Estimating Phylogenies (Evolutionary Trees) II

Biol4230 Thurs, March 1, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Tree estimation strategies:

- Parsimony
  - ?no model, simply count minimum number of changes
  - many sites not "informative"
  - how minimum must minimum be?
- Distance
  - global "distance" between sequences (all sites informative)
  - measured distances underestimate evolutionary change
  - Combined algorithm/criterion approaches (UPGMA, NJ) use distance
  - where distance and parsimony differs
- Statistical (Model based) approaches

To learn more:

- Pevsner Bioinformatics Chapter 6 pp 179–212
- ** Felsenstein, J. Numerical methods for inferring evolutionary trees. *Quart. Review of Biology* 57, 379–404 (1982).
- Graur and Li (2010) "Fundamentals of Molecular Evolution" Sinauer Associates
- Nei (1987) "Molecular Evolutionary Genetics" Columbia Univ. Press
- Hillis, Moritz, and Mable (1996) "Molecular Systematics" Sinauer
- Felsenstein (2003) "Inferring Phylogenies" Sinauer
- Felsenstein (2015) "Systematics and Molecular Evolution: Some history of numerical methods" Lecture at Molecular Evolution Workshop: molevol.mbl.edu/images/e/ed/Felsenstein.15.2.pdf

Finding the best / Estimating trees

- Most strategies to reconstruct evolutionary trees optimize some measure of "goodness"
  - Parsimony methods minimize the number of mutations
  - Distance methods produce trees that match the global distances between the sequences
  - Maximum likelihood methods seek the tree that best fits the data
- What is the "best" method?
  - produces accurate trees with the least data?
  - converges to the correct tree as data increases?
- We cannot know the "correct" tree

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Finding the best / Estimating trees

- An optimality criterion defines how we measure the fit of the data to a given solution
  - parsimony / distance / Maximum likelihood
- Tree searching is a separate step; this is how we search through possible solutions (which we then evaluate with the chosen optimality criterion)
  - Except for Neighbor-Joining and UPGMA, which produce a result based on the search strategy

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Advantages

- Parsimony:
  - Widely applicable to many discrete data types (often used to combine analyses of different data types)
  - Requires no explicit model of evolutionary change
  - Computationally relatively fast
  - Relatively easy interpretation of character change
  - Performs well with many data sets
- Distance:
  - Can be used with pairwise distance data (e.g., non-discrete characters)
  - Can incorporate an explicit model of evolution in estimation of pairwise distances
  - Computationally relatively fast (especially for single-point estimates)
- Likelihood/Bayesian:
  - Fully based on explicit model of evolution
  - Most efficient method under widest set of conditions
  - Consistent (converges on correct answer with increasing data, as long as assumptions are met)
  - Most straight-forward statistical assessment of results; probabilistic assessment of ancestral character states

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Disadvantages:

- Parsimony methods:
  - No explicit model of evolution; often less efficient
  - Nonparametric statistical approaches for assessing results often have poorly understood properties
  - Can provide misleading results under some fairly common conditions
  - Do not provide probablistic assessment of alternative solutions
- Distance methods:
  - Model of evolution applied locally (to pairs of taxa), rather than globally
  - Statistical interpretation not straight-forward
  - Can provide misleading results under some fairly common conditions (but not as sensitive as parsimony)
  - Do not provide probablistic assessment of alternative solutions
- Likelihood/Bayesian:
  - Requires an explicit model of evolution, which may not be realistic or available for some data types
  - Computationally most intense

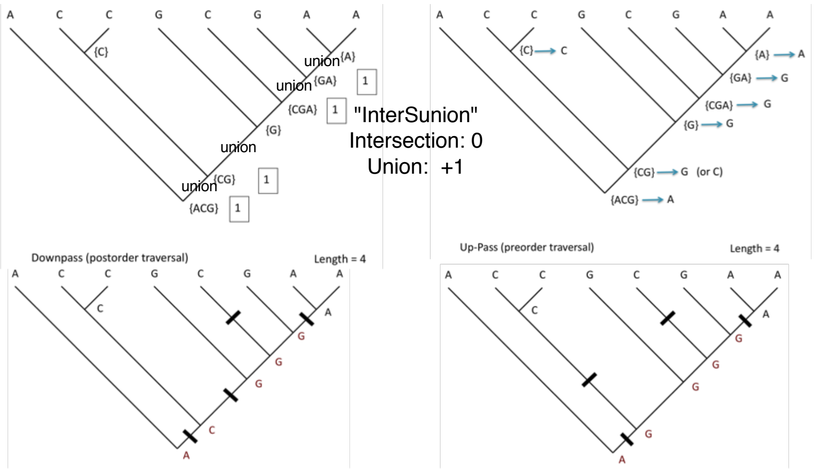
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

The Parsimony Criterion:

- Under the parsimony criterion, the optimal tree (the shortest or minimum length tree) is the one that minimizes the sum of the lengths of all characters in terms of evolutionary steps (a step is a change from one character-state to another).
- For a given tree, find the length of each character, and sum these lengths; this is the tree length.
- The tree with the minimum length is the most parsimonious tree.
- The most parsimonious tree provides the **best fit** of the data set under the parsimony criterion.

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parismony: ancestral states

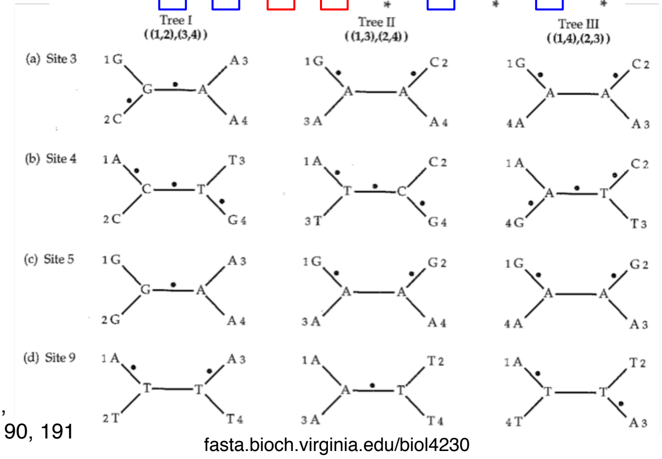


"InterSunion"
Intersection: 0
Union: +1

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parsimony – Informative sites

Sequence	Site 1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T
					*		*		*



