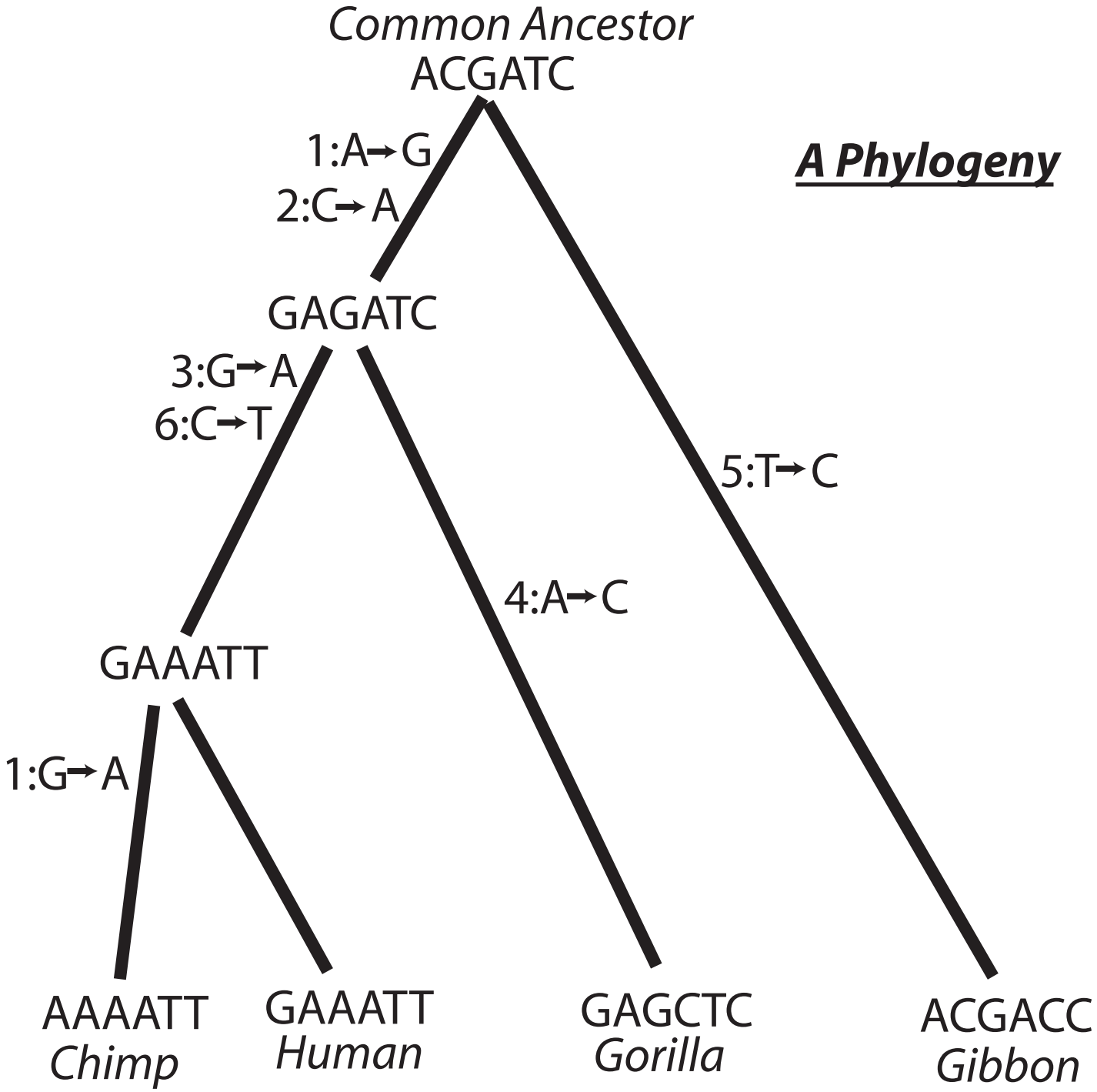
Overview of Molecular/Viral Evolution

Jeff Thorne, Genetics and Statistics

North Carolina State Univ.

These slides can be downloaded from ...
ftp://statgen.ncsu.edu/pub/thorne/samsi-evo-tutorial.pdf


Common Ancestor
ACGATC
1:A→G
2:C→A
A Phylogeny
GAGATC
3:G→A
6:C→T
5:T→C
4:A→C
GAAATT
1:G→A
AAAATT
Chimp
GAAATT
Human
GAGCTC
Gorilla
ACGACC
Gibbon

