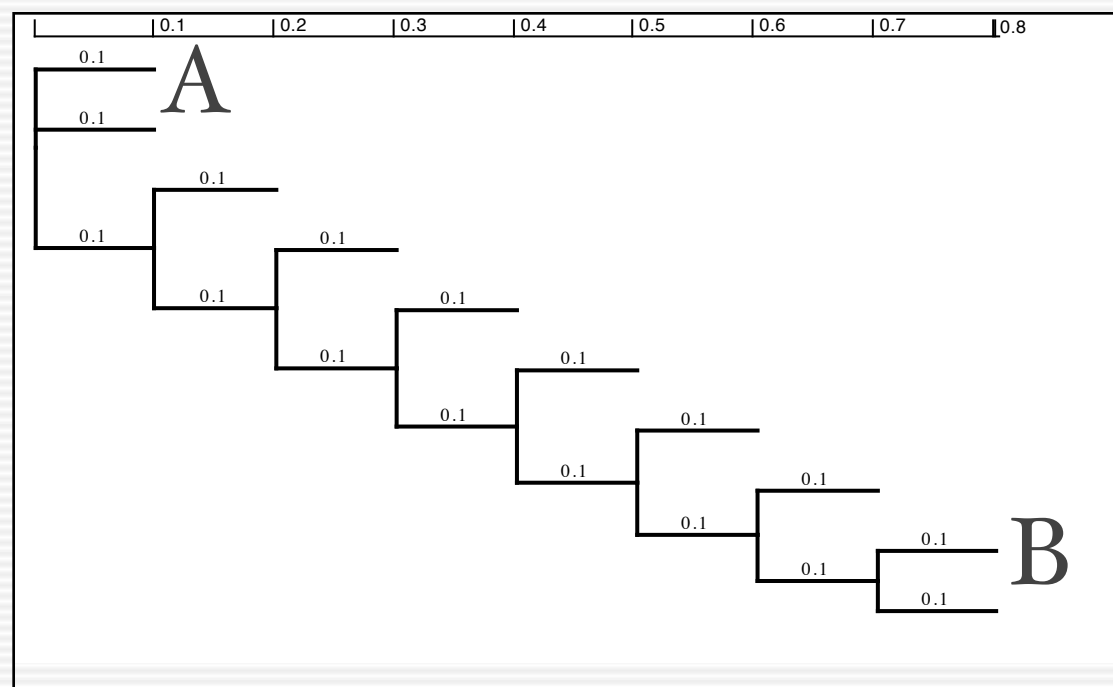


MOLECULAR EVOLUTION AND
PHYLOGENETICS:
AN INTRODUCTION TO
PROBABILISTIC MODELS

COMPUTING DISTANCES BETWEEN SEQUENCES

- **Hamming or p-distance** is the most obvious way to compute distances between two aligned homologous sequences
- p-distances are very simple, but make many hidden assumptions, all of which are violated by biological data. Generally, they work reasonably well only for very closely related sequences.
- In order to reconstruct trees from distance matrices, accurate estimation of **large** distances is necessary as well.



- Each branch is short (0.1)
- But the distance between sequences A and B is actually quite large: 0.9

MULTIPLE HITS AND REVERSIONS

LOW DIVERGENCE

A	T	G	A	A	A	G	C	G	A
A	T	G	A	G	A	G	T	G	A

Substitutions = 2
 $p = 0.2$

HIGH DIVERGENCE

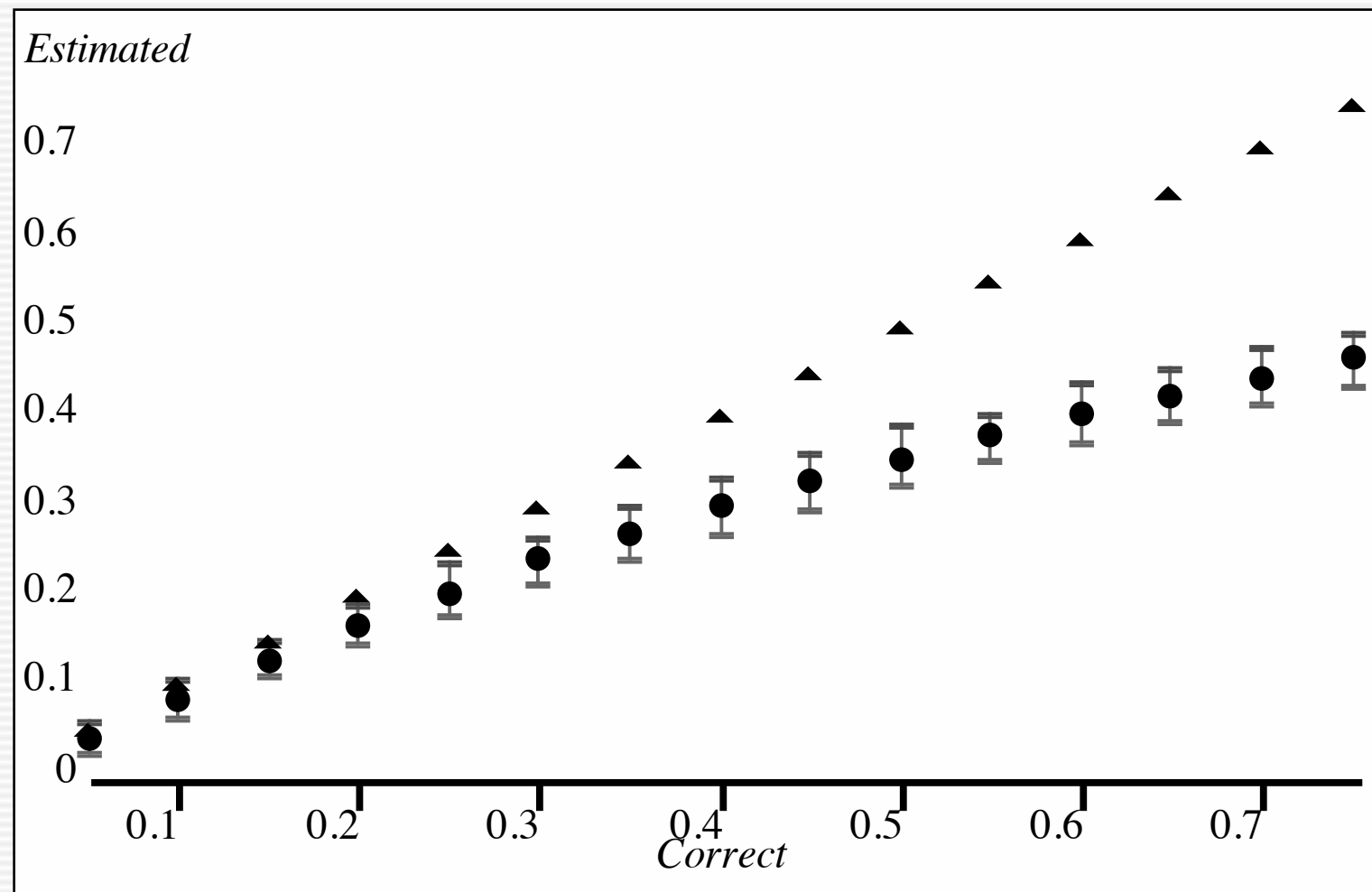
A	T	G	A	A	A	G	C	G	A
				T	C				
A	G	T	A	G	A	G	T	G	A

Substitutions = 7
 $p = 0.4$

Multiple hits

Reversion

EFFECT ON DISTANCE ESTIMATES.



- Simulated 100 replicates of 1000 nucleotide long sequences for various divergence levels (substitutions/site)
- Plotted 'true' divergence vs that estimated by p-distance.
- Even for divergence of 0.25 (1/4 sites have mutation on average), p distance already significantly underestimates the true level: 0.2125 (0.19-0.241 95% range)
- Underestimation becomes progressively worse for larger divergence levels.

JUKES-CANTOR 1969

- The idea is to model substitutions at a site using a **Markov Process**.
- Very much like a Markov chain, except time is now continuous, instead of being measured in discrete steps.
- $X(t)$ defines the *probability distribution* that the observed quantity follows at time $t \geq 0$.
- Markov property (memoryless process):

$$\Pr\{X(t) = x_0 | X(t_1) = x_1, \dots, X(t_n) = x_n, t > t_1 > \dots > t_n \geq 0\} = \Pr\{X(t) = x_0 | X(t_1) = x_1\}$$

MARKOV PROCESSES

- To completely define a Markov process, we need to specify the **transition probability function**: given that the process is in state A at time $\mathbf{u}$, what is the probability that it will be in state B at a later time, $\mathbf{u}+\mathbf{t}$?

- Often written as a matrix, $\mathbf{T}(\mathbf{u},\mathbf{t})$:

$$T(u, t)_{AB} = \Pr\{X(u + t) = B | X(u) = A\}$$

- If one further assumes that the process is homogeneous, i.e. $\mathbf{T}(\mathbf{u},\mathbf{t})$ does not depend on $\mathbf{u}$, then

$$T(t)_{AB} = \Pr\{X(t) = B | X(0) = A\}$$

MARKOV PROCESSES (CONT'D)

- In the homogeneous case, it is easier to define the process in terms of its **rate matrix Q**:

$$Q = \lim_{t \downarrow 0} \frac{T(t) - I}{t}$$

- Given **Q**, it can be shown that for $t \geq 0$,

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

- where the matrix exponential is defined by the standard Taylor series

$$\exp Qt = I + Qt + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \dots$$

- There are abundant numerical algorithms that compute the matrix exponential in $O(C^3)$ time, where C is the dimension of the rate matrix.

HOW DOES THIS RELATE TO GENETIC DISTANCES?

- Consider $\mathbf{t}$ as the evolutionary time for a mutational process that runs at a constant mutation rate $\mathbf{r}$. Divergence is then obtained as $\mathbf{d} = \mathbf{r} \times \mathbf{t}$
- The advantage of using a Markov process is that it automatically computes the probability of all possible paths from A to B over time $\mathbf{t}$, whereas p-distance only considers the direct A to B path.
- The optimal $\mathbf{d}$ can be inferred from the data!

JUKES CANTOR ('69) DISTANCE: JC69

- The Markov process assumes that all four bases are equally probable and that nucleotides mutate to other nucleotides with equal rates.
- Diagonal rates are defined by the requirement that the *transition* matrix forms a valid probability distribution in each row: for this to hold, each row in the *rate* matrix must sum to 0.