Graur and Li,
Chap 5, pp 190, 191

Parsimony – Informative sites

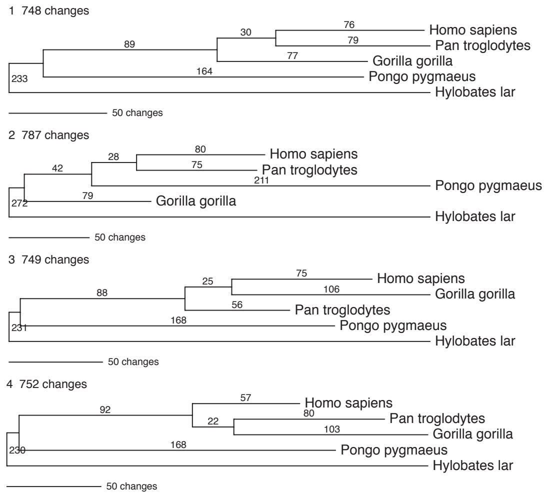
Paup analysis of 3000 sites from primate mitochondrial D-loop

```
 Character-status summary:
    13203 characters are excluded  (selected 1-3000)
    Of the remaining 3000 included characters:
      All characters are of type 'unord'
      All characters have equal weight
      2397 characters are constant
      431 variable characters are parsimony-uninformative
      Number of (included) parsimony-informative characters = 172
  Gaps are treated as "missing"
  Multistate taxa interpreted as uncertainty

Tree #     1   2   3   4   5   6   7   8   9  10
Length   748 787 749 752 792 787 792 789 789 789
```

172/3000 = 5.7% of data used to build tree

Parsimony – Informative sites

Paup analysis of 3000 sites from primate mitochondrial D-loop



Parsimony – Informative sites

```
 Character-status summary:
    13203 characters are excluded  (selected 1-3000)
    Of the remaining 3000 included characters:
      All characters are of type 'unord'
      All characters have equal weight
      2397 characters are constant
      431 variable characters are parsimony-uninformative
      Number of (included) parsimony-informative characters = 172
  Gaps are treated as "missing"
  Multistate taxa interpreted as uncertainty

Tree #    1   2   3   4   5   6   7   8   9  10
Length  748 787 749 752 792 787 792 789 789 789
```

172/3000 = 5.7% of data used to build tree

94.3% of data "not informative"
95% identical??
25% identical??

Distance Methods

- Parsimony methods *ONLY* see informative sites
  - often 20% of the data or less
  - uninformative sites have information:
    - uninformative because no change (short branches)
    - uninformative because lots of change (long branches)
- Distance methods look at *ALL* the data
  - but simply construct pairwise distances
  - must use "transformed" distance, which requires model
  - trees that match pairwise distances need not have a possible evolutionary path

Pairwise Distances

- Distances summarize character differences between objects (terminals, taxa).
- Pairwise distances are computationally quick to calculate.
- Character differences cannot be recovered from distances, because different combinations of character states can yield the same distance (no ancestral states).
- Characters cannot be compared individually, as in discrete character analyses.
- The distances in a matrix are not independent of each other, and errors are often compounded in fitting distances to a tree.

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

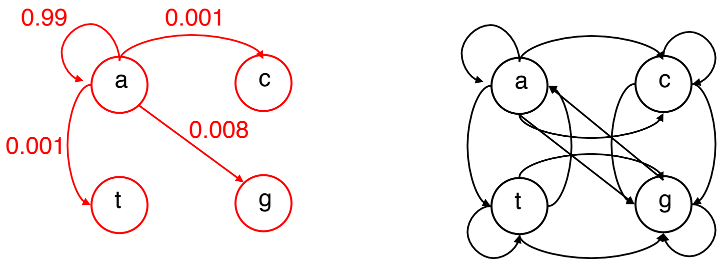
Distance Methods

	Characters (sites)				
Taxa	1	2	3	4	5
one	A	G	C	G	A
two	A	G	C	G	T
three	C	T	C	G	T
four	C	T	C	A	A

	proportional distances			
	one	two	three	four
one	–	0.2	0.6	0.6
two		–	0.4	0.8
three			–	0.4
four				–

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

DNA transition probabilities – 1 PAM



	a	c	g	t	
a	0.99	0.001	0.008	0.001	= 1.0
c	0.001	0.99	0.001	0.008	= 1.0
g	0.008	0.001	0.99	0.001	= 1.0
t	0.001	0.008	0.001	0.99	= 1.0

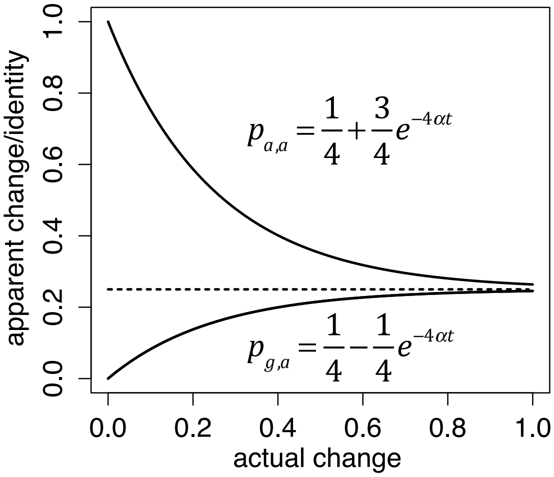
Matrix multiples

can also be calculated from
"instantaneous rate matrix Q"
p(t) = exp(t*Q)

```
M^2={     PAM 2
{0.980, 0.002, 0.016, 0.002},
{0.002, 0.980, 0.002, 0.016},
{0.016, 0.002, 0.980, 0.002},
{0.002, 0.016, 0.002, 0.980}}

M^5={    PAM 5
{0.952, 0.005, 0.038, 0.005},
{0.005, 0.951, 0.005, 0.038},
{0.038, 0.005, 0.952, 0.005},
{0.005, 0.038, 0.005, 0.952}}

M^10={  PAM 10
{0.907, 0.010, 0.073, 0.010},
{0.010, 0.907, 0.010, 0.073},
{0.073, 0.010, 0.907, 0.010},
{0.010, 0.073, 0.010, 0.907}}
```

```
M^100={           PAM 100
{0.499, 0.083, 0.336, 0.083},
{0.083, 0.499, 0.083, 0.336},
{0.336, 0.083, 0.499, 0.083},
{0.083, 0.336, 0.083, 0.499}}

M^1000={          PAM 1000
{0.255, 0.245, 0.255, 0.245},
{0.245, 0.255, 0.245, 0.255},
{0.255, 0.245, 0.255, 0.245},
{0.245, 0.255, 0.245, 0.255}}
```

From differences to distance:
the Jukes-Cantor correction (DNA)