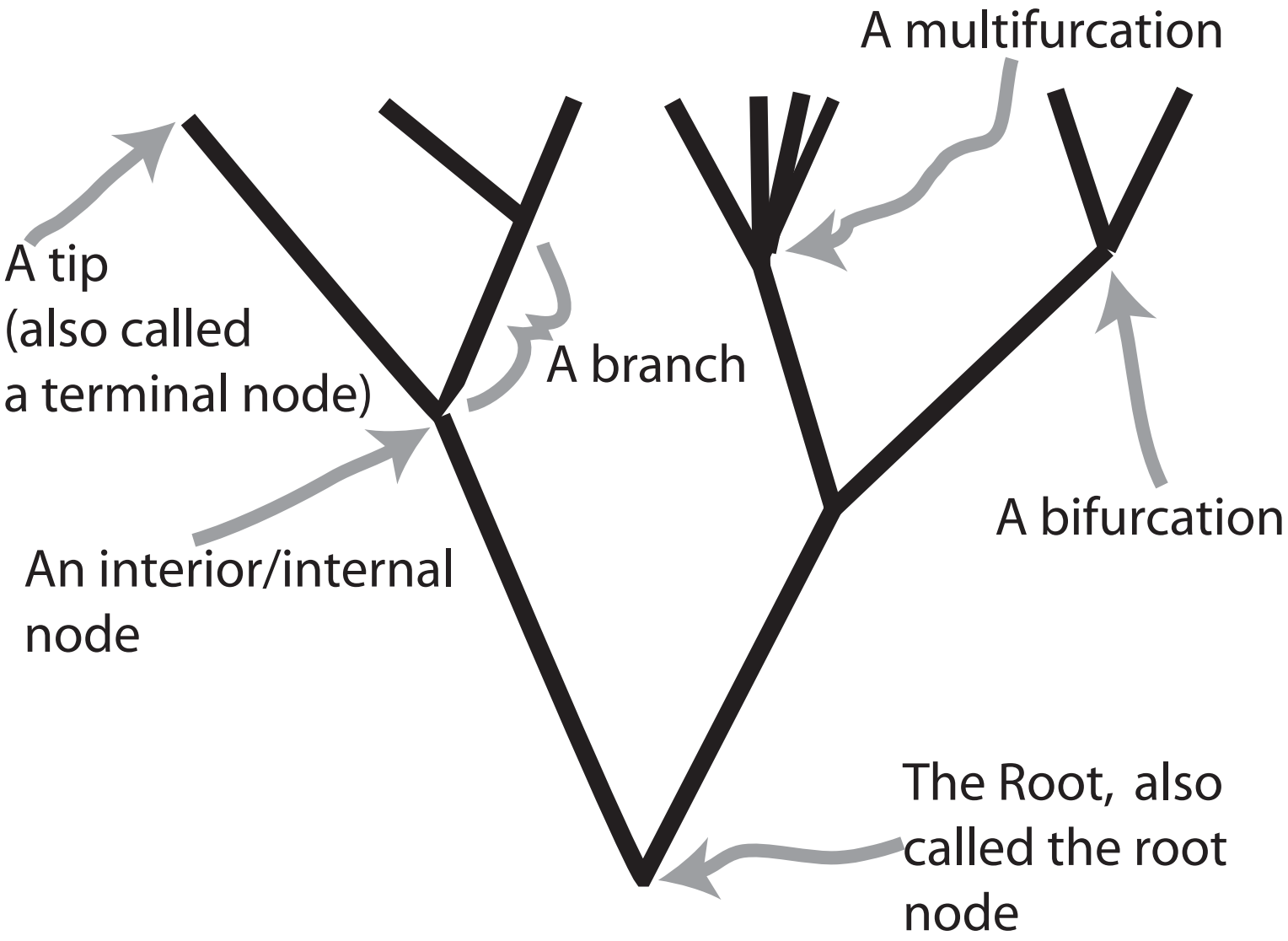
AAAATT
Chimp

GAAATT
Human

GAGCTC
Gorilla

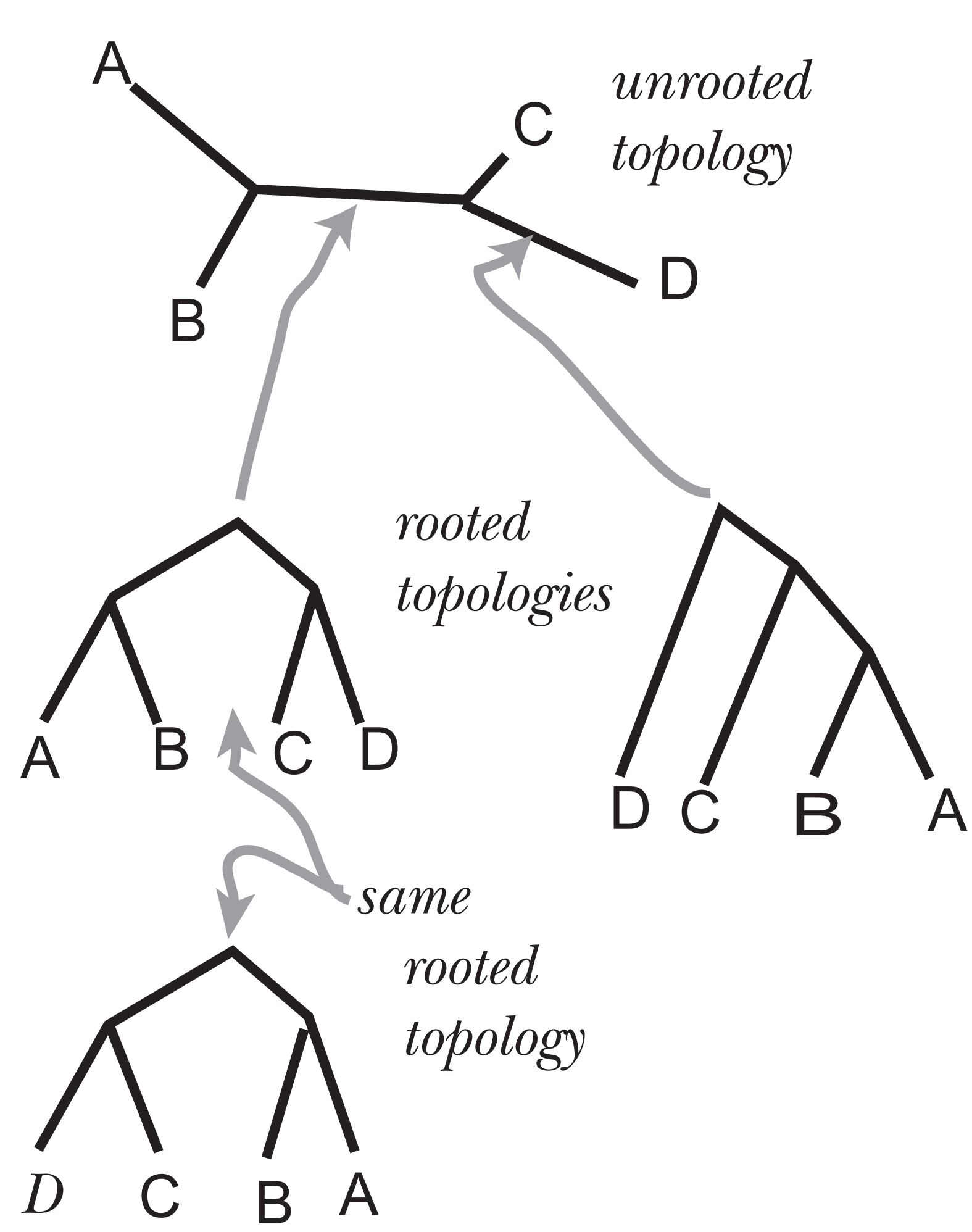
ACGACC
Gibbon

Tree Anatomy



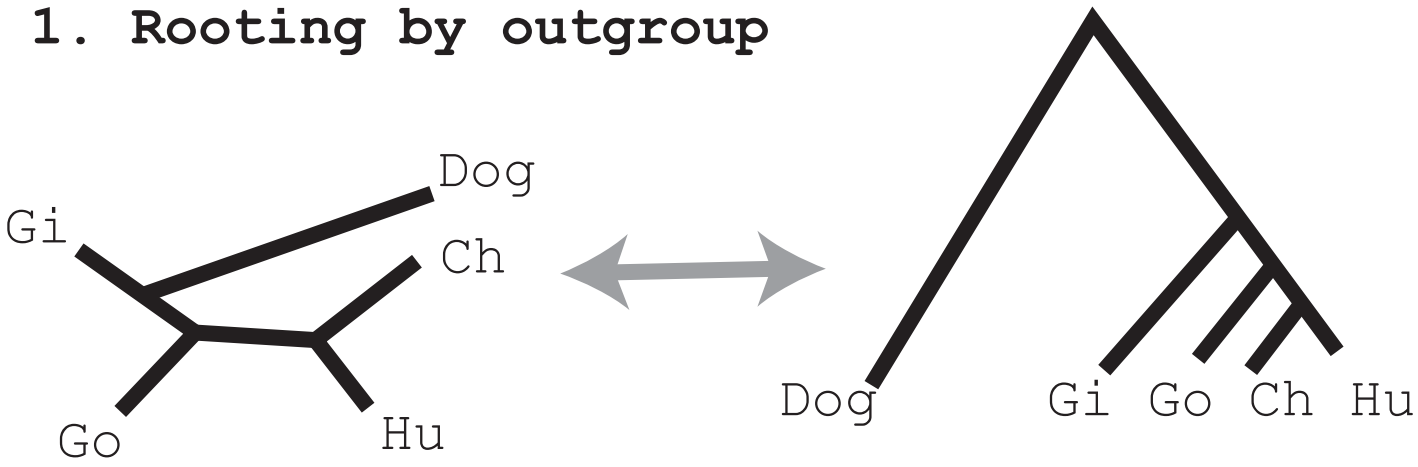
Translating between biology and math jargon

Biology	Math
Tree	Graph
Branch	Edge
Node	Vertice


A
B
C
D
unrooted
topology
rooted
topologies
A B C D
D C B A
same
rooted
topology
D C B A

The two common ways phylogenies are rooted:

1. Rooting by outgroup

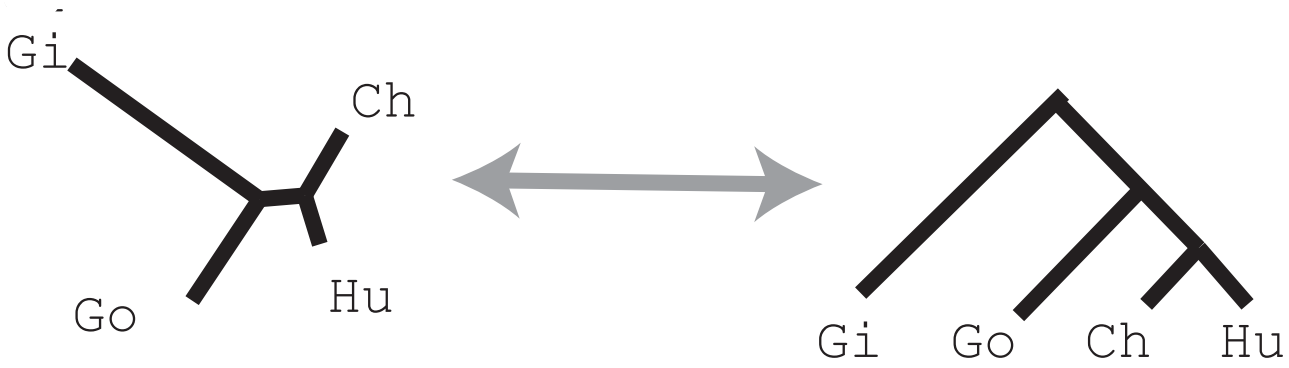


Ingroup species are all more closely related to each other than any are to the outgroup species.