Rate matrix Q

From↓ To →	A	C	G	T
A	-0.75	0.25	0.25	0.25
C	0.25	-0.75	0.25	0.25
G	0.25	0.25	-0.75	0.25
T	0.25	0.25	0.25	-0.75

Transition matrix T (t)

From↓ To →	A	C	G	T
A	$\frac{1}{4}(1 + 3e^{-t})$	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 - e^{-t})$
C	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 + 3e^{-t})$	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 - e^{-t})$
G	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 + 3e^{-t})$	$\frac{1}{4}(1 - e^{-t})$
T	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 - e^{-t})$	$\frac{1}{4}(1 + 3e^{-t})$

$T(o)$

From↓ To →	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

$T(o.5)$

From↓ To →	A	C	G	T
A	0.352	0.216	0.216	0.216
C	0.216	0.352	0.216	0.216
G	0.216	0.216	0.352	0.216
T	0.216	0.216	0.216	0.352

$T(o.1)$

From↓ To →	A	C	G	T
A	0.753	0.082	0.082	0.082
C	0.082	0.753	0.082	0.082
G	0.082	0.082	0.753	0.082
T	0.082	0.082	0.082	0.753

$T(\infty)$

From↓ To →	A	C	G	T
A	0.25	0.25	0.25	0.25
C	0.25	0.25	0.25	0.25
G	0.25	0.25	0.25	0.25
T	0.25	0.25	0.25	0.25

ML FITTING JC69

- The objective is to find the optimal $\mathbf{t}$, given the data
- Use the principle of maximal likelihood to select $\mathbf{t}$, which maximizes the probability of observing the alignment given the model

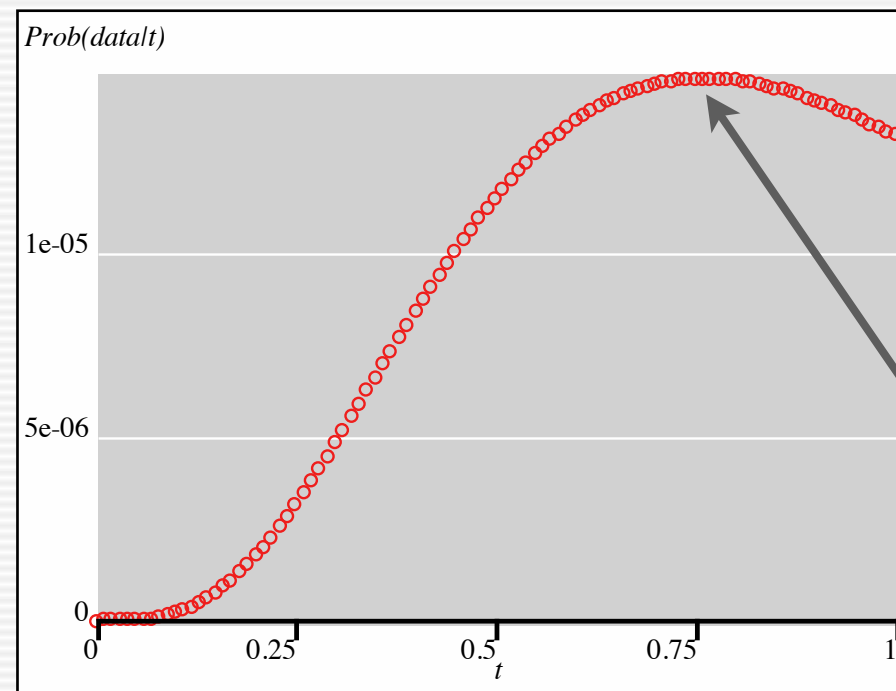
A	T	G	A	A	A	G	C	G	A
A	G	T	A	G	A	G	T	G	A

Independent sites

Simplify...

$$Pr\{data|t\} = \left(\frac{1}{4} [1 + 3e^{-t}]\right)^6 \left(\frac{1}{4} [1 - e^{-t}]\right)^4$$

$$Pr\{data|t\} = Pr\{A \rightarrow A|t\}^4 \times Pr\{T \rightarrow G|t\} \times Pr\{G \rightarrow T|t\} \times Pr\{A \rightarrow G|t\} \times Pr\{G \rightarrow G|t\}^2 \times Pr\{C \rightarrow T|t\}$$



$t=0.76214$

ML FITTING (CON'T)

- For any pair of sequences with $\mathbf{n}$ sites, the JC69 function will only depend on the number of matches ($\mathbf{m}$) and mismatches ($\mathbf{n-m}$):

$$L(t) = Pr\{data|t\} = \frac{1}{4^n} [1 + 3e^{-t}]^m [1 - e^{-t}]^{n-m}$$

- Easier to deal with sums than products: use the log-likelihood function:

$$\log L(t) = -n \log 4 + m \log(1 + 3e^{-t}) + (n - m) \log(1 - e^{-t})$$

- To find the maximum solve (D is the p-distance):

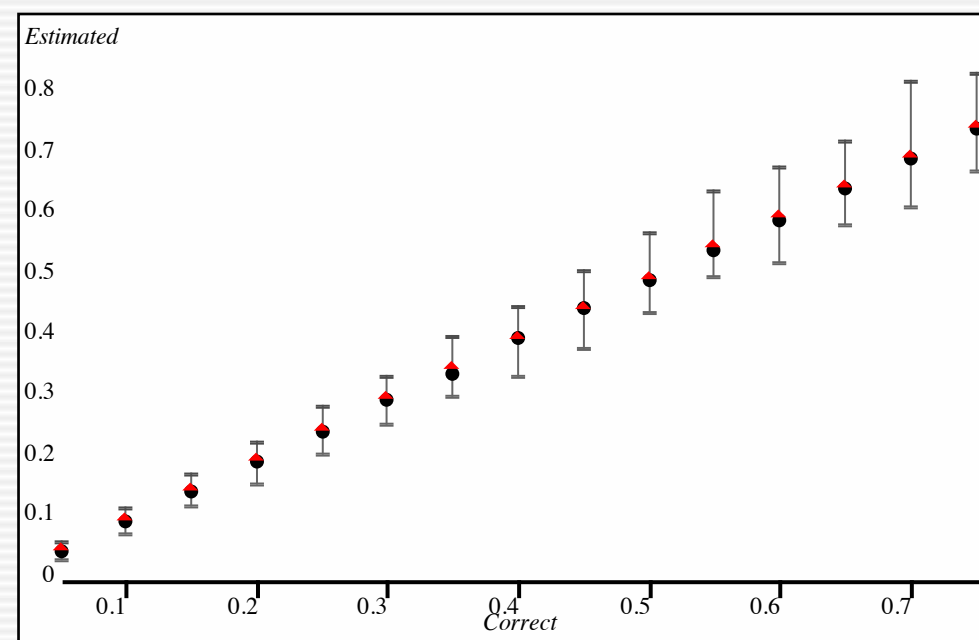
$$\frac{d \log L(t)}{dt} = 0 \implies t = -\log(1 - 4/3D)$$

JC 69 DISTANCE ESTIMATE

- One can show that for JC69, the distance estimator **d** (expected substitutions per site), is related to the time parameter **t**, as **d=3/4 t**

$$d_{JC69} = -\frac{3}{4} \log (1 - 4/3D)$$

- Note that the distance is only defined for divergences up to 0.75 Why does this make sense?



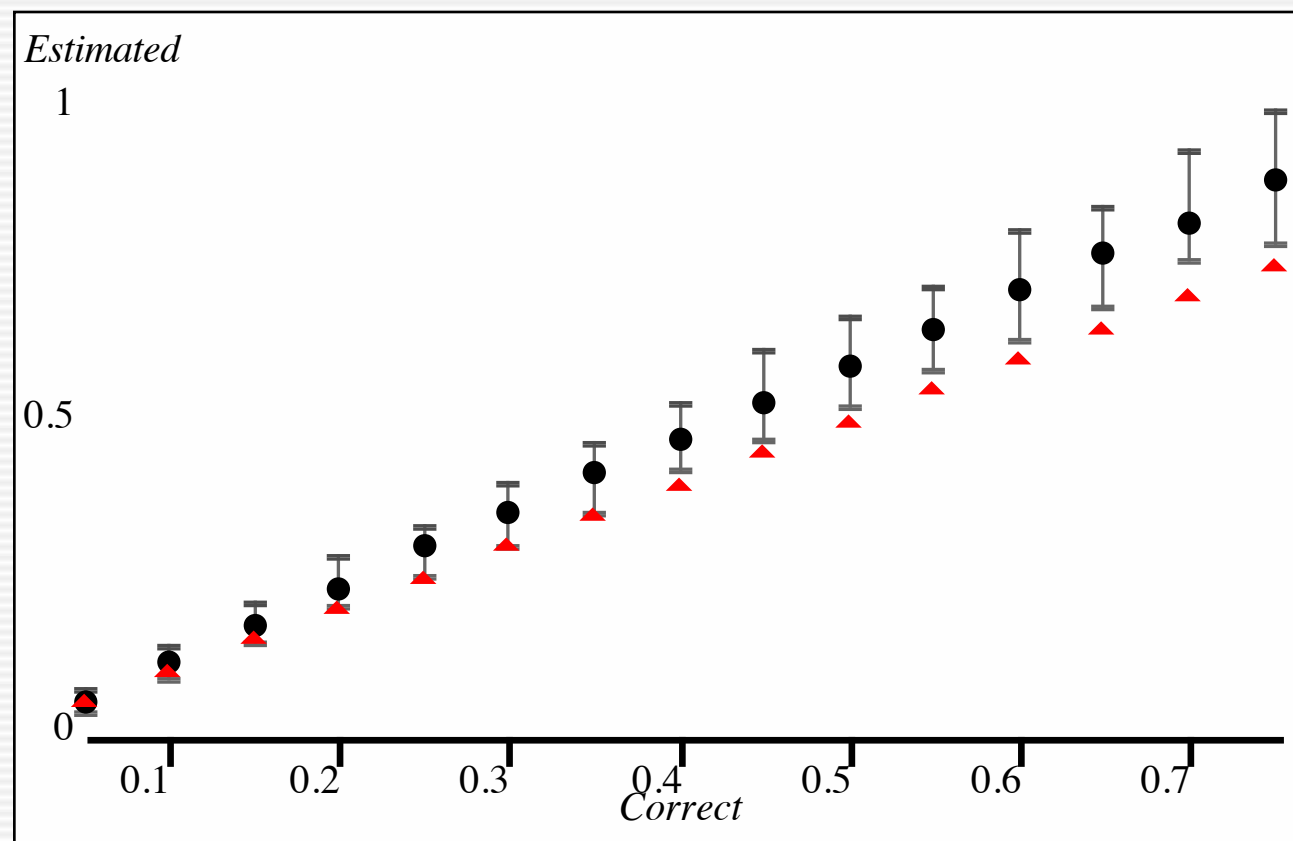
JC69 correction works!

NEXT STEP: FELSENSTEIN 81

- In biological sequences, base frequencies are not equal.
- JC69 can become biased (overestimate distances, see below)
- This is because there are necessarily more substitutions to frequent residues to maintain the frequencies

Base	Frequency
A	0.39
C	0.17
G	0.20
T	0.24

HIV-1 frequencies



F81: rate matrix Q

From ↓ To →	A	C	G	T
A	*	π_C	π_G	π_T
C	π_A	*	π_G	π_T
G	π_A	π_C	*	π_T
T	π_A	π_C	π_G	*

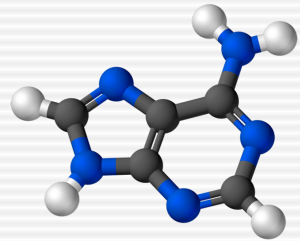
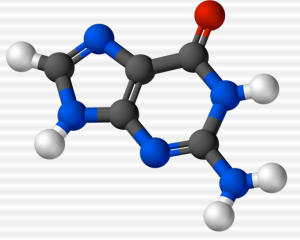
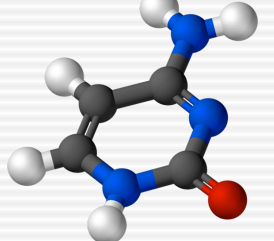
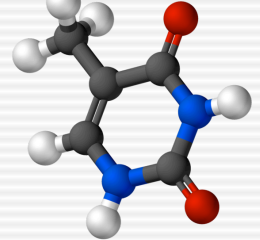
Distance estimator

$$d_{F81} = -F \log(1 - D/F)$$

$$F = 1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2$$

NEXT STEP: DIFFERENT KINDS OF SUBSTITUTIONS

- Nucleotides are split into two chemical groups
 - Adenine and Guanine (purines)
 - Cytosine and Thymine (pyrimidines)
- Substitutions within group (e.g. A to/from G) are called **transitions** and are usually much more frequent than substitutions between groups: **transversions**.