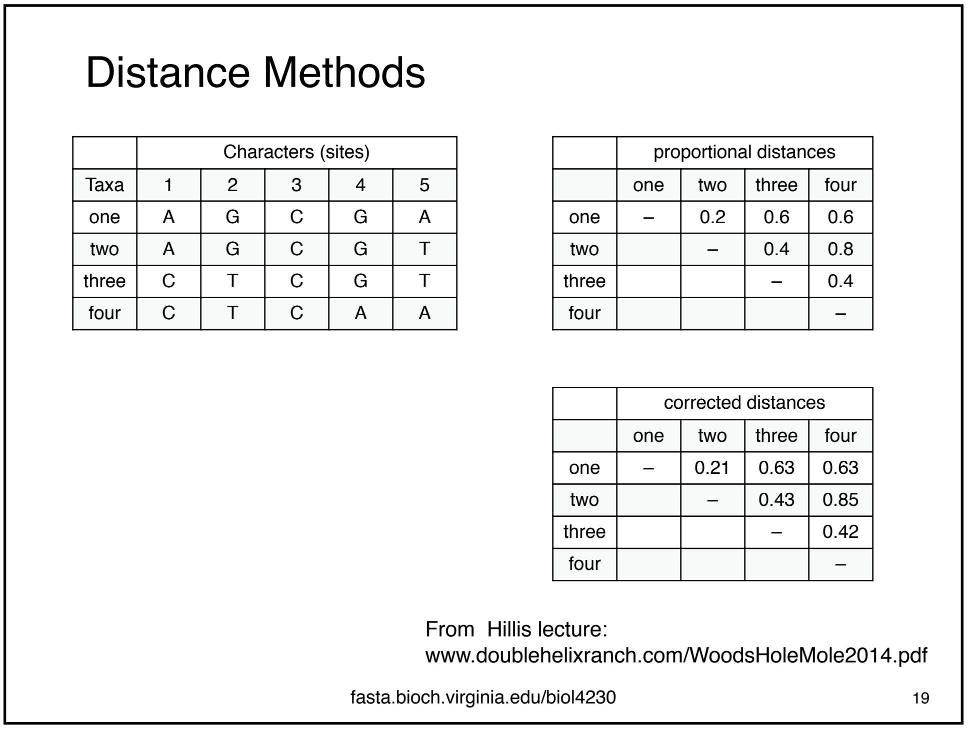


$$p_{a,a} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

A➔A

$$p_{g,a} = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

G➔A



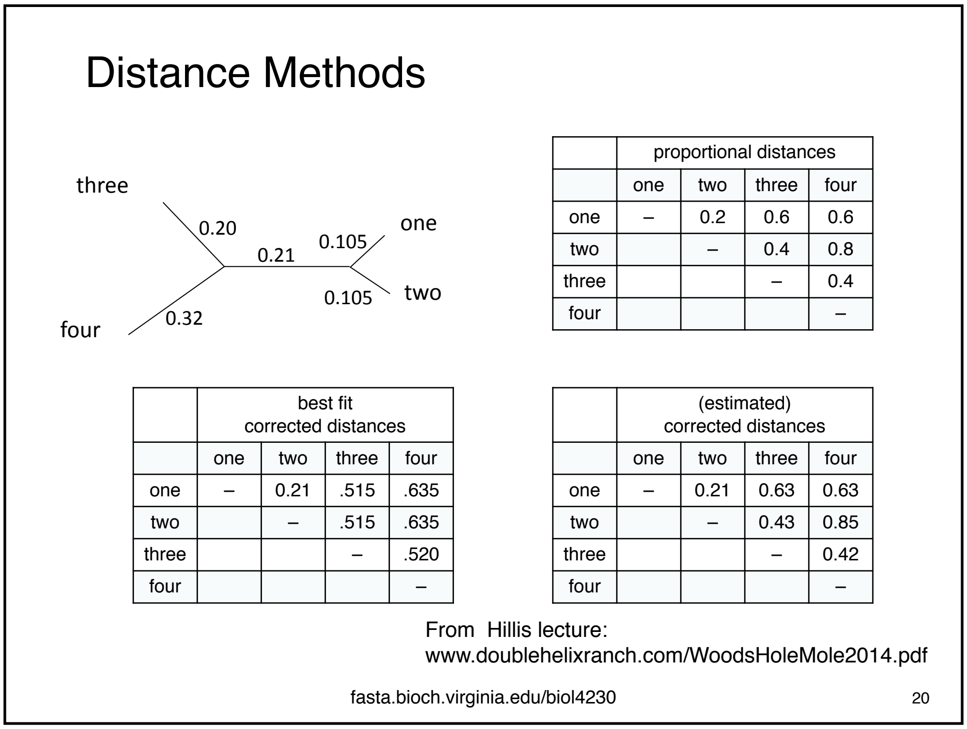
Distance Methods

	Characters (sites)				
Taxa	1	2	3	4	5
one	A	G	C	G	A
two	A	G	C	G	T
three	C	T	C	G	T
four	C	T	C	A	A

	proportional distances			
	one	two	three	four
one	–	0.2	0.6	0.6
two		–	0.4	0.8
three			–	0.4
four				–

	corrected distances			
	one	two	three	four
one	–	0.21	0.63	0.63
two		–	0.43	0.85
three			–	0.42
four				–

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf



Distance Methods

	proportional distances			
	one	two	three	four
one	–	0.2	0.6	0.6
two		–	0.4	0.8
three			–	0.4
four				–

	best fit corrected distances			
	one	two	three	four
one	–	0.21	.515	.635
two		–	.515	.635
three			–	.520
four				–

	(estimated) corrected distances			
	one	two	three	four
one	–	0.21	0.63	0.63
two		–	0.43	0.85
three			–	0.42
four				–

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Pairwise distances: Optimality Criteria

- Two commonly used objective functions:
  - Fitch-Margoliash
  - Minimum Evolution
- The general strategy is to find a set of patristic distances (path-length distances) for the branches that minimize the difference between the evolutionary distances and the patristic distances.