Point on the phylogeny where outgroup attaches to ingroup is root of ingroup

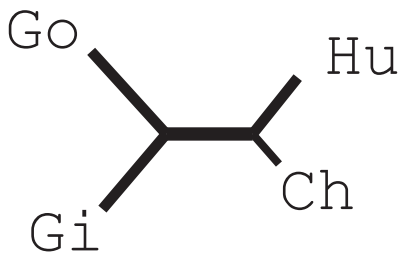
Ingroup root is most accurately inferred when outgroup is not terribly distantly related to ingroup members.

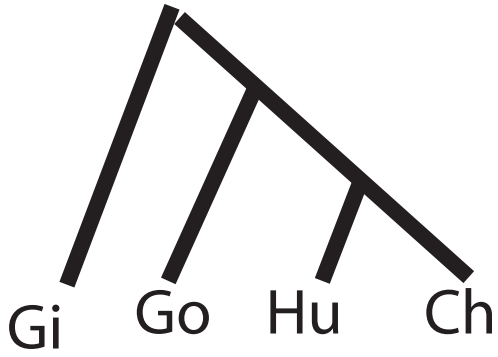
2. Rooting by "Molecular Clock"

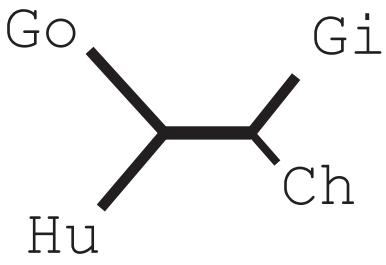


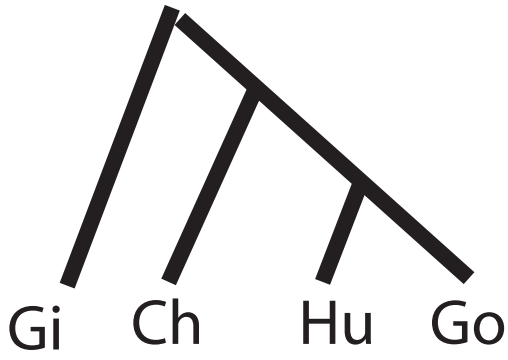
All "tips" should be equally far from root

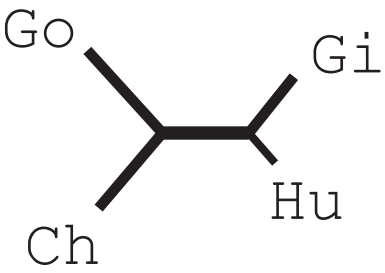
```
   Character:  123456
(Go) Gorilla:  GAGCTC
(Gi)  Gibbon:  ACGACC
(Hu)   Human:  GAAATT
(Ch)   Chimp:  AAAATT
```

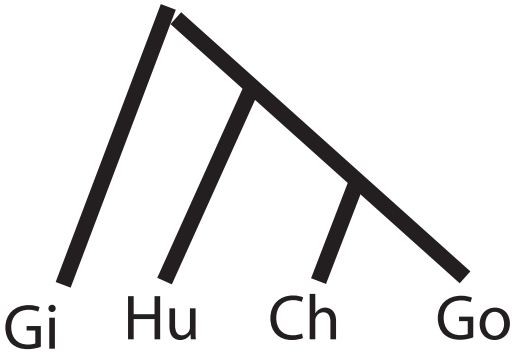


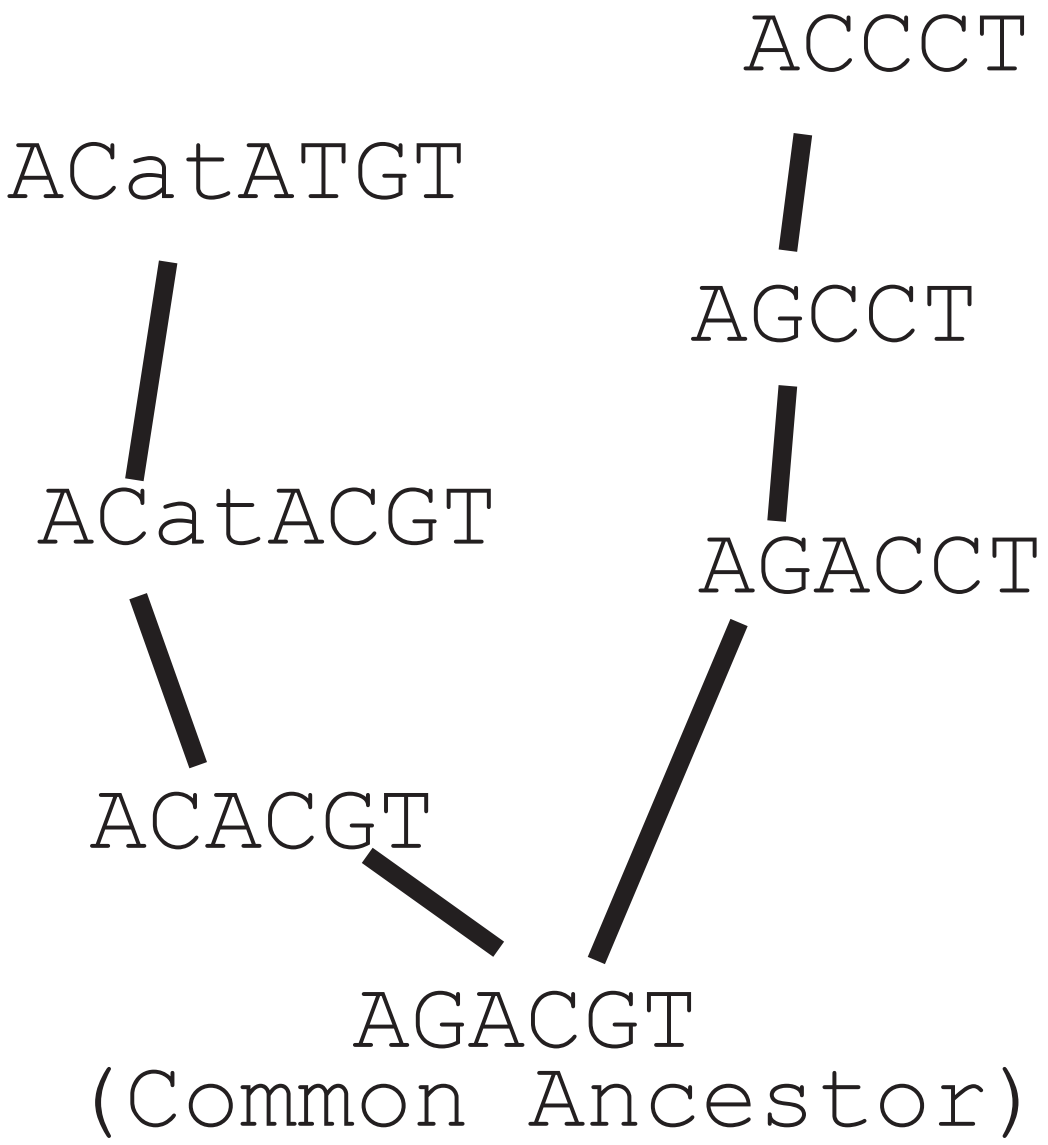












The "true" alignment:

ACATATGT
AC---CCT

Phylogeny Reconstruction is computationally difficult.

Number of Tips	Number of Rooted Trees	Number of Unrooted Trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	20,395
9	2,027,025	135,135
10	34,459,425	2,027,025

A bifurcating unrooted tree with n taxa has (2n-3) branches where $n \geq 2$.

Number of unrooted topologies for n taxa is:

$$(2n-5) \times (2n-7) \times \ldots \times (5) \times (3) \times (1) \; =$$

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad n \geq 3$$

For each unrooted bifurcating topology, there are $(2n-3)$ rooted bifurcating topologies ...

$$= \frac{(2n-3)!}{2^{n-2}(n-2)!} \qquad n \geq 3$$

A good introduction ...

"Inferring Phylogenies"
by Joseph Felsenstein
(published by Sinauer
Associates, August 2003)

covers ...

distance-based
parsimony
maximum likelihood
Bayesian

... phylogeny inference
procedures and more ...

θ – parameters of evolutionary model except for tree topology and branch lengths (e.g., transition/transversion parameter, residue frequencies, rate heterogeneity parameter, etc.)