Adenine	
Guanine	
Cytosine	
Thymine	

From↓ To →	A	C	G	T
A		2	20	1
C	2	*	3	18
G	24	1	*	1
T	1	10	0	*

HIV-1 pol example

HKY85: rate matrix Q

From↓ To →	A	C	G	T
A	*	π_C	$K\pi_G$	π_T
C	π_A	*	π_G	$K\pi_T$
G	$K\pi_A$	π_C	*	π_T
T	π_A	$K\pi_C$	π_G	*

κ : transition/transversion
parameter (=1 to obtain F81)

GTR: rate matrix Q

From↓ To →	A	C	G	T
A	*	$r_{AC}\pi_C$	π_G	$r_{AT}\pi_T$
C	$r_{AC}\pi_A$	*	$r_{CG}\pi_G$	$r_{CT}\pi_T$
G	π_A	$r_{CG}\pi_C$	*	$r_{GT}\pi_T$
T	$r_{AT}\pi_A$	$r_{CT}\pi_C$	$r_{GT}\pi_G$	*

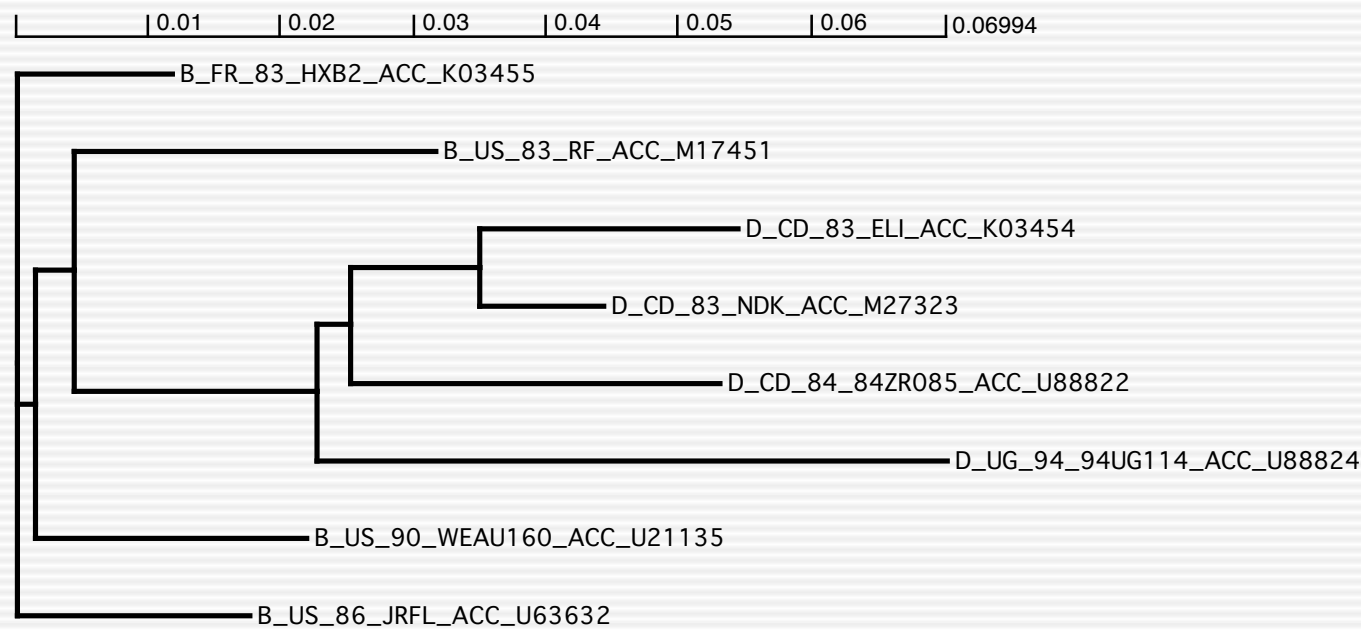
Most general in class: 6
parameters now

Closed form expressions either don't exist (GTR), or are cumbersome (HKY85). Can always estimate numerically.

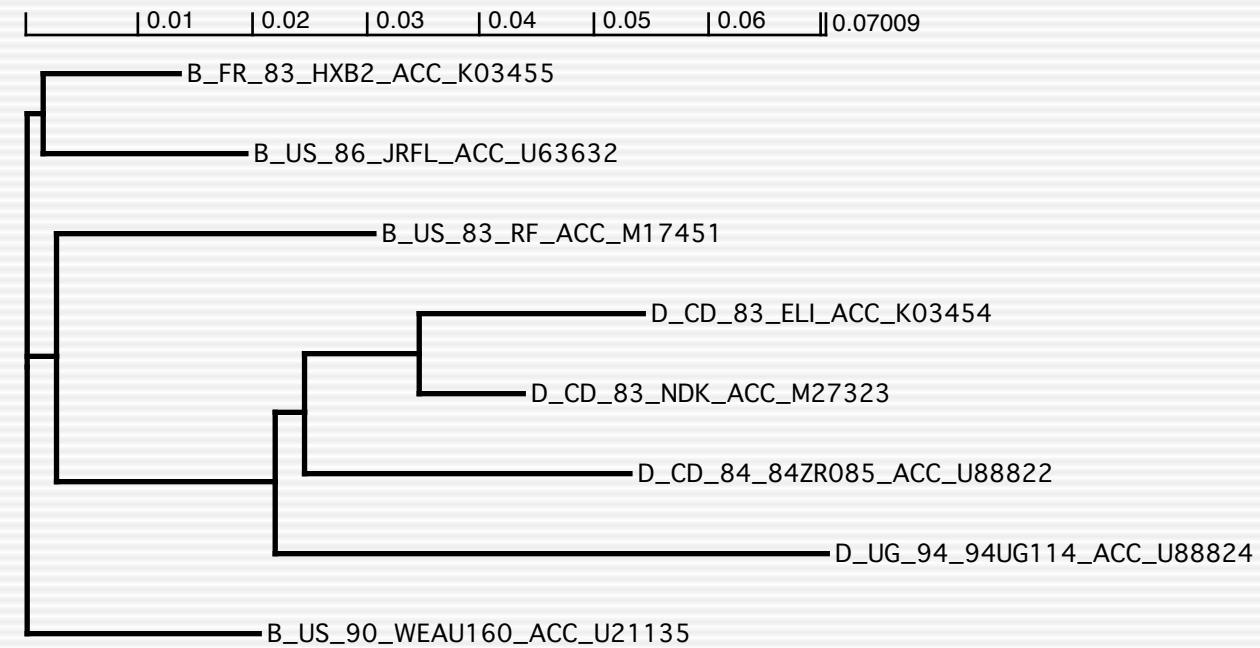
Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA

Masami Hasegawa,¹ Hirohisa Kishino,¹ and Taka-aki Yano²

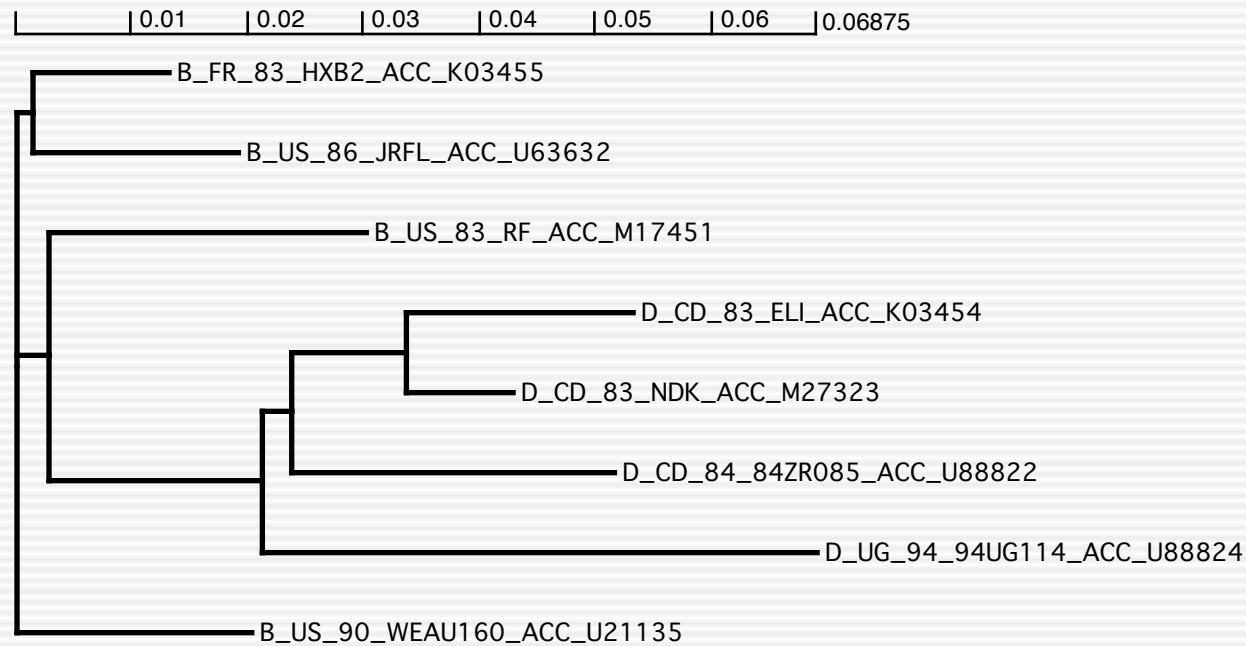
ESTIMATING HIV-1 TREES WITH NJ UNDER DIFFERENT DISTANCES



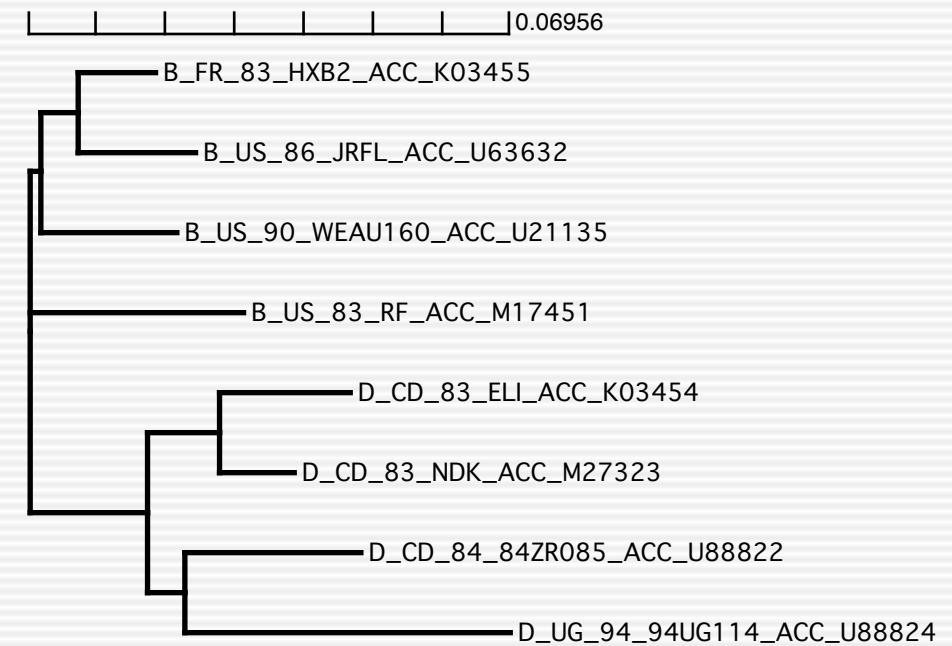
JC69



HKY85

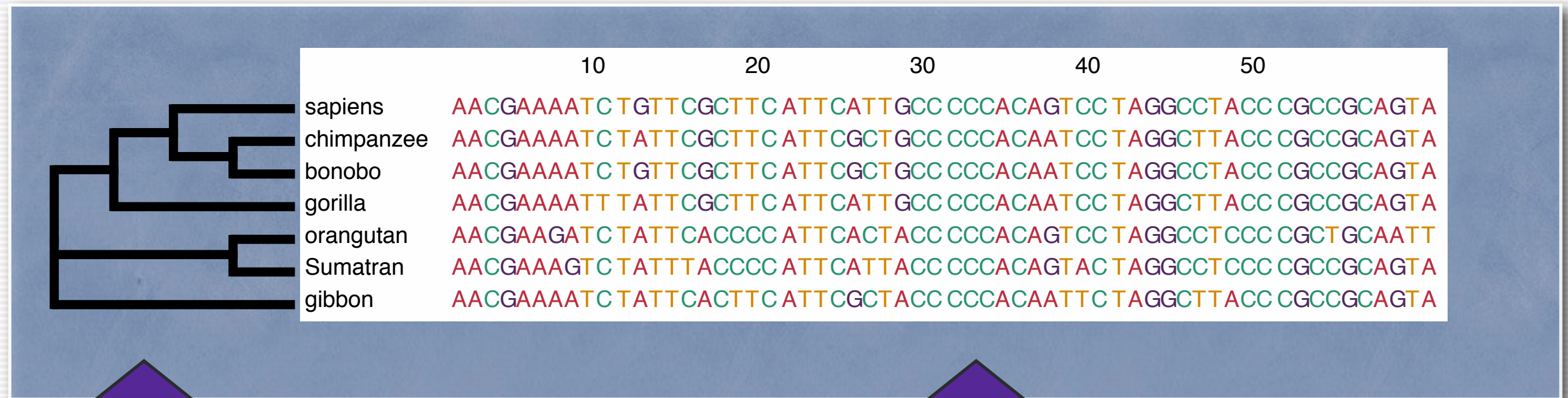


F81



GTR

MAXIMUM LIKELIHOOD SEQUENCE ANALYSIS



Phylogeny

Alignment of homologous sequences (column = site)