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Pairwise distances:

- Fitch-Margoliash (minimize error):

$$Fit = \sum_{j=2}^{n} \sum_{i=1}^{j} \omega_{i,j} \, | d_{i,j} - p_{i,j} |^{\alpha}$$

i = taxon i
j = taxon j, up to n
d = evolutionary distance (from data)
p = patristic or tree distance (from fit)
ω = weight
Exponent α: 2 = least squares
1 = absolute difference

Common weights:
$\omega_{ij} = 1$
$\omega_{ij} = 1/d_{ij}$
$\omega_{ij} = 1/d^2_{ij}$

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Pairwise distances:

- Minimum evolution (minimize tree length):

$$Fit = \sum_{j=2}^{n} \sum_{i=1}^{j} \omega_{i,j} \, | d_{i,j} - p_{i,j} |^{\alpha}$$

1. Use ω=1and α=2 to fit branch lengths
2. Pick the tree that minimizes the sum of the branches (Length of tree, similar to parsimony)

$$L = \sum_{i=1}^{2n-3} l_i$$

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Distance:

Paup analysis of 3000 sites from primate mitochondrial D-loop

```
Uncorrected      1        2        3        4        5
Hylobates        -
Human      0.11182        -
Chimp      0.10851  0.05186        -
Gorilla    0.11422  0.06069  0.06136        -
Pongo      0.13056  0.10548  0.10414  0.10901        -


Corrected         1        2        3        4        5
Hylobates         - 0.120941 0.117090 0.123937 0.143651
Human      0.120941        - 0.053528 0.063076 0.113246
Chimp      0.117090 0.053528        - 0.063769 0.111617
Gorilla    0.123937 0.063076 0.063769        - 0.117366
Pongo      0.143651 0.113246 0.111617 0.117366        -
```

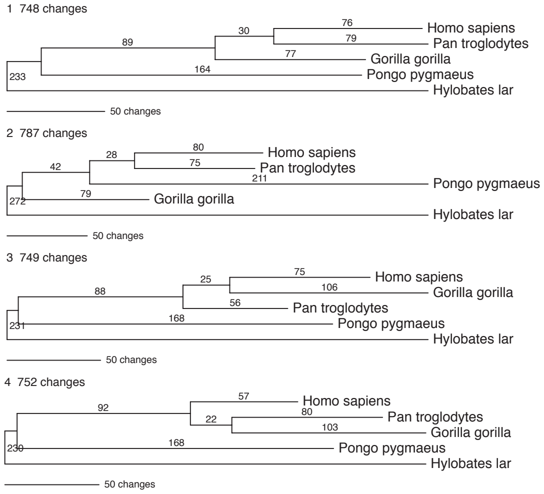
Distance:

Paup analysis of 3000 sites from primate mitochondrial D-loop

```
                          1        2        3        4        5
    1 Hylobates lar       -
    2 Homo sapiens     0.11182        -
    3 Pan troglodytes  0.10851  0.05186        -
    4 Gorilla gorilla  0.11422  0.06069  0.06136        -
    5 Pongo pygmaeus   0.13056  0.10548  0.10414  0.10901        -

Heuristic search settings:
  Optimality criterion = distance (unweighted least squares (power=0))
    Negative branch lengths allowed, but set to zero for tree-score
calculation
    Distance measure = uncorrected ("p")
3000 characters are included
  Starting tree(s) obtained via neighbor-joining
  Branch-swapping algorithm: tree-bisection-reconnection (TBR) with
reconnection limit = 8
    Steepest descent option not in effect
    Saving 5 best trees found by branch-swapping (on best trees only)
Trees are unrooted
Heuristic search completed
  Total number of rearrangements tried = 12
  Score of best tree(s) found = 3.9665e-06 (%SD=1.20072, g%SD=0.11499[k=7])
  Number of trees retained = 5
```

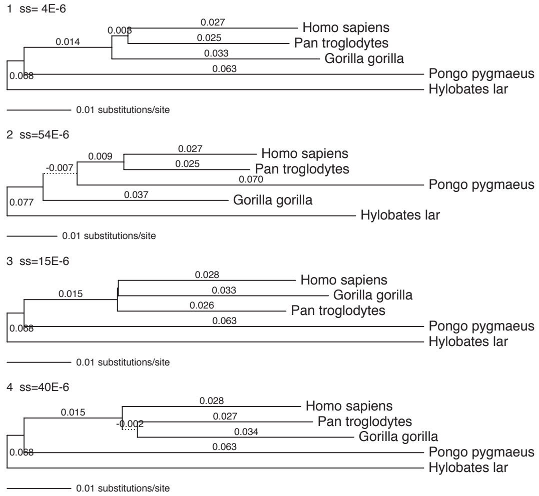
Parsimony – Informative sites

Paup analysis of 3000 sites from primate mitochondrial D-loop



Distance:

Paup analysis of 3000 sites from primate mitochondrial D-loop



Distance defined by an algorithm

- UPGMA – Unweighted Pair Group Mean Arithmetic

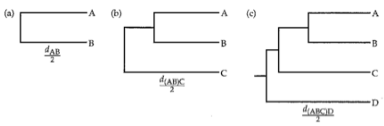


FIGURE 5.10 Diagram illustrating the stepwise construction of a phylogenetic tree for four OTUs by using UPGMA (see text).

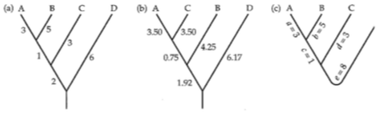
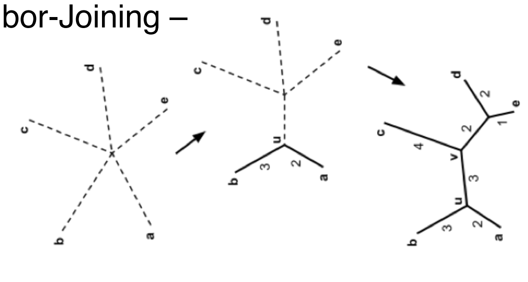


FIGURE 5.11 (a) The true phylogenetic tree. (b) The erroneous phylogenetic tree reconstructed by using UPGMA, which does not take into account the possibility of unequal substitution rates along the different branches. (c) The tree inferred by the transformed distance method. The root must be on the branch connecting OTU D and the node of the common ancestor of OTUs A, B, and C, but its exact location cannot be determined by the transformed distance method.