τ – evolutionary tree topology and branch lengths

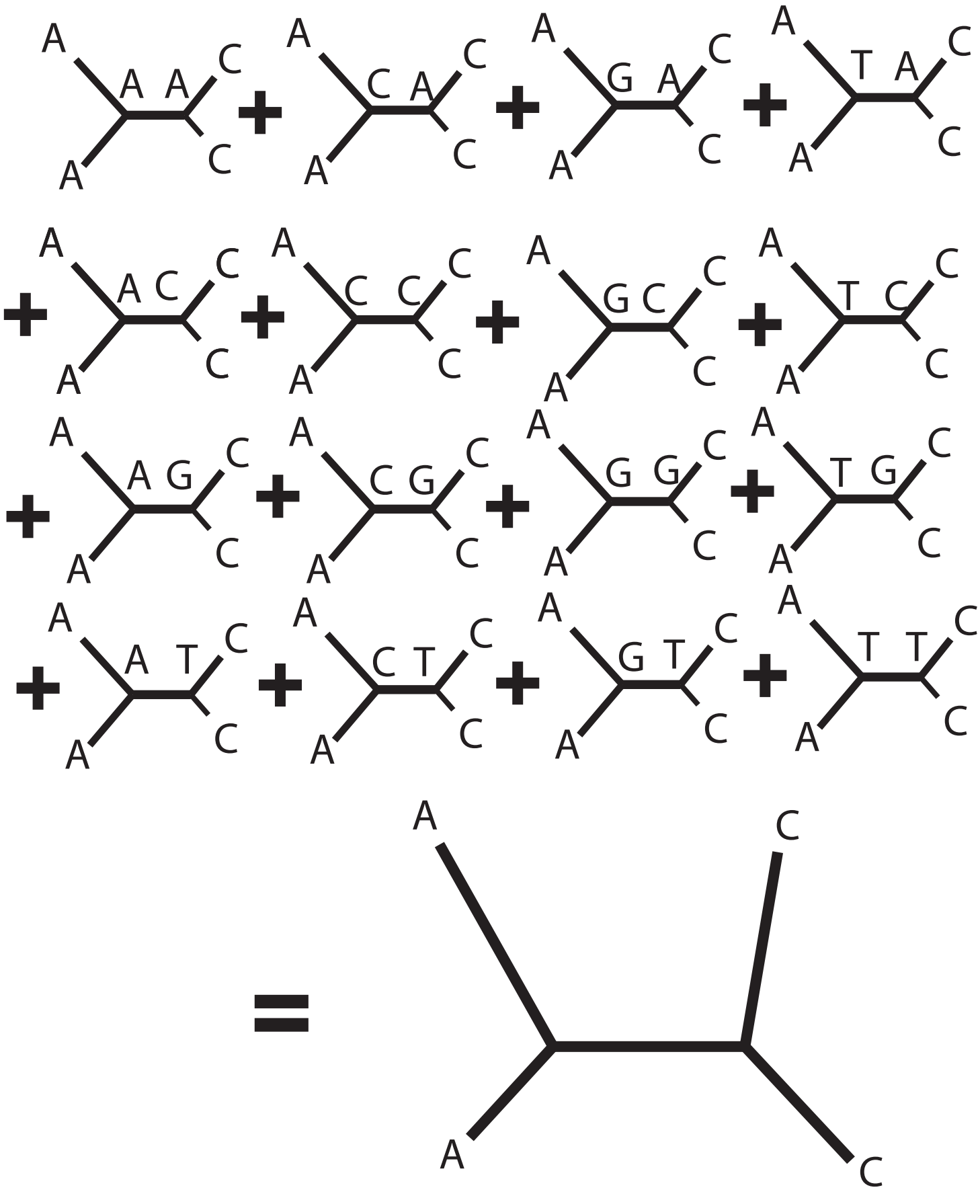
X – aligned sequence data

$\Pr(X \mid \theta, \tau)$ is the likelihood

$\max_\tau \max_\theta \Pr(X \mid \theta, \tau)$ is the maximum likelihood

the topology that represents the τ that maximizes the above is the maximum likelihood estimate of topology

Likelihood Idea:



To calculate likelihood by summing over possible internal node states,

"pruning algorithm"

(Felsenstein, 1981, J.Mol.Evol.,17:368-376)

is available.

4-state substitution model

From \ To	A	C	G	T
A	−	+	+	+
C	+	−	+	+
G	+	+	−	+
T	+	+	+	−

Q will represent a matrix of instantaneous rates of change. For the general time reversible model, the entries of Q are:

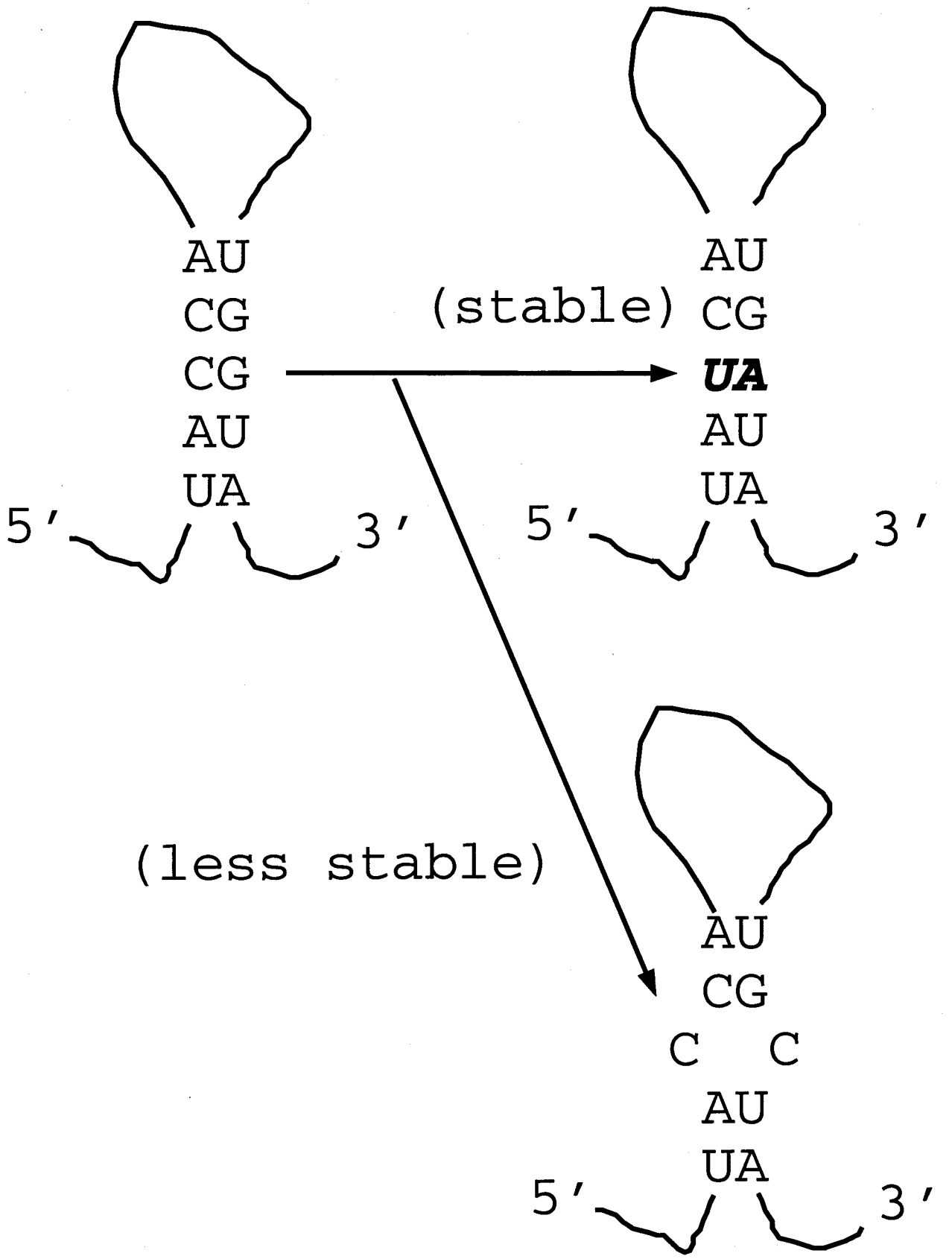
		To		
From	A	C	G	T
A	$-(a\pi_C + b\pi_G + c\pi_T)$	$a\pi_c$	$b\pi_G$	$c\pi_T$
C	$a\pi_A$	$-(a\pi_A + d\pi_G + e\pi_T)$	$d\pi_G$	$e\pi_T$
G	$b\pi_A$	$d\pi_C$	$-(b\pi_A + d\pi_C + f\pi_T)$	$f\pi_T$
T	$c\pi_A$	$e\pi_C$	$f\pi_G$	$-(c\pi_A + e\pi_C + f\pi_G)$