Substitution Models

“All models are wrong, but some models are useful” Box (1976)

MAXIMUM LIKELIHOOD MODELS

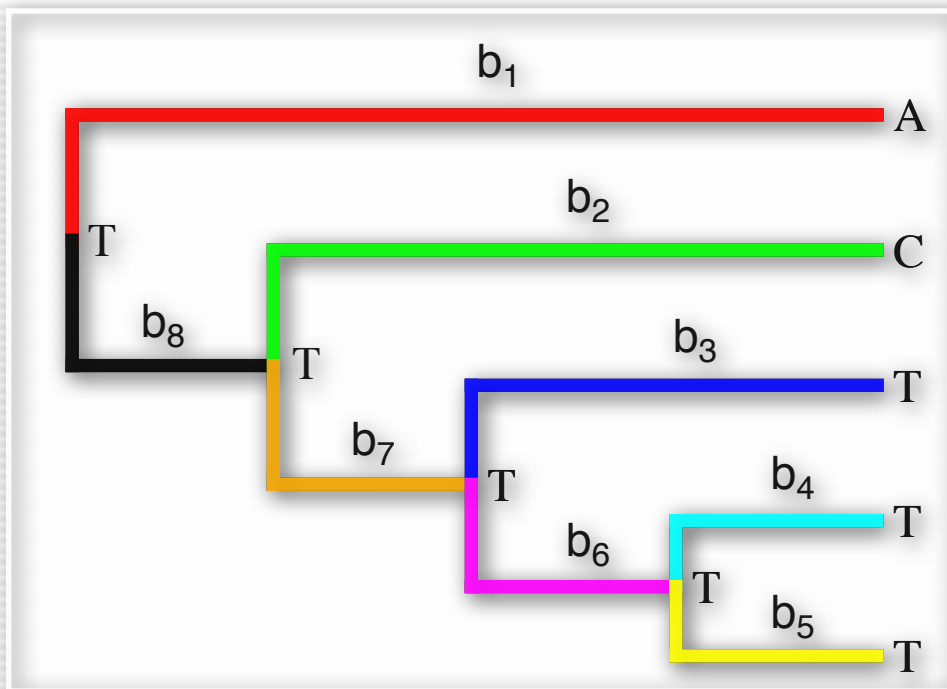
(PLEASE SEE [HTTP://WWW.HYPHY.ORG/DOCS/MAXIMUMLIKELIHOOD.PDF](http://www.hypHY.org/docs/maximumlikelihood.pdf) FOR DETAILS)

- **Define the probability of a point substitution along a branch at a given site:**

- $Q_{x,y}^i(t; \theta) = \text{Pr}_\theta \{x \text{ is replaced with } y \text{ in time } t : x, y \in \mathcal{C}\}$

- **Continuous time Markov chains.**

- **Typically, the models are stationary and time-reversible.**

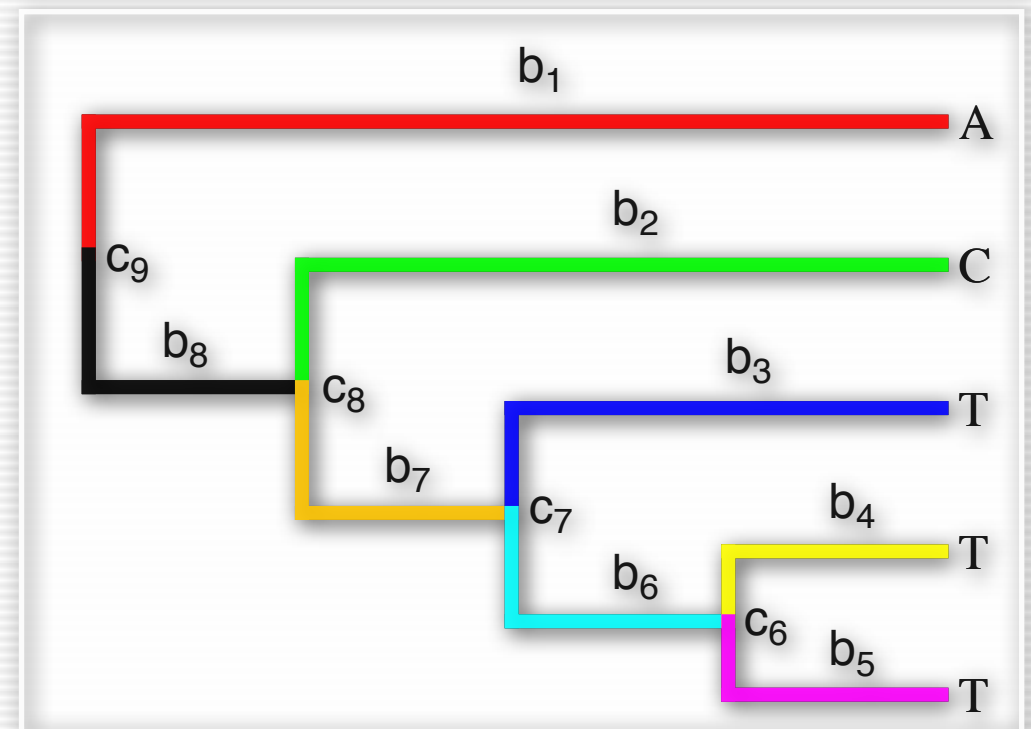


$$L(\mathcal{D}_s; \mathcal{T}, \theta) = Q_{T,A}^1(t_1; \theta) Q_{T,T}^8(t_8; \theta) Q_{T,C}^2(t_2; \theta) Q_{T,T}^7(t_7; \theta) \\ Q_{T,T}^3(t_3; \theta) Q_{T,T}^6(t_6; \theta) Q_{T,T}^4(t_4; \theta) Q_{T,T}^5(t_5; \theta)$$

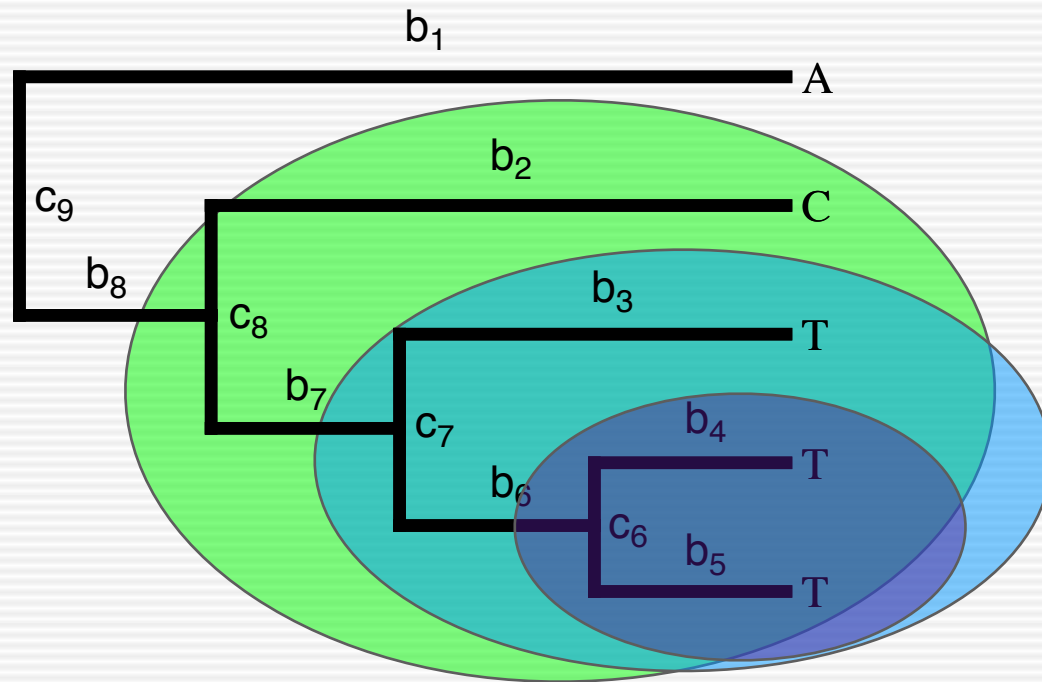
If ancestral states were known.

COMPUTING LIKELIHOOD

- Ancestral states are almost always unknown - must sum over all possible internal node character assignments.
- Computations can be done efficiently, in $O(C^2N)$ time, as opposed to $O(C^{2N})$ “brute force” time, using Felsenstein’s pruning algorithm (1981) that takes advantage of conditional independence of evolution along tree branches



$$L(\mathcal{D}_s; \mathcal{T}, \theta) = \sum_{c_9 \in \mathcal{C}} \sum_{c_8 \in \mathcal{C}} \sum_{c_7 \in \mathcal{C}} \sum_{c_6 \in \mathcal{C}} \pi(c_9) Q_{c_9, A}^1(t_1; \theta) Q_{c_9, c_8}^8(t_8; \theta) Q_{c_8, C}^2(t_2; \theta) \\ Q_{c_8, c_7}^7(t_7; \theta) Q_{c_7, T}^3(t_3; \theta) Q_{c_7, c_6}^6(t_6; \theta) Q_{c_6, T}^4(t_4; \theta) Q_{c_6, T}^5(t_5; \theta)$$



$$L(\mathcal{D}_s; \mathcal{T}, \theta) = \sum_{c_9 \in \mathcal{C}} \sum_{c_8 \in \mathcal{C}} \sum_{c_7 \in \mathcal{C}} \sum_{c_6 \in \mathcal{C}} \pi(c_9) Q_{c_9, A}^1(t_1; \theta) Q_{c_9, c_8}^8(t_8; \theta) Q_{c_8, C}^2(t_2; \theta) \\ Q_{c_8, c_7}^7(t_7; \theta) Q_{c_7, T}^3(t_3; \theta) Q_{c_7, c_6}^6(t_6; \theta) Q_{c_6, T}^4(t_4; \theta) Q_{c_6, T}^5(t_5; \theta)$$

Only depends on c_9

$$L(\mathcal{D}_s; \mathcal{T}, \theta) = \sum_{c_9 \in \mathcal{C}} \pi(c_9) Q_{c_9, A}^1(t_1; \theta) \sum_{c_8 \in \mathcal{C}} Q_{c_9, c_8}^8(t_8; \theta) Q_{c_8, C}^2(t_2; \theta) \times \\ \sum_{c_7 \in \mathcal{C}} \left(Q_{c_8, c_7}^7(t_7; \theta) Q_{c_7, T}^3(t_3; \theta) \sum_{c_6 \in \mathcal{C}} (Q_{c_7, c_6}^6(t_6; \theta) Q_{c_6, T}^4(t_4; \theta) Q_{c_6, T}^5(t_5; \theta)) \right)$$

Only depends on c_8

Only depends on c_7

FELSENSTEIN'S PRUNING ALGORITHM

- Idea: for each node $\mathbf{n}$ in the tree, maintain a C (number of characters) - dimensional vector $\mathbf{L}_n$, whose i -th element records the probability of the subtree rooted at $\mathbf{n}$, given that the character at node $\mathbf{n}$ is $\mathbf{i}$.
- For leaves, $\mathbf{L}_n$ is easy to compute. $\mathbf{L}_n[i] = 1$ if $\mathbf{n}$ is labeled with character i , and $\mathbf{L}_n[i] = 0$, otherwise
- For interior nodes, $\mathbf{L}_n[i]$ is computed by iterating over all children of $\mathbf{n}$, and computing the cumulative probability of changing from $\mathbf{i}$ to any other state at child $\mathbf{m}$ (this uses $\mathbf{L}_m$), and then taking the product over all children
- At the root node, $\mathbf{r}$, compute the likelihood of the site, by summing over all characters $\mathbf{L}_r[i] \times \pi(i)$, where $\pi(i)$ is the (supplied) distribution of characters at the root.