  - strongly assumes clock-like tree

Li and Graur, p. 184, 185

- Neighbor-Joining –

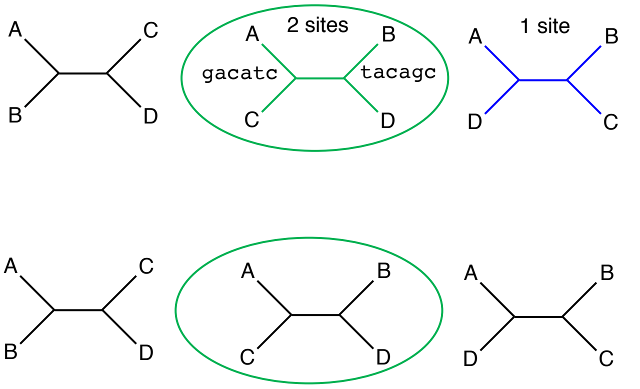


Wikipedia

Parsimony vs Distance – a data set

A: gtgttc
B: taccgt
C: gacatc
D: tagcgc

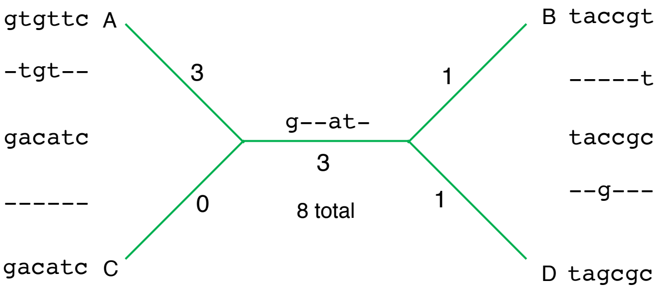
	A	B	C	D
A	0	6	3	4
B		0	4	2
C			0	2
D				0

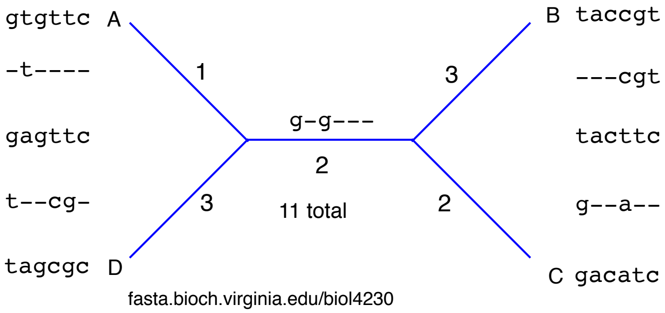


Are there ancestral nodes with correct distances?

Parsimony solutions

A: gtgttc
B: taccgt
C: gacatc
D: tagcgc

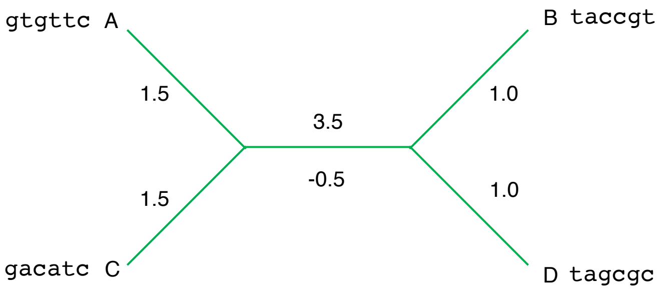




Distance solution

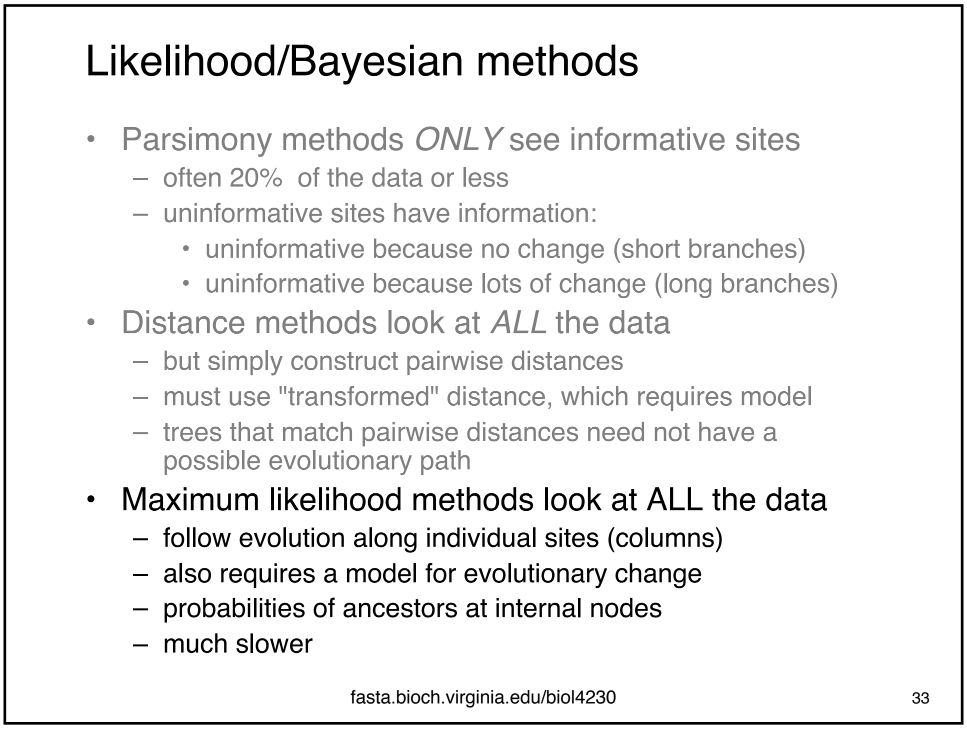
```
A: gtgttc
B: taccgt
C: gacatc
D: tagcgc
```

