen from the full set.

Write these as ((A,B),(C,D)); ((A,B),(D,E)); ((B,C),(D,E)); etc

Randomize the order of the species. Begin with the quartet tree for the first four species. Add the rest one at a time in such a way that the topology is maximally compatible with what is already there.

Repeat this with many randomized orders. Form consensus of all the resulting trees.



31

What is this ?

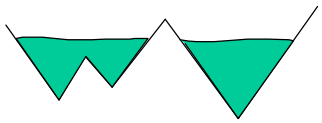
Tarsier *(Tarsius syrichta)*

32

Markov Chain Monte Carlo method: MCMC

We don't want a single tree, we want to sample over all tree-space with a weighting determined by the likelihood.

Physics analogy: finite temperature simulation rather than search for groundstate.



F = - ln L

Run a long simulation using the Metropolis-Hastings algorithm.

Move set = parameter changes + branch length changes + topology changes.

Trees generated are an equilibrium sample, therefore average properties of interest over the whole set with equal weighting.

Bayes' theorem: $$L(tree_i \mid data) = \frac{L(data \mid tree_i)}{\sum_k L(data \mid tree_k)}$$

$$\overline{Obs} = \sum_k Obs(tree_k) L(tree_k \mid data)$$

33

Topology probabilities for the primates according to MCMC

	Prob.	Cum.	Tree topology
1.	0.316	0.316	((L,T),O) + (G,(C,H)) = NJ
2.	0.196	0.512	((L,T),O) + ((G,C),H)
3.	0.145	0.657	(L,(T,O)) + (G,(C,H))
4.	0.121	0.778	(L,(T,O)) + ((G,C),H)
5.	0.113	0.891	(T,(L,O)) + ((G,C),H)
6.	0.101	0.992	(T,(L,O)) + (G,(C,H))
7.	0.007	0.999	(T,(L,O)) + ((G,H),C)
8.	0.001	0.999	((L,T),O) + ((G,H),C)
9.	0.001	1.000	(L,(T,O)) + ((G,H),C)

Only 9 topologies arose (= 3 x 3). The two uncertain points in the tree are almost independent.

The NJ tree has maximum posterior probability even though it was not the ML tree.

The ML tree has the second highest posterior probability.

?? Broad, Low peaks versus Narrow, High peaks.

34

Clade probabilities compared from three methods

	MCMC	NJ Bootstrap	QuartetPuzzling
Clades in the top tree			
(Human,Chimp)	**56%**	**84%**	**98%**
(Lemur,Tarsier)	**51%**	**46%**	**81%**
Clades not in the top tree			
(Gorilla,Chimp)	**43%**	**9%**	**2%**
(Tarsier,Other Pri.)	**27%**	**29%**	**18%**
(Lemur,Other Pri.)	**18%**	**24%**	**2%**
(Human,Gorilla)	**1%**	**7%**	**1%**

35

What is this ?

Homo sapiens

Lunchtime !

36