Data : A tree T , with leaves labeled with integers from 0 to $C - 1$, root r , transition probabilities function for each branch (identified uniquely by the node m where the branch terminates) $\Pr(x \rightarrow y|m)$; $\pi(i)$: probability of observing character each i at the root.

Result: The likelihood score of the tree

```

1  $n \leftarrow$  the first node in the post-order traversal of  $T$  starting at  $r$ ;
2  $C \leftarrow$  the dimension of the alphabet;
3 while 1 do
4   if  $n$  is a leaf then
5      $L_n \leftarrow$  zero - vector( $C$ );
6      $L_n[\text{label } n] \leftarrow 1$ ;
7   end
8   else
9      $L_n \leftarrow$  vector of ones ( $C$ );
10    for  $p$  from 0 to  $C-1$  do
11      for node  $m$  is a child of  $n$  do
12         $s \leftarrow 0$ ;
13        for  $q$  from 0 to  $C-1$  do
14           $s+ = \Pr(p \rightarrow q|m)L_m[q]$ ;
15        end
16      end
17       $L_n[p] \leftarrow L_n[p] \times s$ ;
18    end
19  end
20  if  $n = r$  then
21    exit loop;
22  end
23   $n \leftarrow$  the next node in the post-order traversal of  $T$ ;
24 end
25  $L \leftarrow 0$ ;
26 for  $p$  from 0 to  $C-1$  do
27    $L \leftarrow L_r[p] \times \pi(p)$ ;
28 end
29 return  $L$ ;

```

Leaf initialization

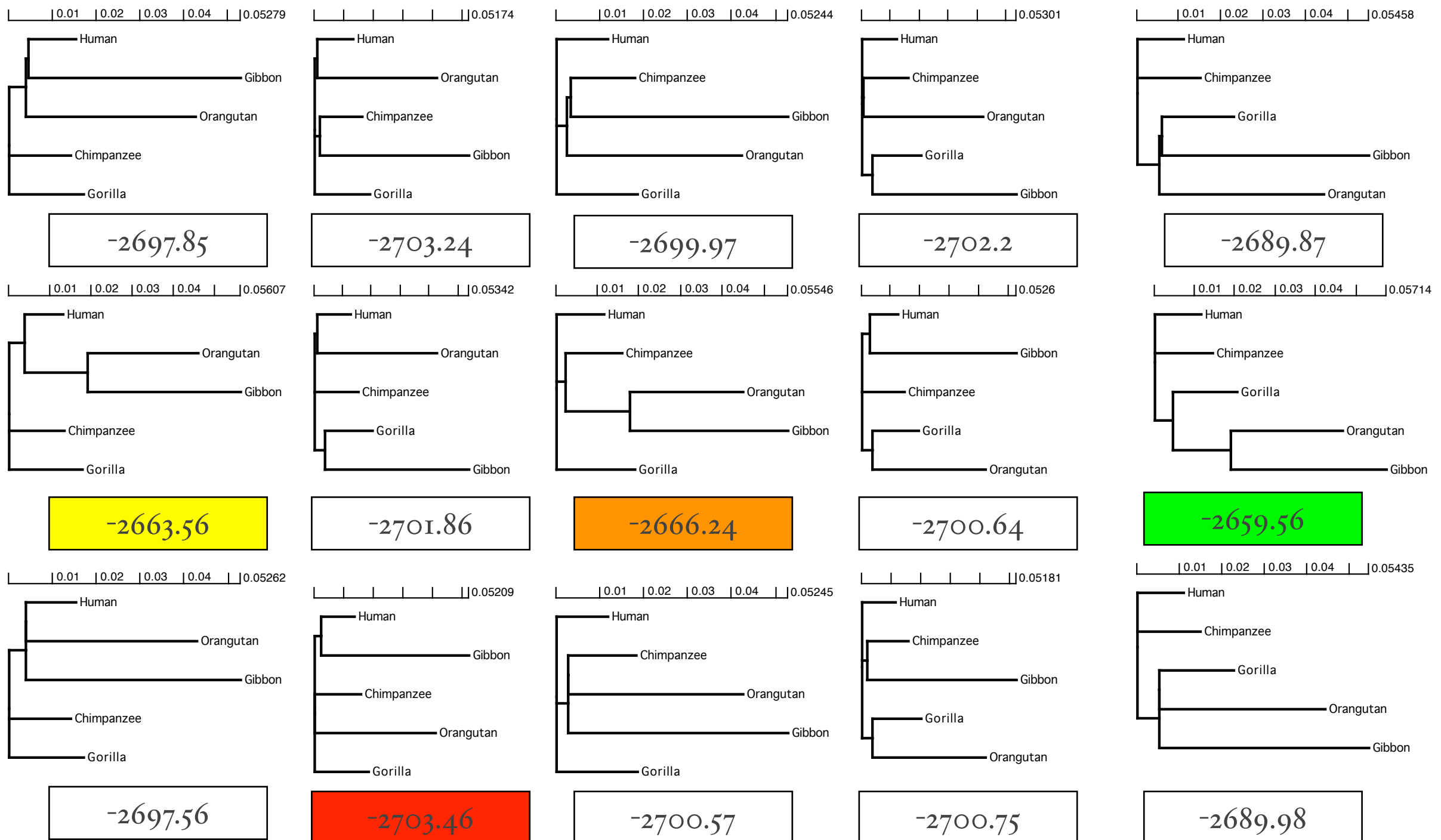
Interior node
calculation

Root clause

Felsenstein's Pruning Algorithm

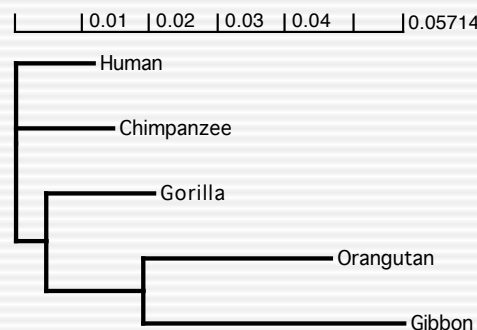
ML TREE COMPARISON

□ Partial mitochondrial DNA (5 taxa). Exhaustive tree search. GTR.



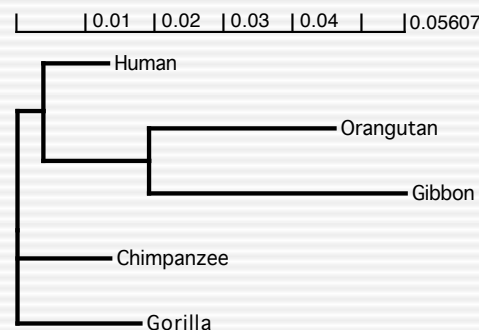
SELECTING THE BEST TREE.

- ☐ The tree with the best log-likelihood score can be reported as the true tree.
- ☐ What if there are other trees whose scores are not too much worse?
- ☐ How does one assess significance?
- ☐ How does one compare trees?



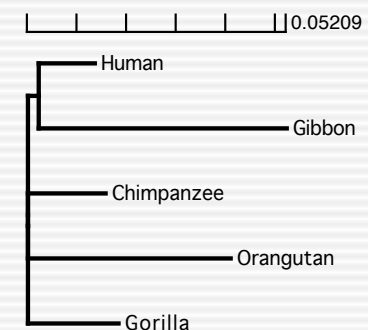
-2659.56

Best



-2663.56

2nd best



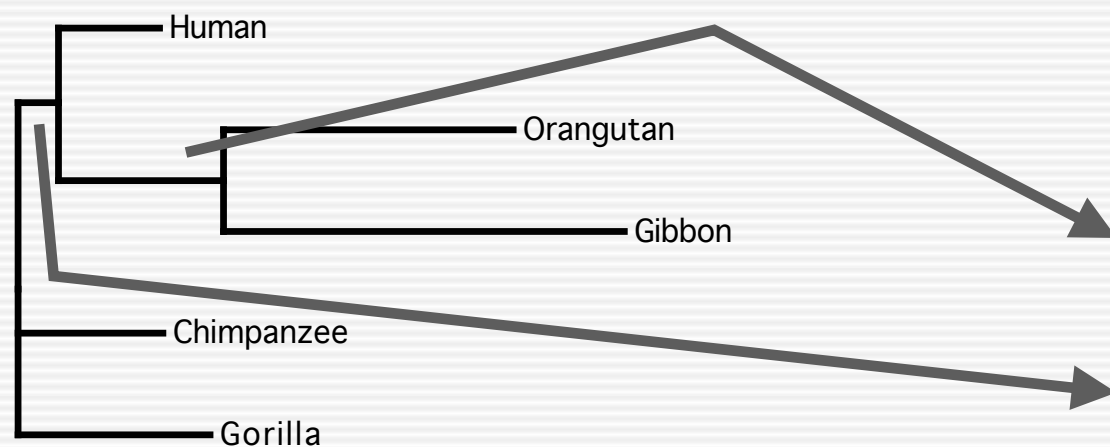
-2703.46

Worst

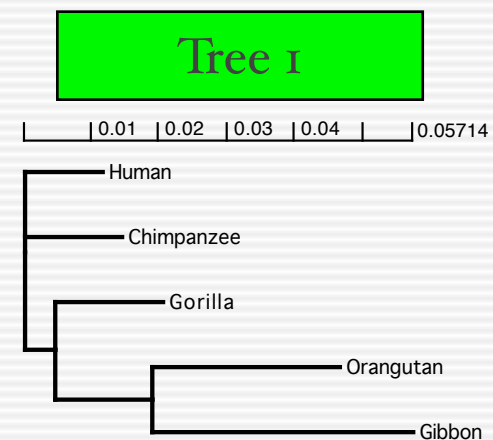
COMPARING TREES

- Most widely used tree distance measures are based on the concept of splits:
- Each interior branch partitions all leaves into two disjoint sets.
- The Robinson-Foulds (RF) distance between trees on N leaves is the number of different splits between the trees.
- For binary unrooted trees, can range from **0** to **$N-3$**

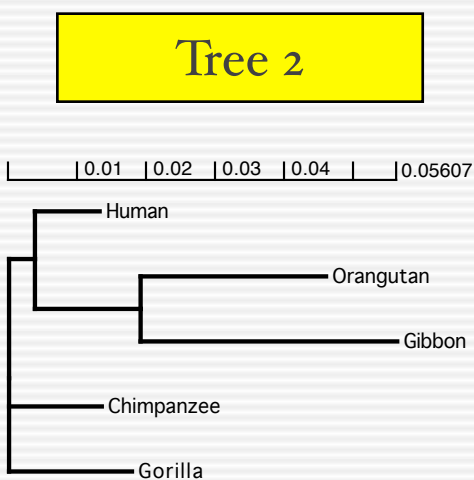
| 0.01 | 0.02 | 0.03 | 0.04 | 0.05607



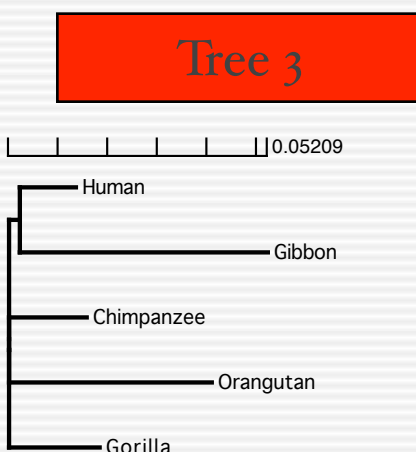
Split 1	Split 2
O,Gib	H,C,Gor
H,O,Gib	C,Gor