	A	B	C	D
A	0	6	3	4
B		0	4	2
C			0	2
D				0



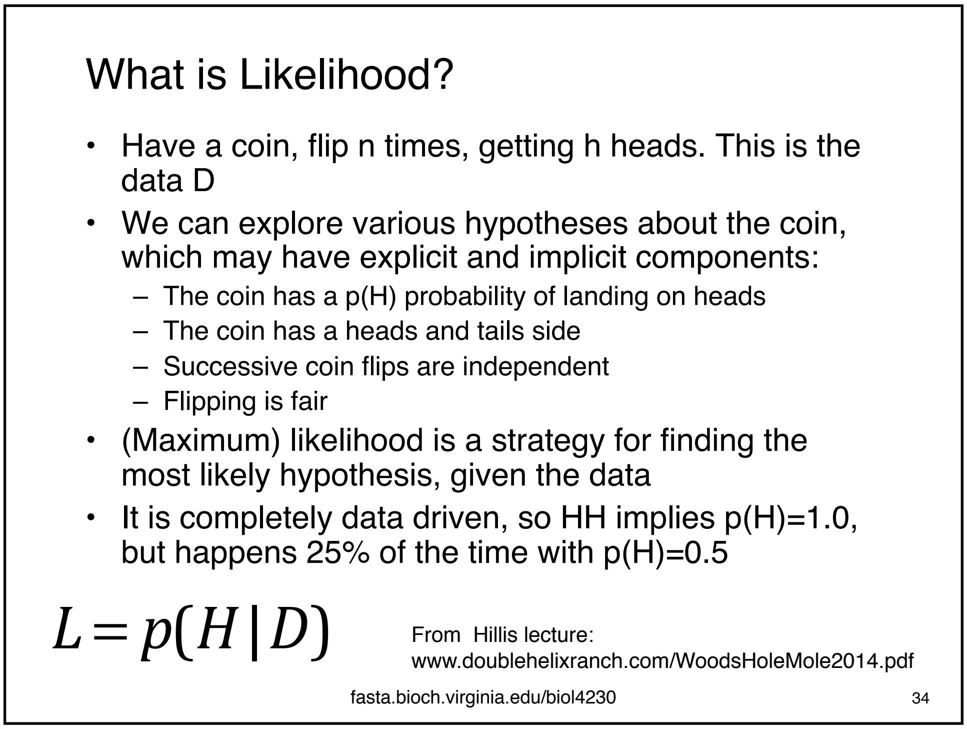
Likelihood/Bayesian methods

- Parsimony methods *ONLY* see informative sites
  - often 20% of the data or less
  - uninformative sites have information:
    - uninformative because no change (short branches)
    - uninformative because lots of change (long branches)
- Distance methods look at *ALL* the data
  - but simply construct pairwise distances
  - must use "transformed" distance, which requires model
  - trees that match pairwise distances need not have a possible evolutionary path
- Maximum likelihood methods look at ALL the data
  - follow evolution along individual sites (columns)
  - also requires a model for evolutionary change
  - probabilities of ancestors at internal nodes
  - much slower



Likelihood/Bayesian methods

- Parsimony methods *ONLY* see informative sites
  - often 20% of the data or less
  - uninformative sites have information:
    - uninformative because no change (short branches)
    - uninformative because lots of change (long branches)
- Distance methods look at *ALL* the data
  - but simply construct pairwise distances
  - must use "transformed" distance, which requires model
  - trees that match pairwise distances need not have a possible evolutionary path
- Maximum likelihood methods look at ALL the data
  - follow evolution along individual sites (columns)
  - also requires a model for evolutionary change
  - probabilities of ancestors at internal nodes
  - much slower



What is Likelihood?

- Have a coin, flip n times, getting h heads. This is the data D
- We can explore various hypotheses about the coin, which may have explicit and implicit components:
  - The coin has a p(H) probability of landing on heads
  - The coin has a heads and tails side
  - Successive coin flips are independent
  - Flipping is fair
- (Maximum) likelihood is a strategy for finding the most likely hypothesis, given the data
- It is completely data driven, so HH implies p(H)=1.0, but happens 25% of the time with p(H)=0.5

$$L = p(H|D)$$

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Coin flipping

- The likelihood (*L*) is proportional to the probability of observing our data, given our hypothesis:

$$L(H \mid D) \propto P(D \mid H)$$

- The probability of getting the outcome *h* heads on *n* flips is given by the binomial distribution:

$$P(h,n \mid p_h) = \binom{n}{h} (p_h)^h (1-p_h)^{n-h}$$

- The combinatorial term gives the binomial coefficients, for the number of ways to get 4 heads in 10 flips
- We will ignore that term and look at a particular sequence of H's and T's (more like a specific sequence of nucleotides)

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Coin flipping

- Let's apply likelihood to specific data:
  - Dataset 1: A particular run of tosses

    H T T H T T H T T H
- Assume a hypothesis, $p_h = 0.5$
- This gives a likelihood score of:

$$L(p_h = 0.5 \mid obs) = (0.5)^4 (1-0.5)^6 = 0.000976563$$

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

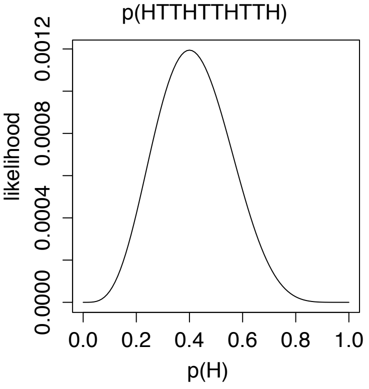
Coin flipping

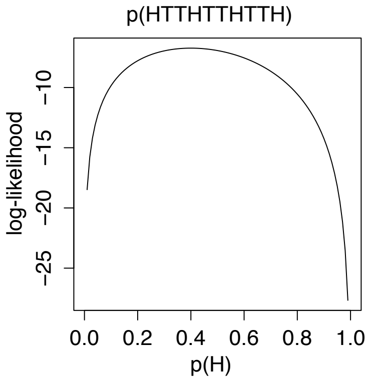
- What does the likelihood score tell us about the likelihood of our hypothesis? In isolation, nothing, because the score is dependent on the particular data set. The score will get smaller as we collect more data (flip the coin more times).
- Only the *relative* likelihood scores for various hypotheses, evaluated using the same data, are useful to us.
- What are some other models?

$$L(p_h = 0.6|obs) = (0.6)^4(0.4)^6 = 0.000530842$$
$$L(p_h = 0.4|obs) = (0.4)^4(0.6)^6 = 0.001194394$$

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

The likelihood surface





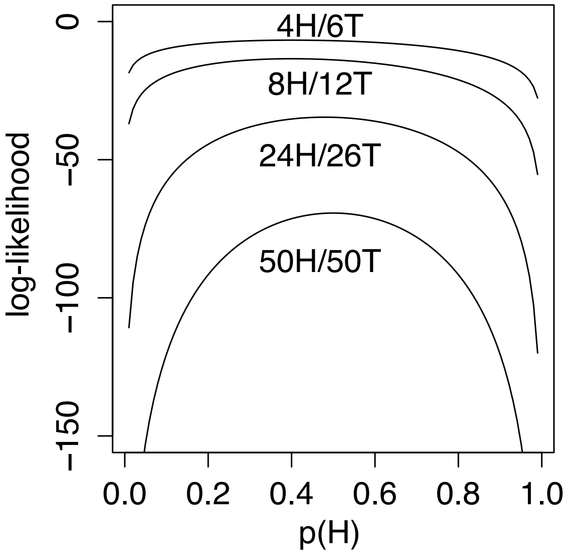
log() is ln()
ln(20)~3

Likelihood

- Likelihood (*H*|*D*) is proportional to *P*(*D*|*H*)
- Components of the hypothesis can be explicit and implicit
- Only relative likelihoods are important in evaluating hypotheses
- The point on the likelihood curve that maximizes the likelihood score (the MLE) is our best estimate given the data at hand
- Likelihood scores shouldn't be compared between datasets
- More data lead to more peaked surfaces (i.e., better ability to discriminate among hypotheses)