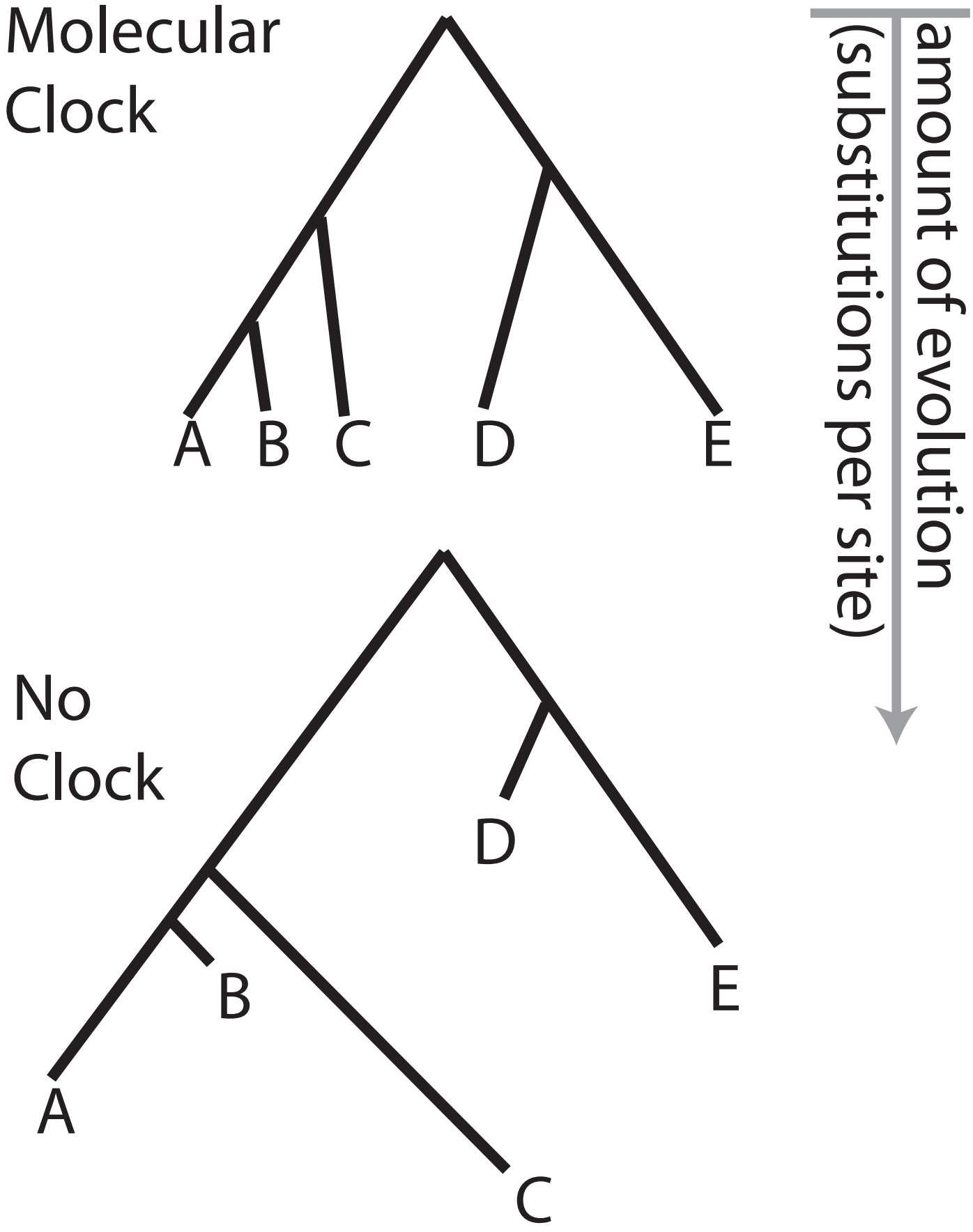
In above matrix: a, b, c, d, e, and f cannot be negative.

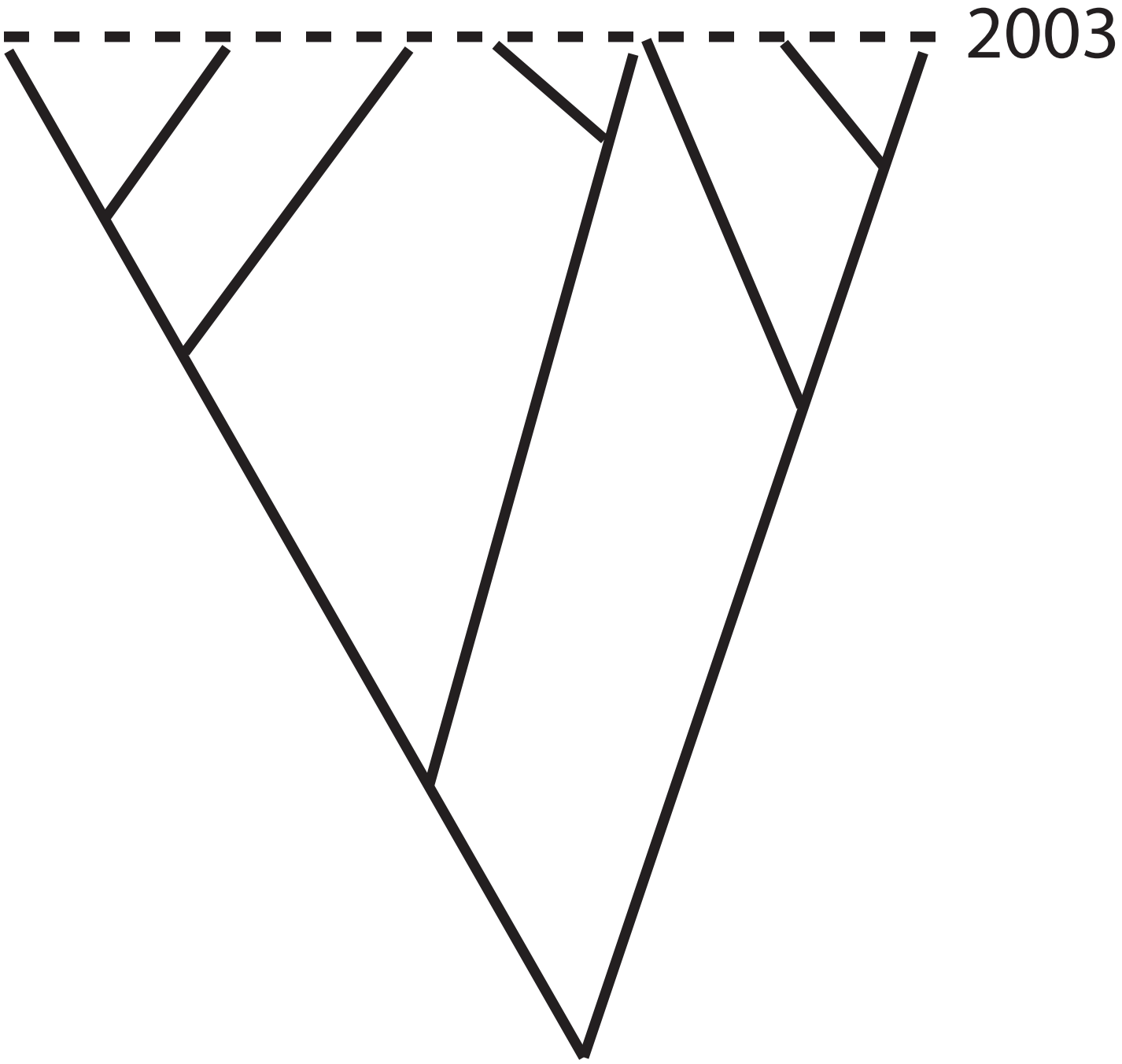
With any rate matrix (including above), the transition probabilities $P(t)$ can be determined from the rate matrix Q and the amount of evolution t via

$$P(t) = e^{Qt} = I + \frac{(Qt)}{1!} + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \ldots,$$