Split 1	Split 2
Gor,O,Gib	C,H
O,Gib	H,C,Gor



Split 1	Split 2
H,O,Gib	C,Gor
O,Gib	H,C,Gor



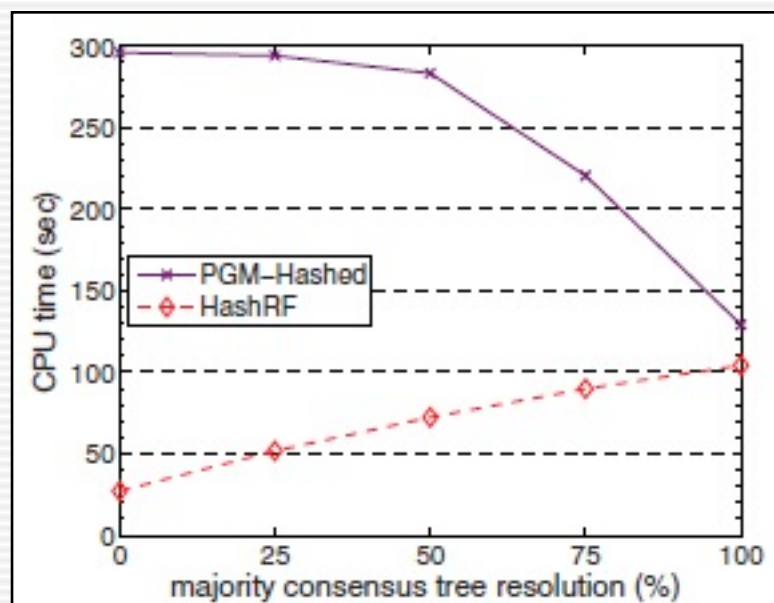
Split 1	Split 2
H,Gib	C,Gor,O
C,O	H,Gib,Gor

Robison Foulds
Distance Matrix

	T2	T3
T1	1	2
T2	-	2

RF CALCULATION ALGORITHMS

- Pairwise Robinson Foulds distance calculations can be performed exactly in $O(nk^2)$ time on **k** -leaf trees
- Probabilistic algorithms with a tunable error parameter (ϵ) can run very quickly in practice:



2048 trees x 2048 taxa

An Experimental Analysis of Robinson-Foulds Distance Matrix Algorithms

Seung-Jin Sul and Tiffani L. Williams

Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
{sulsj,tlw}@cs.tamu.edu

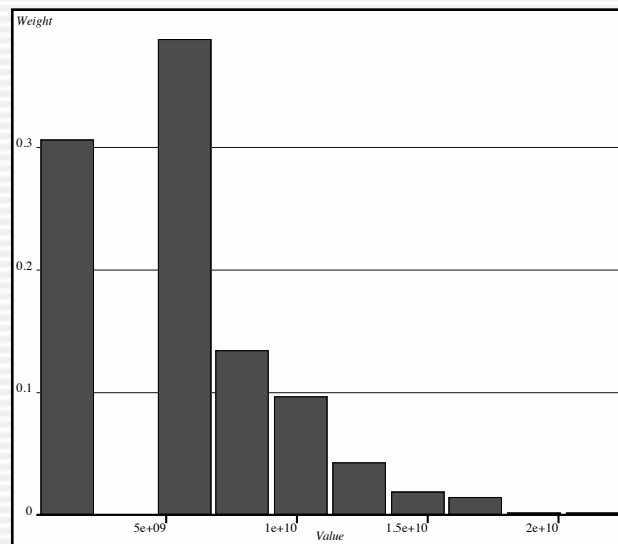
CONFIDENCE ASSESSMENT VIA BOOTSTRAP

- Biological inference is performed on limited length, noisy data
- **Sampling variance:** the error in our ability to infer a quantity (e.g. the tree topology) based on limited data
- Extreme example: outlier effects.
 - Consider trying to approximate the mean net worth in the US from a sample of 10 tax returns.
 - Assume that one of 10 is Bill Gates; hence we may have a data matrix that looks like the one on the left:
 - The mean estimated from the table is ~\$5B
 - We could resample the database of tax returns to get many tables and average over them
 - In biology, frequently all we have is one sample (an alignment); hence it's impossible to obtain another set

Bill G	\$50B
1	\$100K
2	\$20K
3	\$120K
4	\$30K
5	\$200K
6	\$10K
7	\$25K
8	\$40K
9	\$5K

BOOTSTRAP

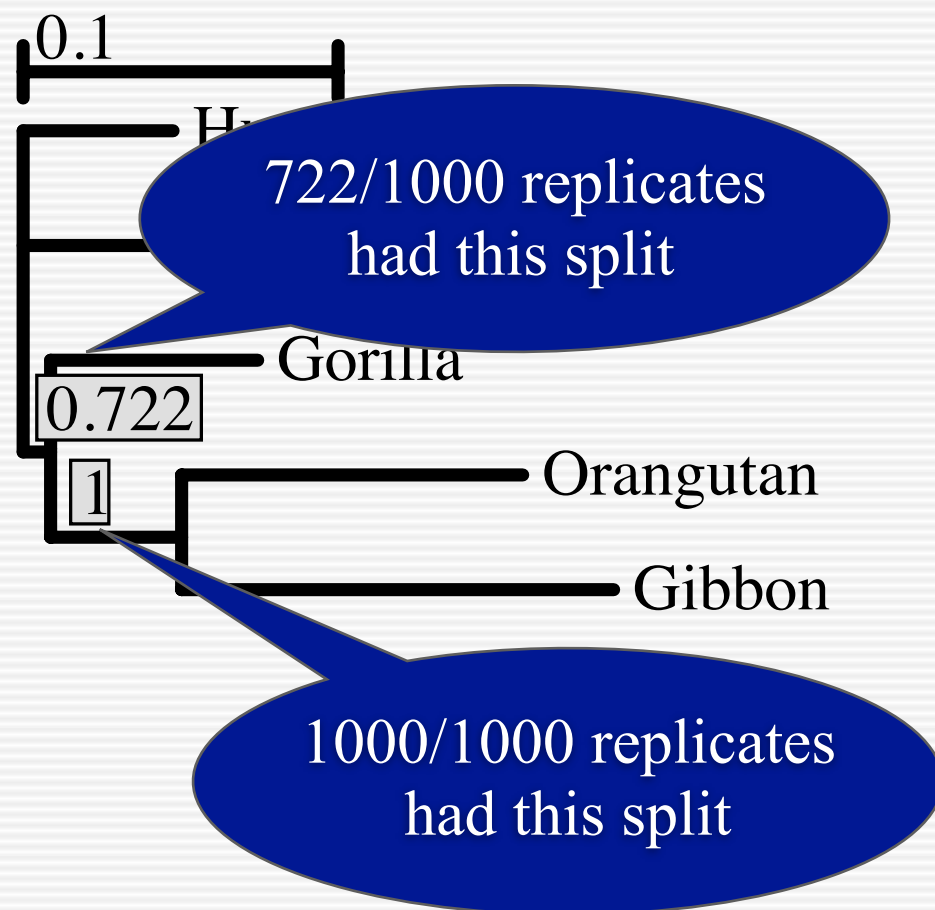
- We can approximate the true underlying distribution that generated our observed sample by resampling it with replacement:
 - Draw a value from a sample at random; replace it in the sample
 - Repeat N times (N - size of the sample).
- Back to the Bill Gates example
 - 1000 bootstrap replicates indicate that the standard deviation of our estimate is about the same as the mean: \$5B!
 - In other words, we have very little confidence in the obtained mean (~34% of cases have mean < 200K for example).



PHYLOGENETIC BOOTSTRAP

- ❑ Infer a phylogeny using your favorite method
- ❑ Generate bootstrap replicates: in this case our samples are alignment columns
- ❑ Repeat the inference procedure
- ❑ Tabulate the number of phylogenetic splits that are recovered in the replicates
- ❑ Branches with high bootstrap support values are those that have strong signal in the alignment
- ❑ Joe Felsenstein is credited with introducing the phylogenetic bootstrap

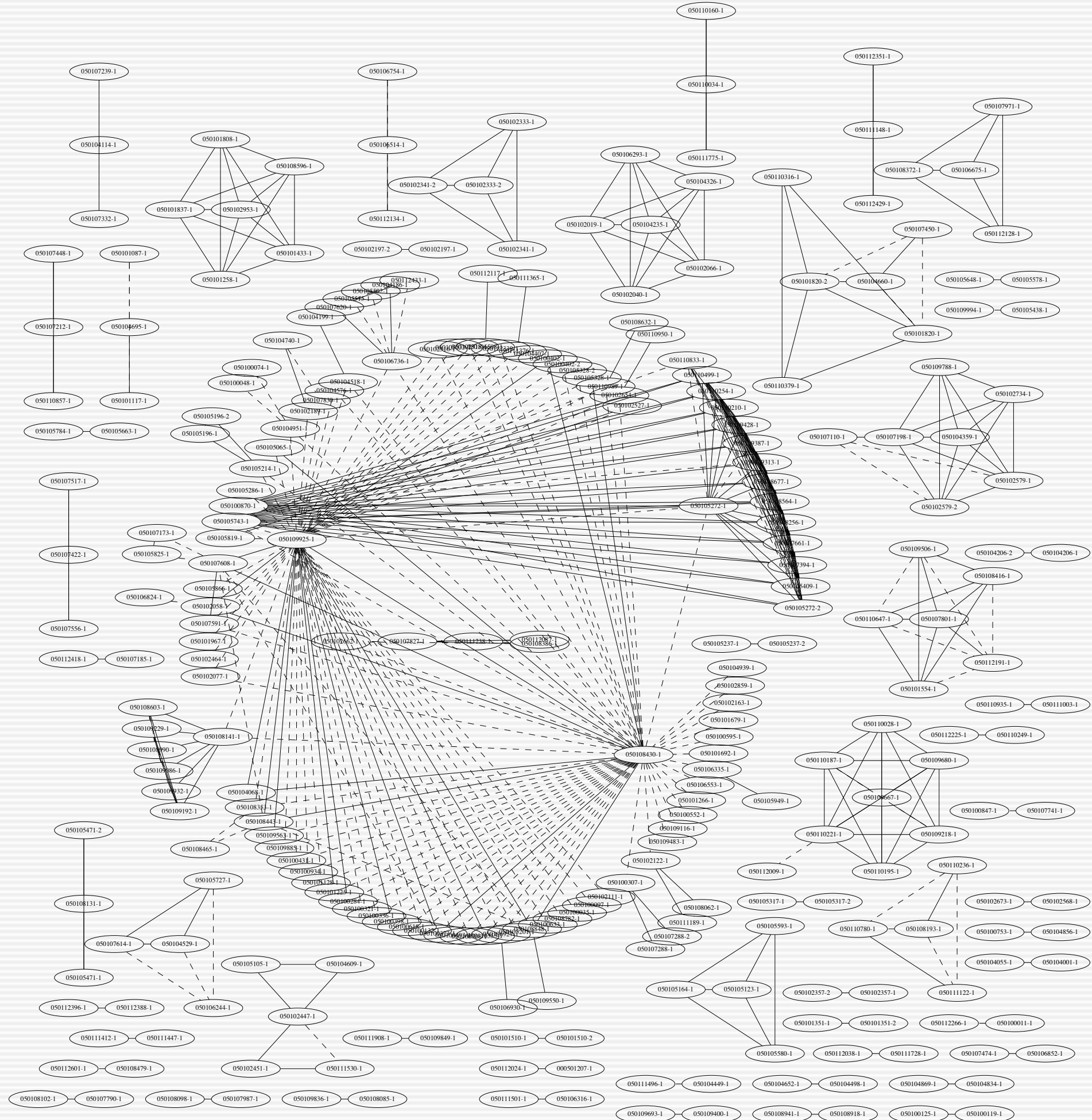
EXAMPLE MTDNA IN PRIMATES



- ☐ Used NJ to reconstruct a tree
- ☐ Drew 1000 alignment samples using the bootstrap
- ☐ Inferred 1000 trees: one from each resampled alignment
- ☐ If a given split from the **inferred tree** was found in a replicate, then its count was incremented by 1.
- ☐ The accepted number for good support is 0.7 or greater, but 1.0 is sometimes desired.
- ☐ This alignment is too short (or not informative enough) to unequivocally support the Gorilla-Chimp-Human branching order.
- ☐ This echoes the observation that the 2nd best ML tree changed that order.