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Likelihood



Likelihood in Phylogenetics

- In phylogenetics, the data are the observed characters (e.g., DNA sequences) as they are distributed across taxa
- The hypothesis consists of the tree topology, a set of specified branch lengths, and an explicit model of character evolution.
- Calculating the likelihood score for a tree requires a very large number of calculations

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

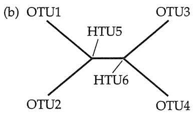
Likelihood in Phylogenetics

(a)

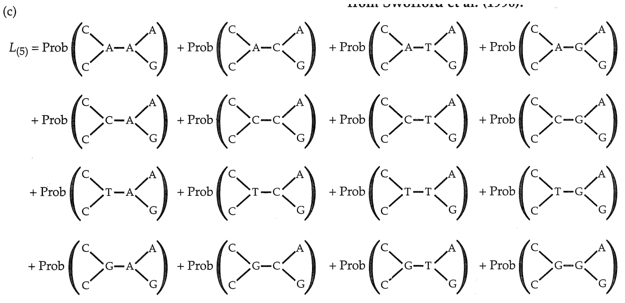
	1	2	3	4	5	6	7	8	9	...*n*
OTU1	A	A	G	A	C	T	T	C	A	...N
OTU2	A	G	C	C	C	T	T	C	T	...N
OTU3	A	G	A	T	A	T	C	C	A	...N
OTU4	A	G	A	G	G	T	C	C	T	...N

$$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \ldots \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$$

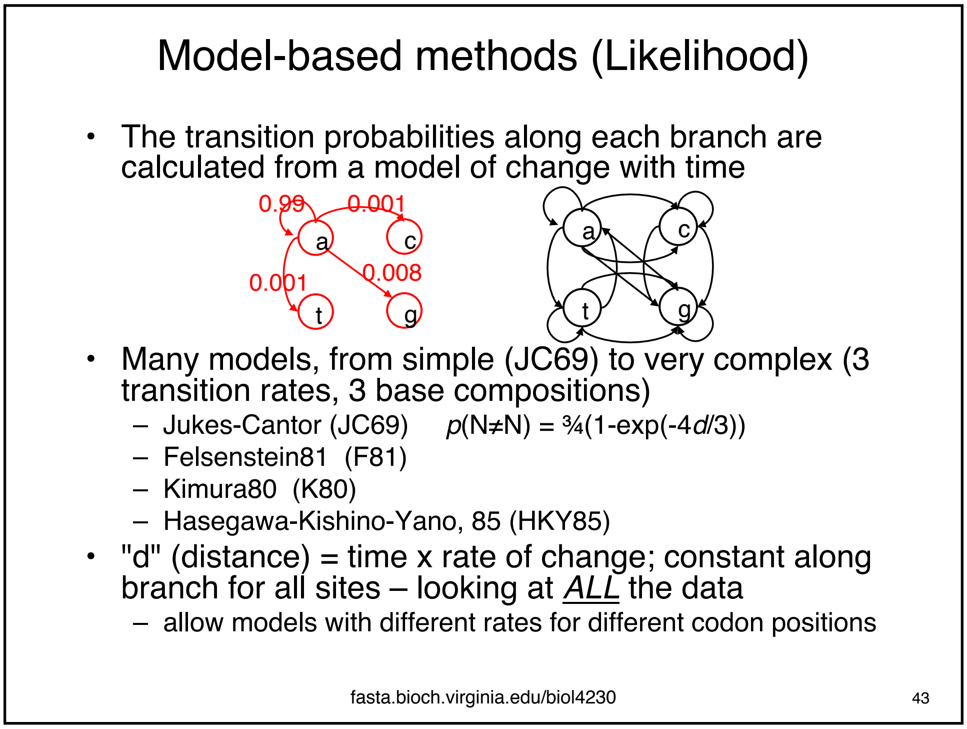
$$\ln(L) = \ln(L_{(1)}) + \ldots + \ln(L_{(n)}) = \sum_{i=1}^{n} \ln(L_{(i)})$$



from Swofford et al. (1996).



- One tree topology 16 ancestral states at HTU5/HTU6 (4x4)
- What about branch lengths?

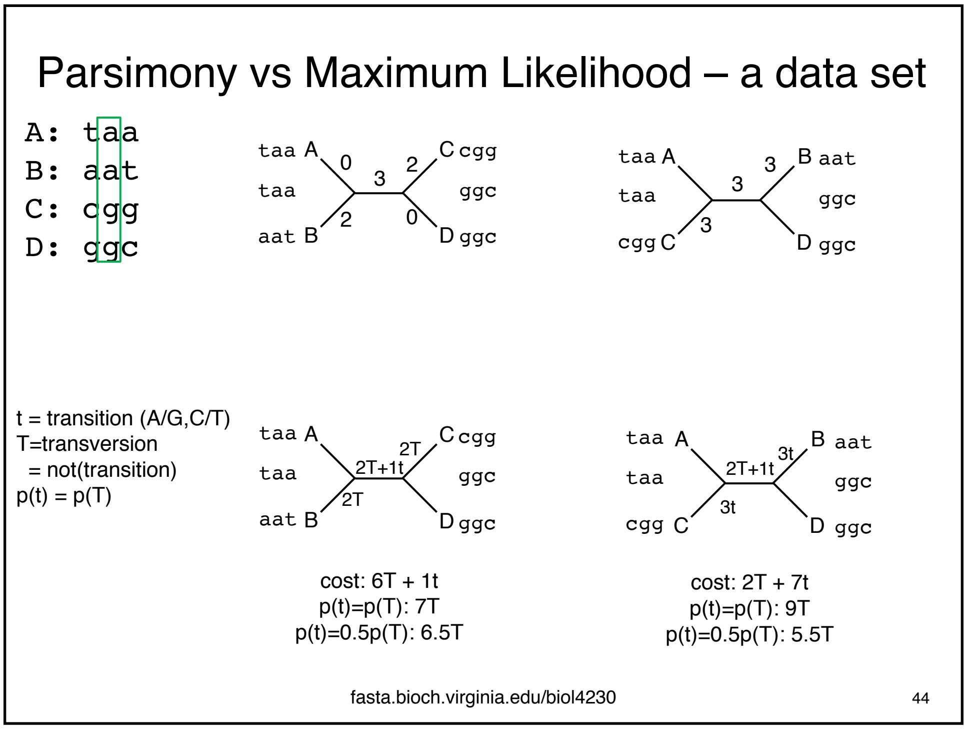


Model-based methods (Likelihood)