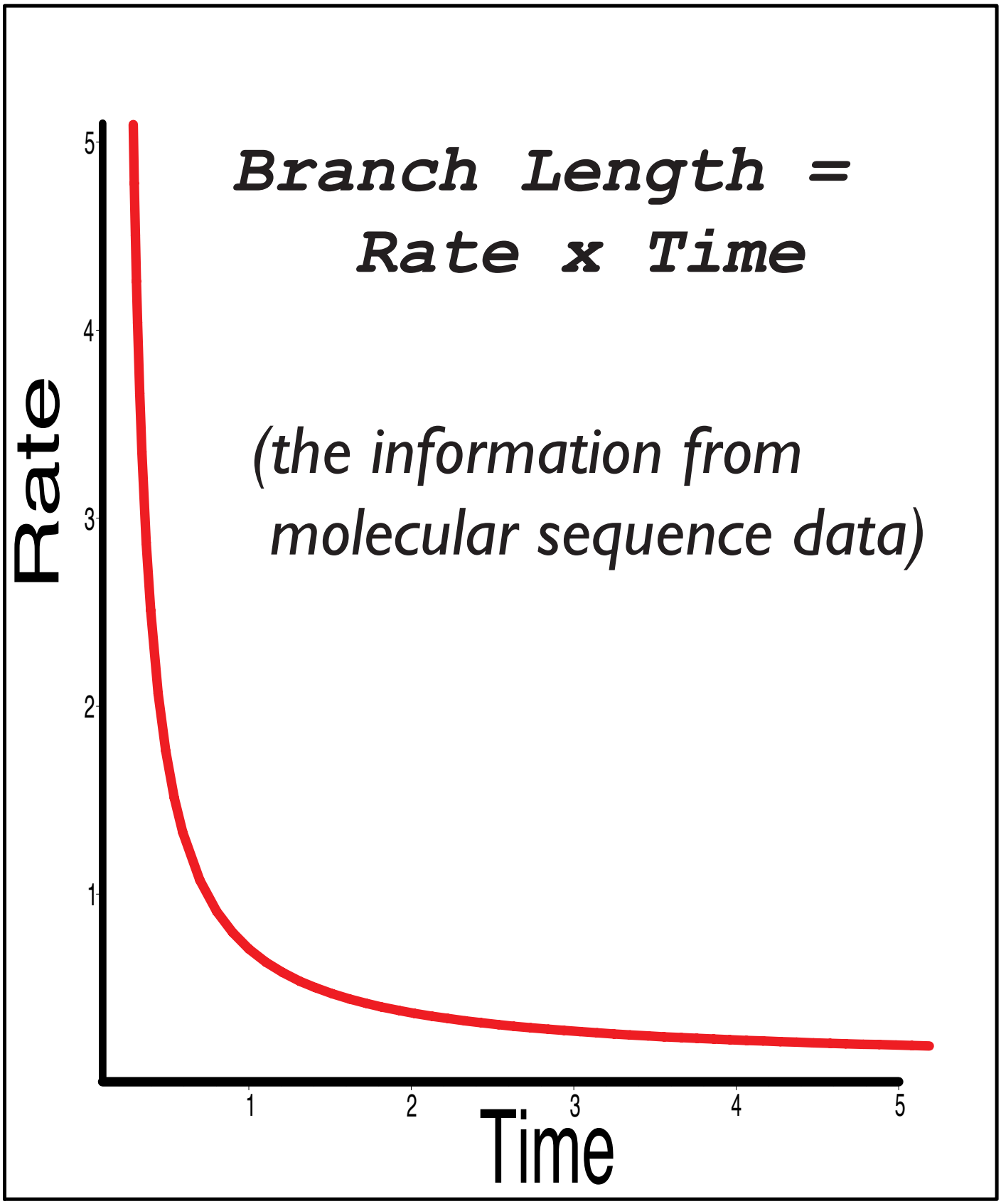
where I is the identity matrix.


AU
CG
CG
AU
UA
5′
3′
(stable)
AU
CG
UA
AU
UA
5′
3′
(less stable)
AU
CG
C C
AU
UA
5′
3′


Molecular Clock
A
B
C
D
E
amount of evolution (substitutions per site)
No Clock
D
E
B
A
C

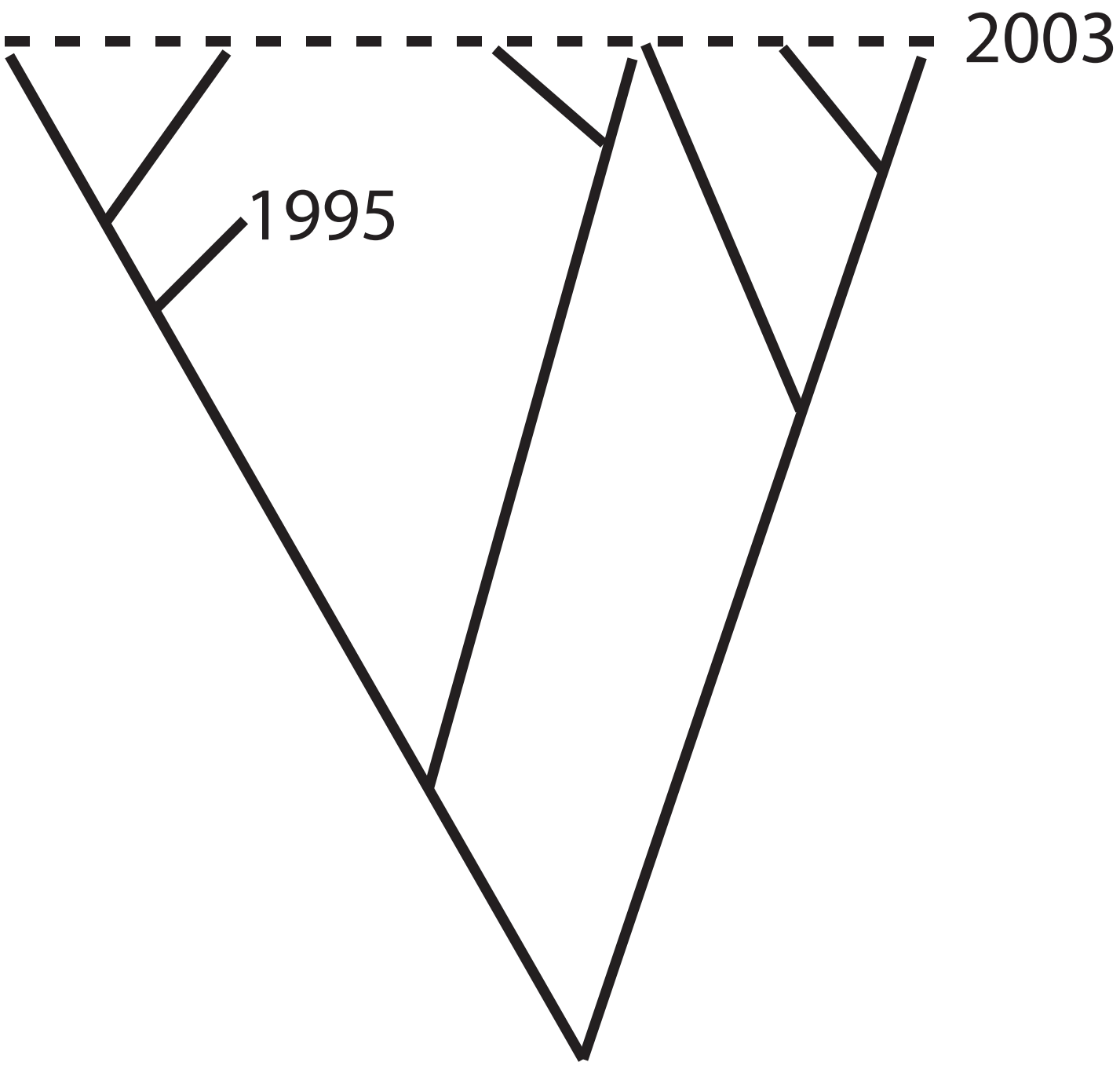


Contemporaneously Sampled Data


Branch Length =
Rate x Time
(the information from
molecular sequence data)
Rate
5
4
3
2
1
1
2
3
4
5
Time