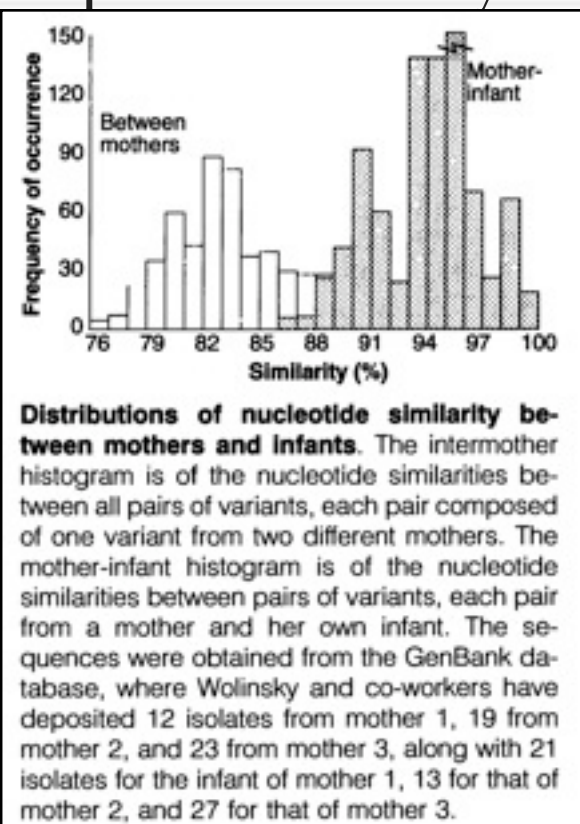
ESTABLISHING HIV TRANSMISSION CHAINS

- One of the major applications of molecular phylogenetics in HIV research is to attempt to identify putative transmission events (i.e. A may have infected B).
- This is obviously useful for epidemiological purposes: to estimate transmission networks, transmission cluster sizes etc
- A pop-sci application: court cases to prove HIV transmission.
- Generally the problem is quite difficult.



ESTABLISHING TRANSMISSION?

- It is nearly impossible to use phylogenetic trees alone to **definitively** prove transmission.
- What may a transmission event look like in a tree?
- HIV has a high mutation rate, so viruses from two 'unrelated' individuals will be quite divergent (upwards of 20% in some genomic regions)
- Direct transmissions tend to have lower pairwise distances than unrelated cases (this depends on many factors, most importantly, sampling time relative to transmission)



The Continuing Case of the Florida Dentist

Temple F. Smith and Michael S. Waterman

Molecular Epidemiology of HIV Transmission in a Dental Practice

Chin-Yih Ou, Carol A. Ciesielski, Gerald Myers,
Claudiu I. Bandea, Chi-Cheng Luo, Bette T. M. Korber,
James I. Mullins, Gerald Schochetman, Ruth L. Berkelman,
A. Nikki Economou, John J. Witte, Lawrence J. Furman,
Glen A. Satten, Kersti A. MacInnes, James W. Curran,
Harold W. Jaffe, Laboratory Investigation Group*,
Epidemiologic Investigation Group†

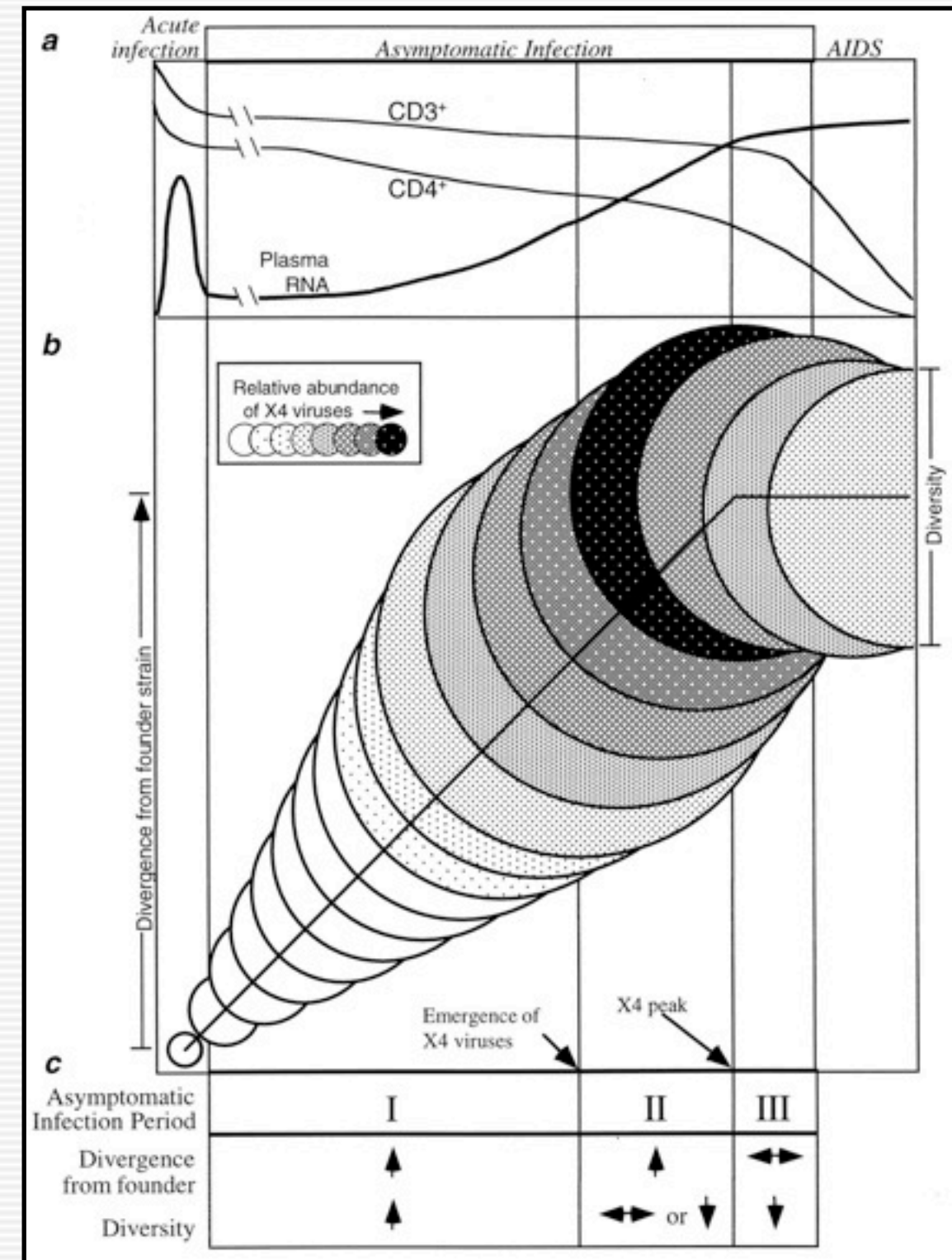
- David Acer - a Florida dentist became infected with HIV in late 1986 (Kaposi's sarcoma) and developed AIDS in 1987. He continued to practice general dentistry for two years following and performed invasive procedures on patients.
- In 1990, a young woman in her 20s with AIDS but no identifiable risk factors, and a patient of Acer's (patient A), prompted the CDC to investigate his practice.
- Based on epidemiological evidence and sketchy molecular data, the CDC concluded that Acer might have infected patient A.
- The dentist then requested that his former patients be tested for HIV.
- Out of 1100 tested persons, two patients (B and C) were found to be HIV positive.
- An additional HIV+ person (D - a former patient of Acer) was identified by using the AIDS case registry and two more patients (E,F) contacted the CDC to report that they were HIV infected.

CAVEAT

- HIV evolves **a lot** in a single patient, so it's a moving target, and diverse at any given time point.
- Consider the pattern of HIV-1 envelope evolution over time proposed by Shakarappa et al.

Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection

RAJ SHANKARAPPA,¹ JOSEPH B. MARGOLICK,² STEPHEN J. GANGE,³ ALLEN G. RODRIGO,¹ DAVID UPCHURCH,¹ HOMAYOON FARZADEGAN,³ PHALGUNI GUPTA,⁴ CHARLES R. RINALDO,⁴ GERALD H. LEARN,¹ XI HE,¹ XIAO-LI HUANG,⁴ AND JAMES I. MULLINS^{1*}



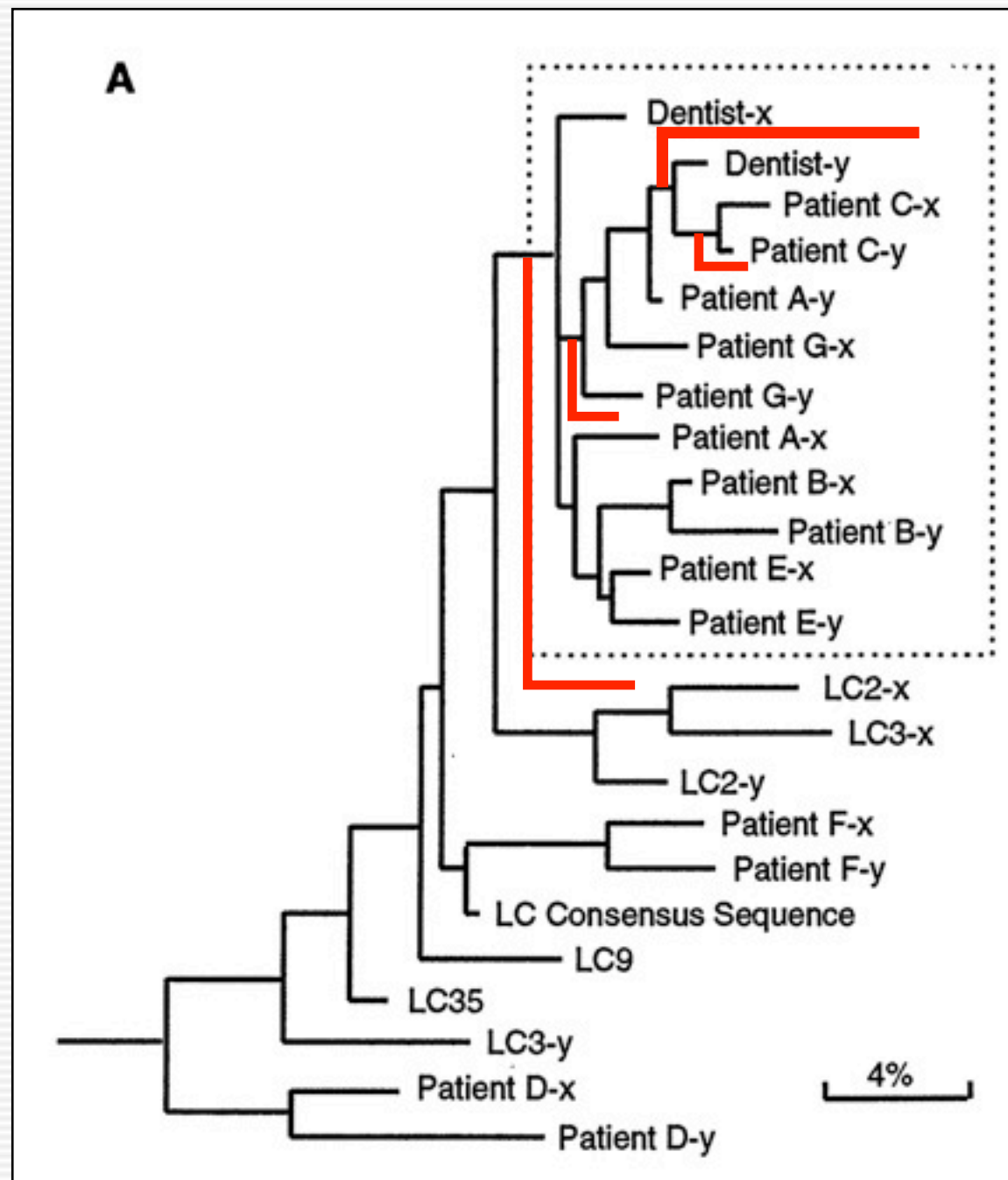
MOLECULAR DATA

- Multiple clones from a short region of the envelope gene (C2-V₃ region) was sequenced from each patient (and the dentist)
- 30 local controls (i.e. epidemiologically unrelated cases) were also included

Table 1. Dental cohort clinical information and HIV nucleotide variation in the C2-V3 domain of the envelope gene.