- The transition probabilities along each branch are calculated from a model of change with time

0.99
0.001
a
c
0.008
0.001
t
g
a
c
t
g

- Many models, from simple (JC69) to very complex (3 transition rates, 3 base compositions)
  - Jukes-Cantor (JC69) *p*(N≠N) = ¾(1-exp(-4*d*/3))
  - Felsenstein81 (F81)
  - Kimura80 (K80)
  - Hasegawa-Kishino-Yano, 85 (HKY85)
- "d" (distance) = time x rate of change; constant along branch for all sites – looking at *ALL* the data
  - allow models with different rates for different codon positions



Parsimony vs Maximum Likelihood – a data set

```
A: taa
B: aat
C: cgg
D: ggc
```

taa A
taa
aat B
0
2
3
2
0
C cgg
ggc
D ggc
taa A
taa
cgg C
3
3
3
B aat
ggc
D ggc
taa A
taa
aat B
2T+1t
2T
2T
C cgg
ggc
D ggc
taa A
taa
cgg C
2T+1t
3t
3t
B aat
ggc
D ggc

t = transition (A/G,C/T)
T=transversion
= not(transition)
p(t) = p(T)

cost: 6T + 1t
p(t)=p(T): 7T
p(t)=0.5p(T): 6.5T

cost: 2T + 7t
p(t)=p(T): 9T
p(t)=0.5p(T): 5.5T

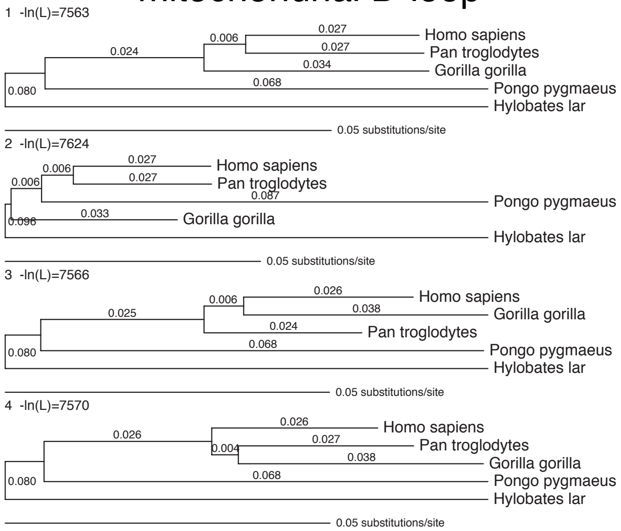
Maximum Likelihood

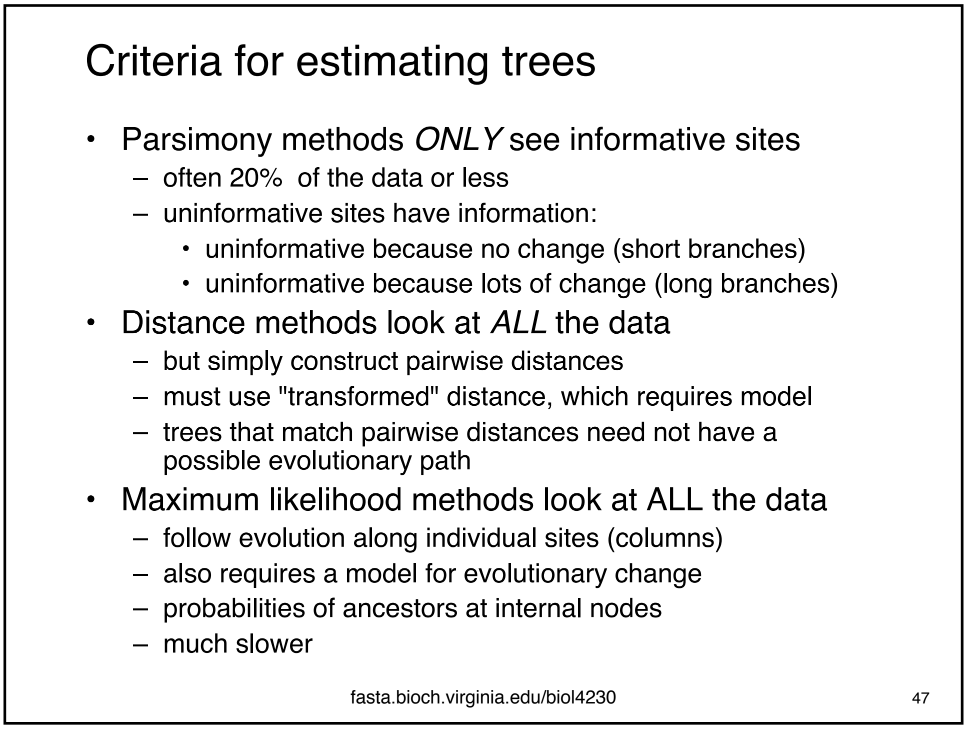
Paup analysis of 3000 sites from primate mitochondrial D-loop

```
3000 characters are included
  Likelihood settings:
    Current model:
                         Data type = nucleotide
                Substitution types = 2 (HKY85 variant)
                       Ti/tv ratio = 2
       State frequencies = empirical: A=0.33701 C=0.27103 G=0.17279 T=0.21917
        Proportion of invariable sites = none
               Rates at variable sites = equal
                  Model correspondence = HKY85
    Number of distinct data patterns under this model = 140
    Molecular clock not enforced
    Starting branch lengths obtained using Rogers-Swofford approximation method
    Branch-length optimization = one-dimensional Newton-Raphson Likelihood
calculations performed in single precision
    Vector processing enabled
    Conditional-likelihood rescaling threshold = 1e-20
    Using 1 thread on 4 physical (8 logical) processors

Tree             1           2           3           4           5
------------------------------------------------------------------
-ln L     7563.309    7614.123    7566.153    7570.346    7614.714
```

Maximum Likelihood

Paup analysis of 3000 sites from primate mitochondrial D-loop





Criteria for estimating trees

- Parsimony methods *ONLY* see informative sites
  - often 20% of the data or less
  - uninformative sites have information:
    - uninformative because no change (short branches)
    - uninformative because lots of change (long branches)
- Distance methods look at *ALL* the data
  - but simply construct pairwise distances
  - must use "transformed" distance, which requires model
  - trees that match pairwise distances need not have a possible evolutionary path
- Maximum likelihood methods look at ALL the data
  - follow evolution along individual sites (columns)
  - also requires a model for evolutionary change
  - probabilities of ancestors at internal nodes
  - much slower