Substitution rates of RNA viruses (Suzuki & Gojobori 1999)

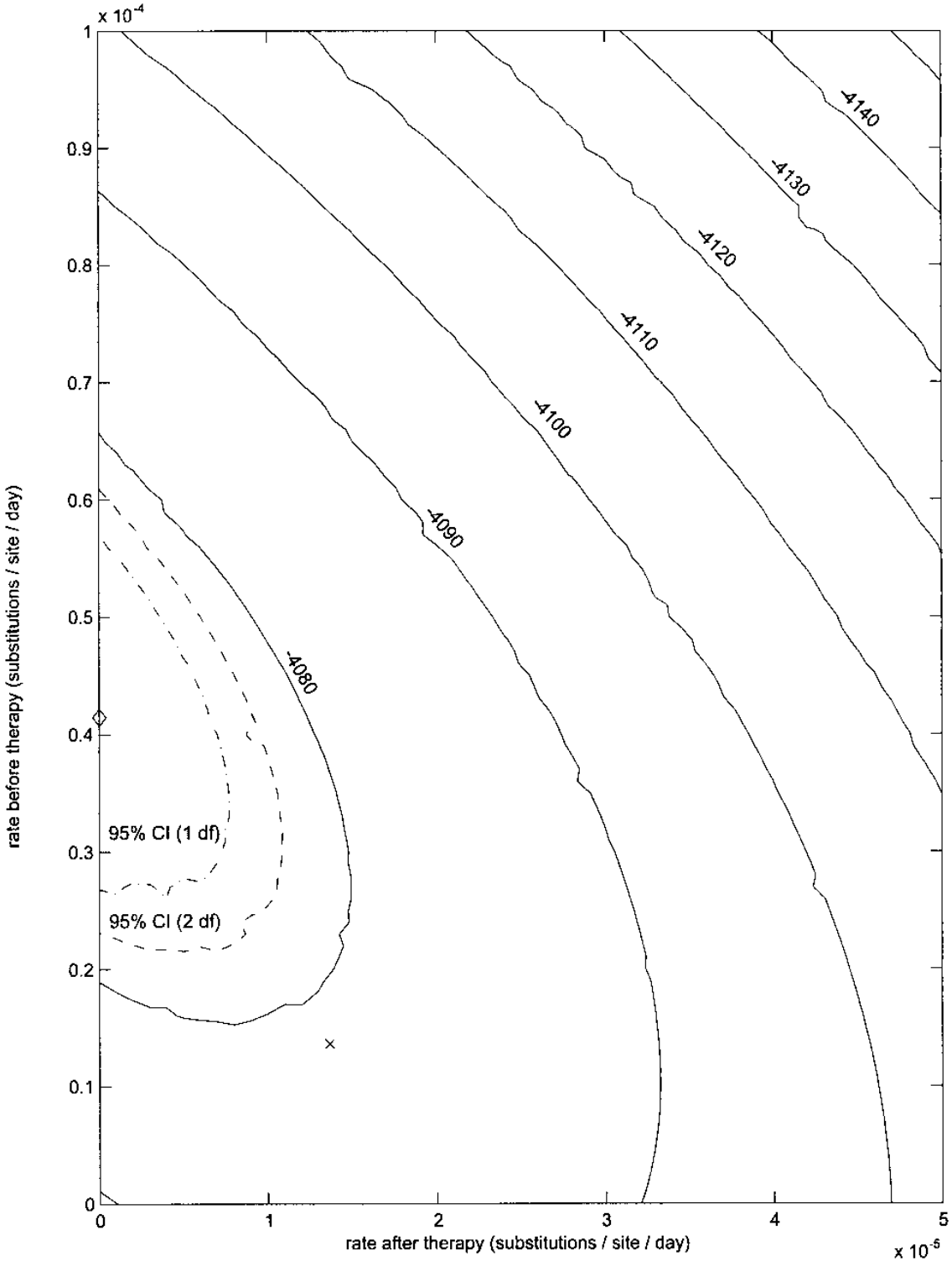
Virus and Organism	Gene	Substitution Rate (/site/year)	Reference
Ebola virus	GP	3.6×10^{-5}	
Marburg virus	VP35	3.6×10^{-4}	
	VP30	3.8×10^{-5}	
	VP24	1.3×10^{-4}	
HIV-1[a]	*gag*	$(1.0 - 3.9) \times 10^{-3}$	Li, Tanimura, and Sharp (1988); Gojobori, Moriyama, and Kimura (1990); Gojobori et al. (1994)
	pol	1.6×10^{-3}	Li, Tanimura, and Sharp (1988)
	env	$(3.9 - 5.1) \times 10^{-3}$	Li, Tanimura, and Sharp (1988); Gojobori et al. (1994)
	envhv	14.0×10^{-3}	Li, Tanimura, and Sharp (1988)
Human influenza A virus	HA (H3)	$(2.9 - 3.6) \times 10^{-3}$	Gojobori, Moriyama, and Kimura (1990); Hayashida et al. (1985)
	NA (N1)	3.7×10^{-3}	Hayashida et al. (1985)
	NA (N2)	2.8×10^{-3}	Hayashida et al. (1985)
MMSV[b]	v-*mos*	8.2×10^{-4}	Gojobori, Moriyama, and Kimura (1990)
MMLV[c]	*gag*	5.4×10^{-4}	Gojobori and Yokoyama (1985)
HCV[d]	C	6.3×10^{-4}	Ina et al. (1994)
	E	3.2×10^{-4}	Ina et al. (1994)
	NS1	7.5×10^{-4}	Ina et al. (1994)
	NS3	3.3×10^{-4}	Ina et al. (1994)
	NS5	2.2×10^{-4}	Ina et al. (1994)
HBV[e]	P	1.5×10^{-5}	Orito et al. (1989)
	pre-S	2.6×10^{-5}	Orito et al. (1989)
	C	1.8×10^{-5}	Orito et al. (1989)
	X	5.5×10^{-5}	Orito et al. (1989)
Mammals	α-globin	5.6×10^{-10}	Li, Luo, and Wu (1985)