Person	Sex	Known risk factor	Clinical status*	M13 clones (no.)	Intraperson variation† (%)	Interperson variation† (%)	
						To dentist	To 30 LCs‡
Dentist	M	Yes	AIDS	6	3.3 (0.8–5.4)		11.0 (5.8–16.0)
Patient A	F	No	AIDS	6	2.0 (0.0–4.5)	3.4 (0.8–6.2)	10.9 (5.4–14.8)
Patient B	F	No	Asymptomatic (CD4 = 222/μl)	12	1.9 (0.4–3.7)	4.4 (2.1–7.0)	11.2 (6.2–16.5)
Patient C	M	No§	Asymptomatic (CD4 = <50/μl)	5	1.2 (0.4–1.6)	3.4 (2.1–4.9)	11.1 (7.0–15.6)
Patient E	F	No	Asymptomatic (CD4 = 567/μl)	6	2.1 (0.4–3.7)	3.4 (1.2–6.6)	10.8 (5.8–14.8)
Patient G	M	No	Asymptomatic (CD4 = 400/μl)	5	2.8 (1.6–3.7)	4.9 (2.9–7.0)	11.8 (6.2–16.9)
Patient D	M	Yes	AIDS	5	7.5 (0.0–9.9)	13.6 (11.5–15.6)	13.1 (7.8–17.3)
Patient F	M	Yes	Asymptomatic (CD4 = 253/μl)	6	3.0 (0.8–5.8)	10.7 (8.2–13.6)	11.9 (7.0–17.3)

PHYLOGENETIC ANALYSIS: PARSIMONY



- ☐ Looks fairly convincing?
- ☐ This is inferred from 279 sites (146 informative). How reliable is the tree?
- ☐ But the malpractice insurer (CIGNA) hired their own experts to argue that the analysis is flawed because the local controls are not representative.
- ☐ Other types of evidence (e.g. molecular signature, known sexual contacts etc) were used to bolster the case.

Molecular evidence of HIV-1 transmission in a criminal case

Michael L. Metzker^{**}, David P. Mindell[‡], Xiao-Mei Liu^{*§}, Roger G. Ptak[¶], Richard A. Gibbs^{*}, and David M. Hillis^{**}

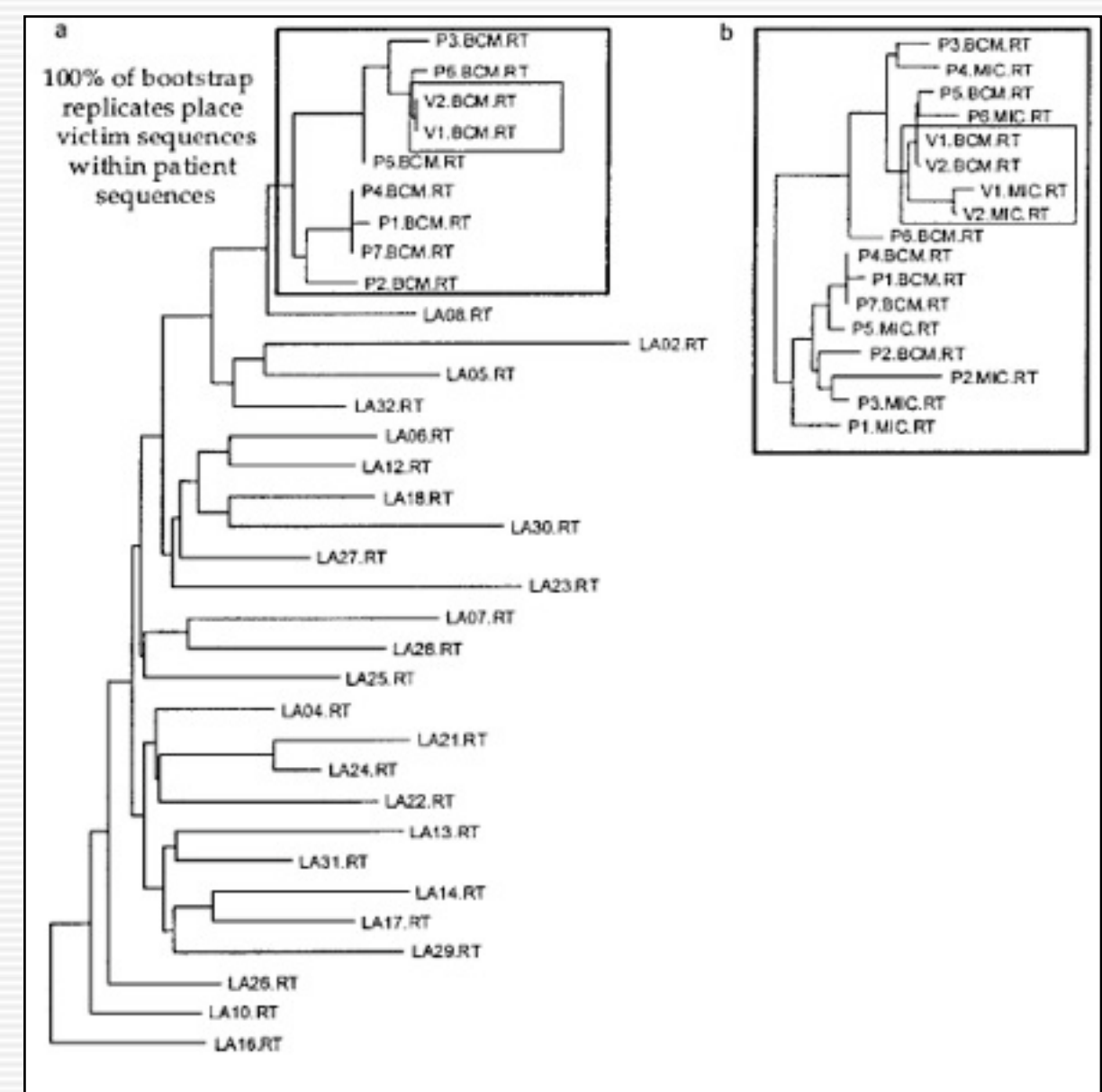
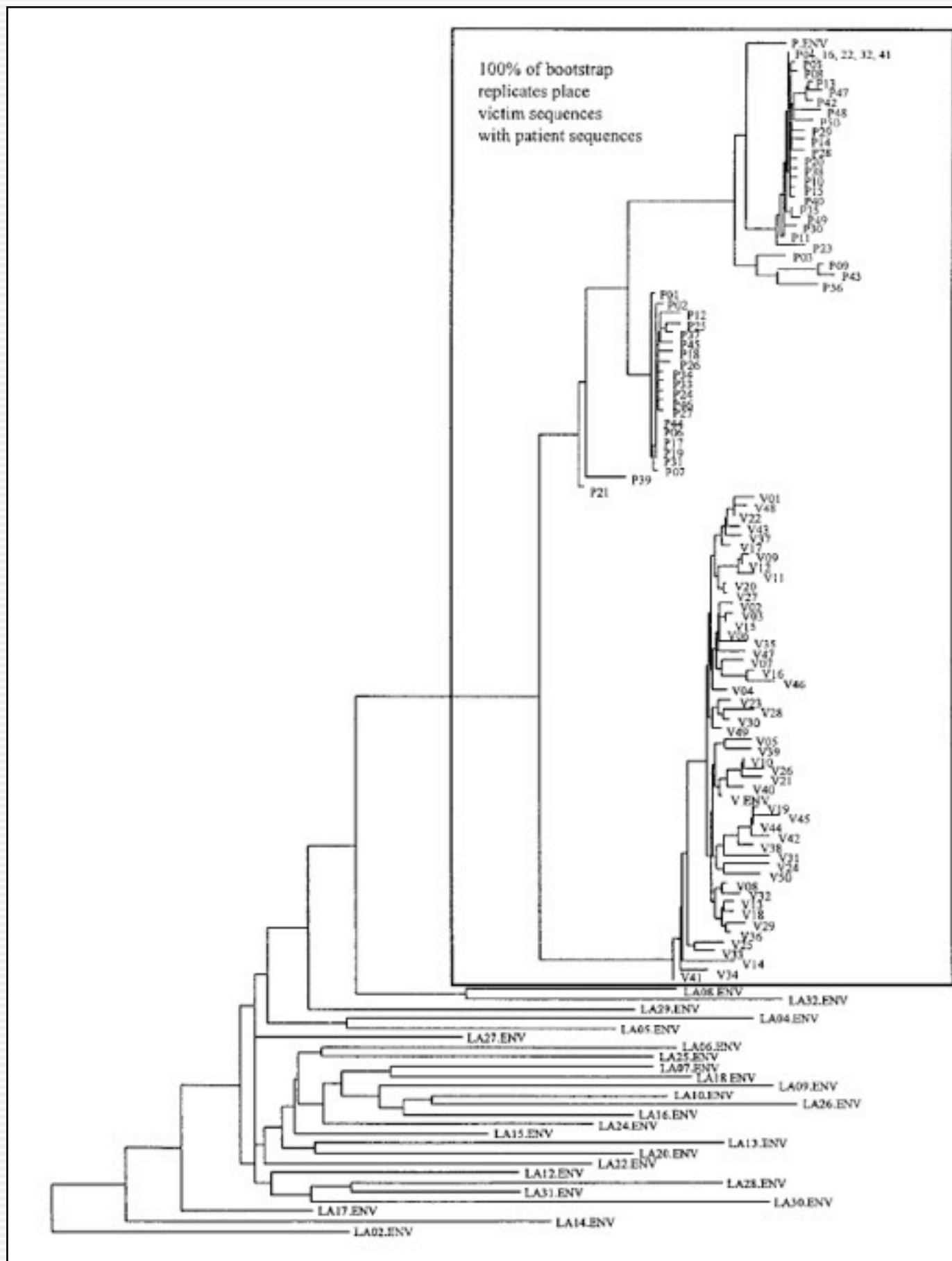
^{*}Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030; [‡]Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079; [§]School of Dentistry, Biologic and Materials Sciences, University of Michigan, Ann Arbor, MI 48109; and [¶]Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

Edited by Walter M. Fitch, University of California, Irvine, CA, and approved September 4, 2002 (received for review May 2, 2002)

- ☐ The prosecution argued that on August 4, 1994 a Lafayette, LA gastroenterologist made a mixture of blood or blood-products from two patients under the doctor's care, one infected with HIV-1 and the other with hepatitis C, and infected his former girlfriend by intramuscular injection.
- ☐ Molecular/epidemiological data were used to support the case
- ☐ Other risk factors for the victim were examined (sexual transmission, IDU, occupational hazards – the victim was a nurse) and determined unlikely.
- ☐ In January 1995 the victim tested positive for HIV-1 and accused the physician of infecting her.
- ☐ Law enforcement identified the potential source – a homosexual patient infected with HIV-1 in 1990.
- ☐ The hypothesis of transmission was tested.

DUE DILIGENCE

- This study used three different approaches: parsimony, maximum likelihood and Bayesian to build evolutionary trees from two different gene regions (RT and env)
- 28 local (Louisiana) HIV+ controls were used together with background sequences from GenBank.
- Confidence values (bootstrap or posterior probability) were obtained.



- ☐ The analysis does not preclude with 100% confidence that no intermediates were involved
- ☐ It also does not establish directionality (although in this case there are other factors that do)
- ☐ The fact that victim sequences cluster WITHIN the clade of source sequences makes a compelling case for transmission.