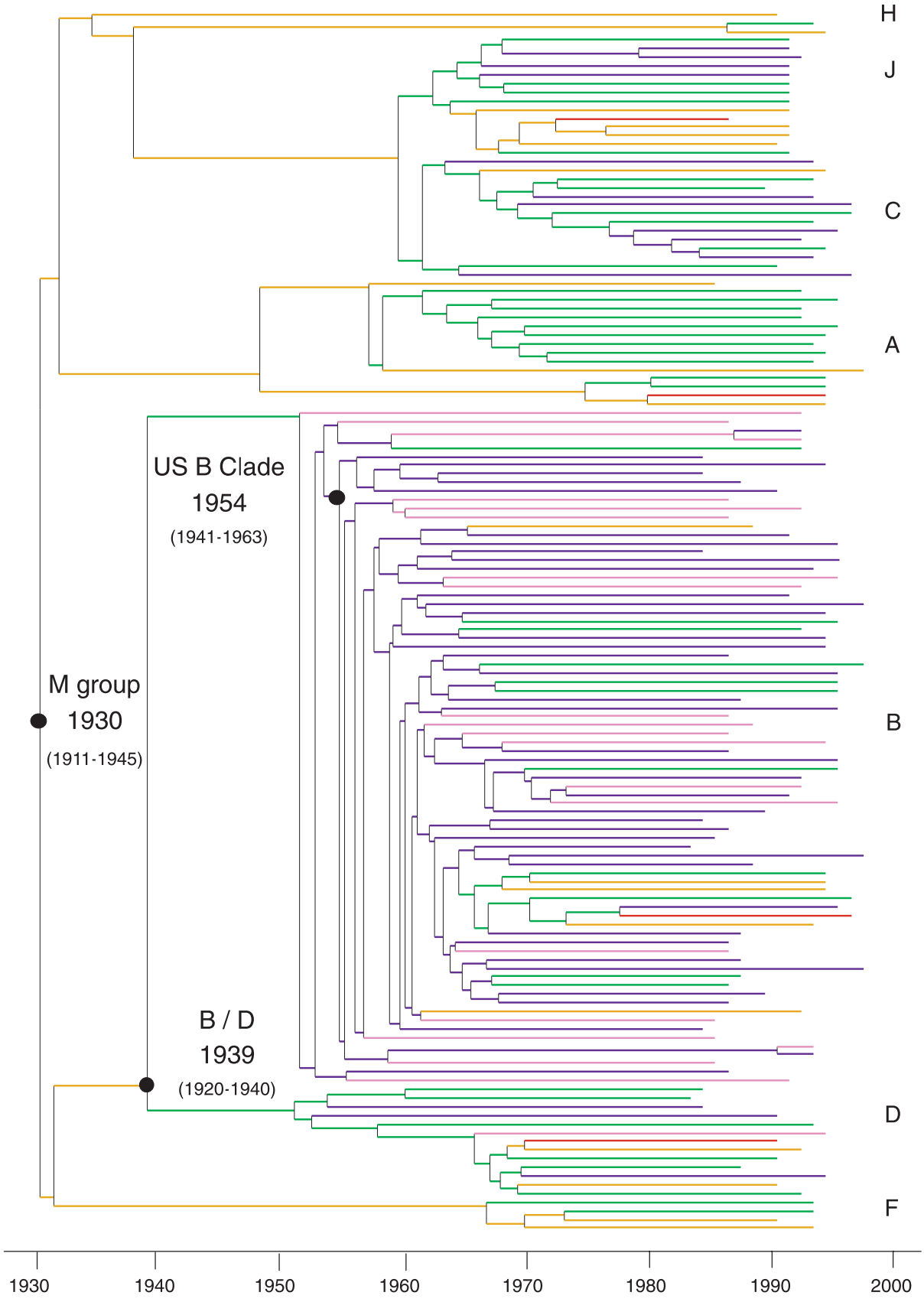
Serially Sampled Data

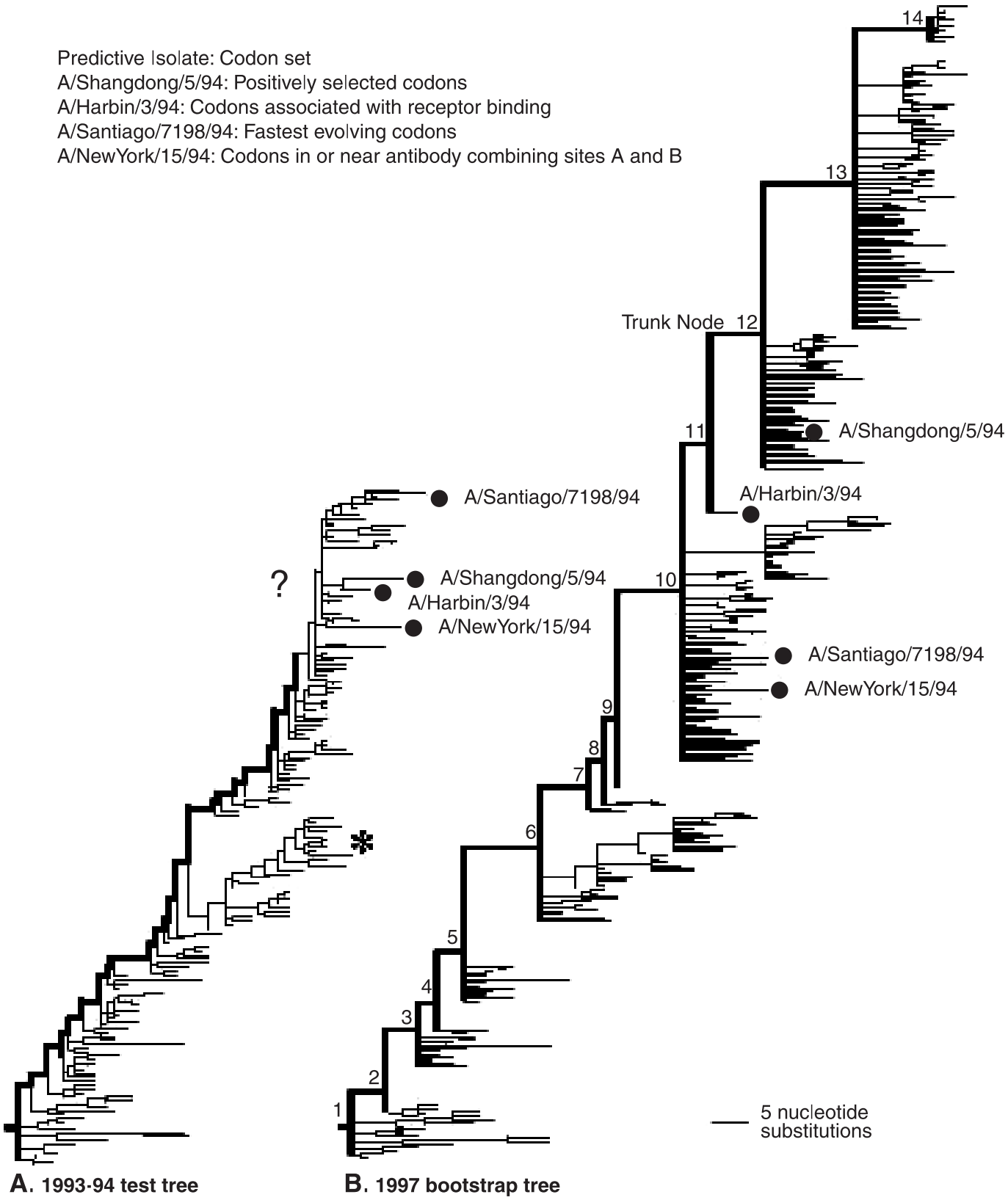
Drummond et al.2001. Mol.Biol.Evol.18:1365

HIV evolutionary rates before & after drug treatment as estimated from serially sampled sequence data



Korber et al.2000.Timing the Ancestor of the HIV-1 Pandemic Strains. Science 288:1789





Bush et al. 1999. Science 286:1921-1925

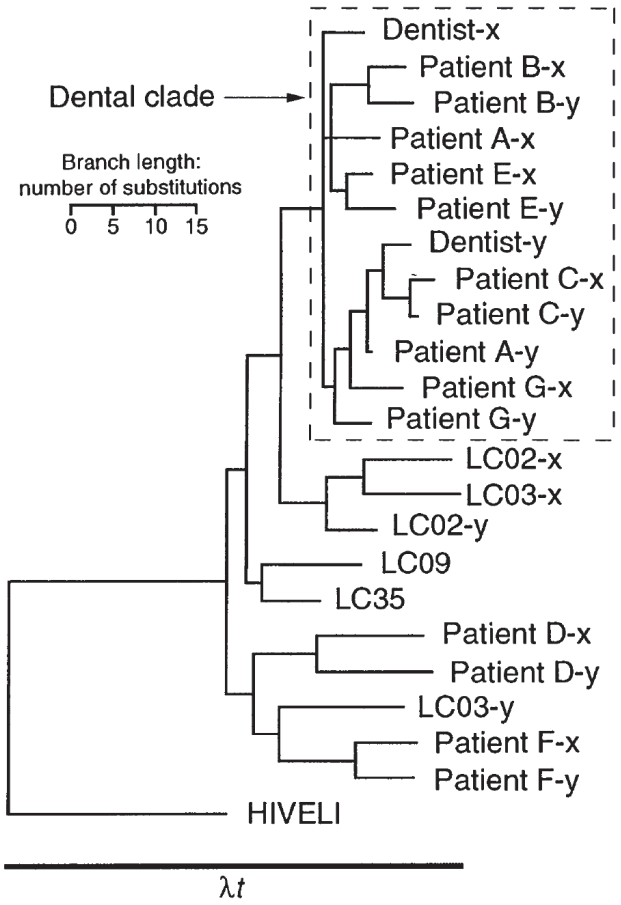


Fig. 3. Estimated phylogeny of HIV sequences from a Florida dentist, seven of his HIV-seropositive patients, and four individuals from the local population (LC) whose HIV sequences were most similar to those of the dentist (*47*). The outgroup (HIVELI) is an African HIV-1 sequence. Two divergent HIV sequences (labeled x and y) were examined from most individuals. The dental clade consists of patients whose HIV sequences are closer to those of the dentist than to those of any of the local controls. Branch lengths are proportional to the number of inferred evolutionary changes averaged across all possible character reconstructions (from *MacClade*) (*20*). The bar labeled λt is the distance from the root to the most divergent tip; it also indicates the divergence scale for the simulations in Fig. 4.

From Hillis et al. 1994