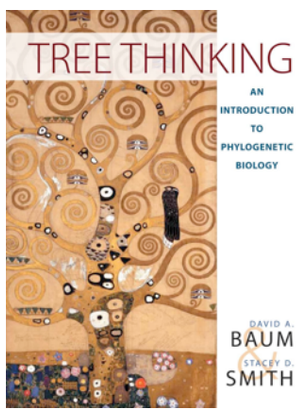
Phylogenetic Biology
Week 3

Biology 1425
Professor: Casey Dunn, dunnlab.org
Brown University
2013

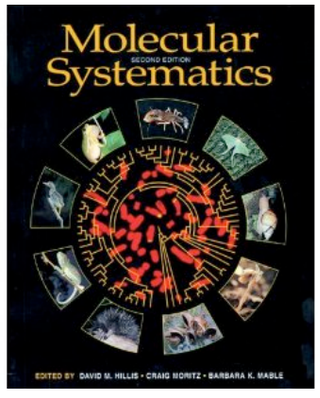
Front matter...

Sources

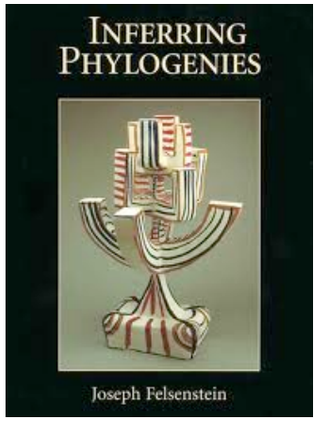
Some non-original content is drawn from:



Baum, D and S. Smith (2012) Tree Thinking: and Introduction to Phylogenetic Biology. Roberts and Company Publishers. ISBN 9781936221165



Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). Phylogenetic inference. In: Molecular Systematics, Second Edition. eds: D. M. Hillis, C Moritz, & B. K. Mable. Sinauer Associates. ISBN 9780878932825



Felsenstein, J. (2003) Inferring Phylogenies. Sinauer Associates. ISBN 978-0878931774

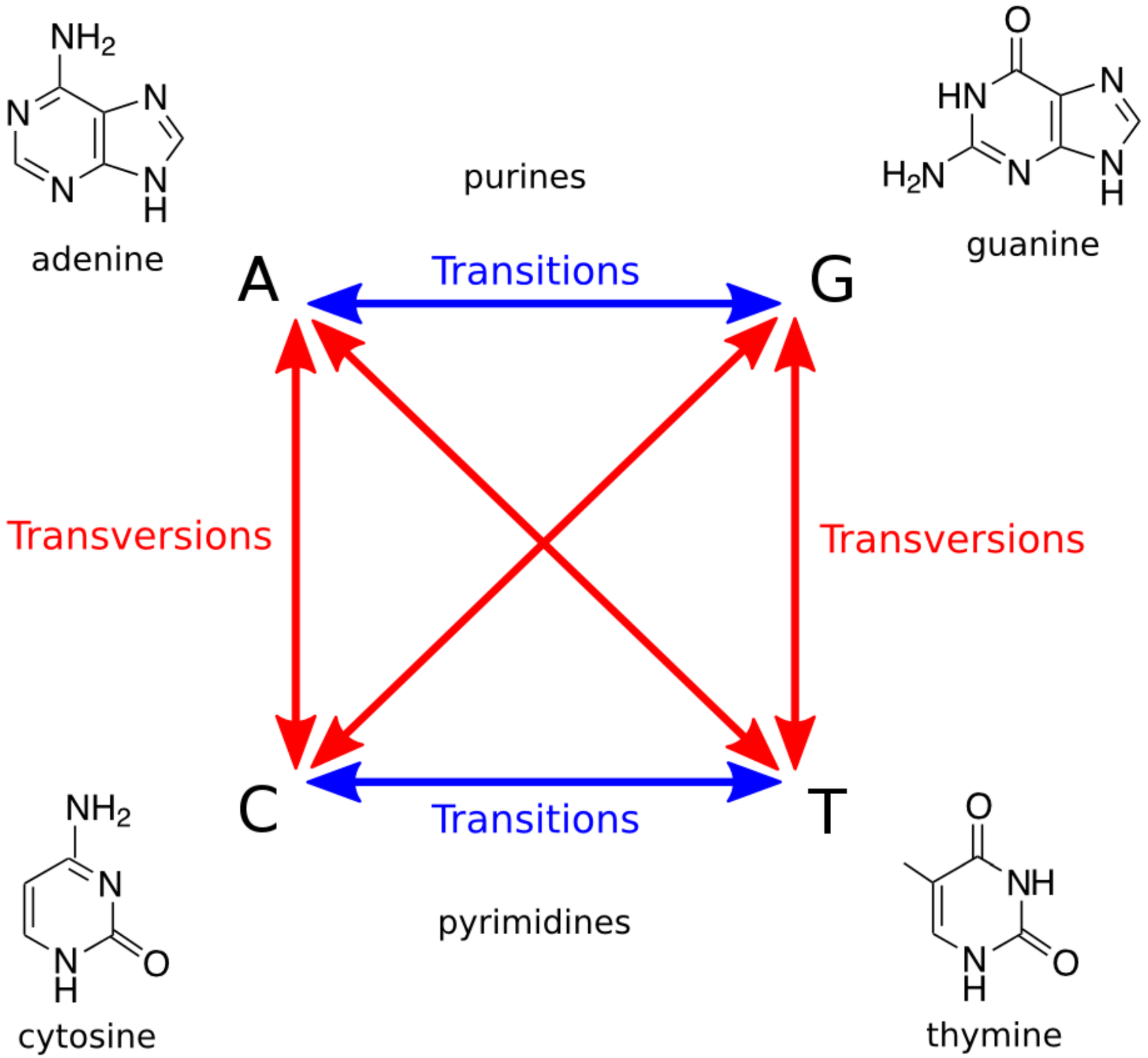
Other non-original content is referenced by url.

Other resources

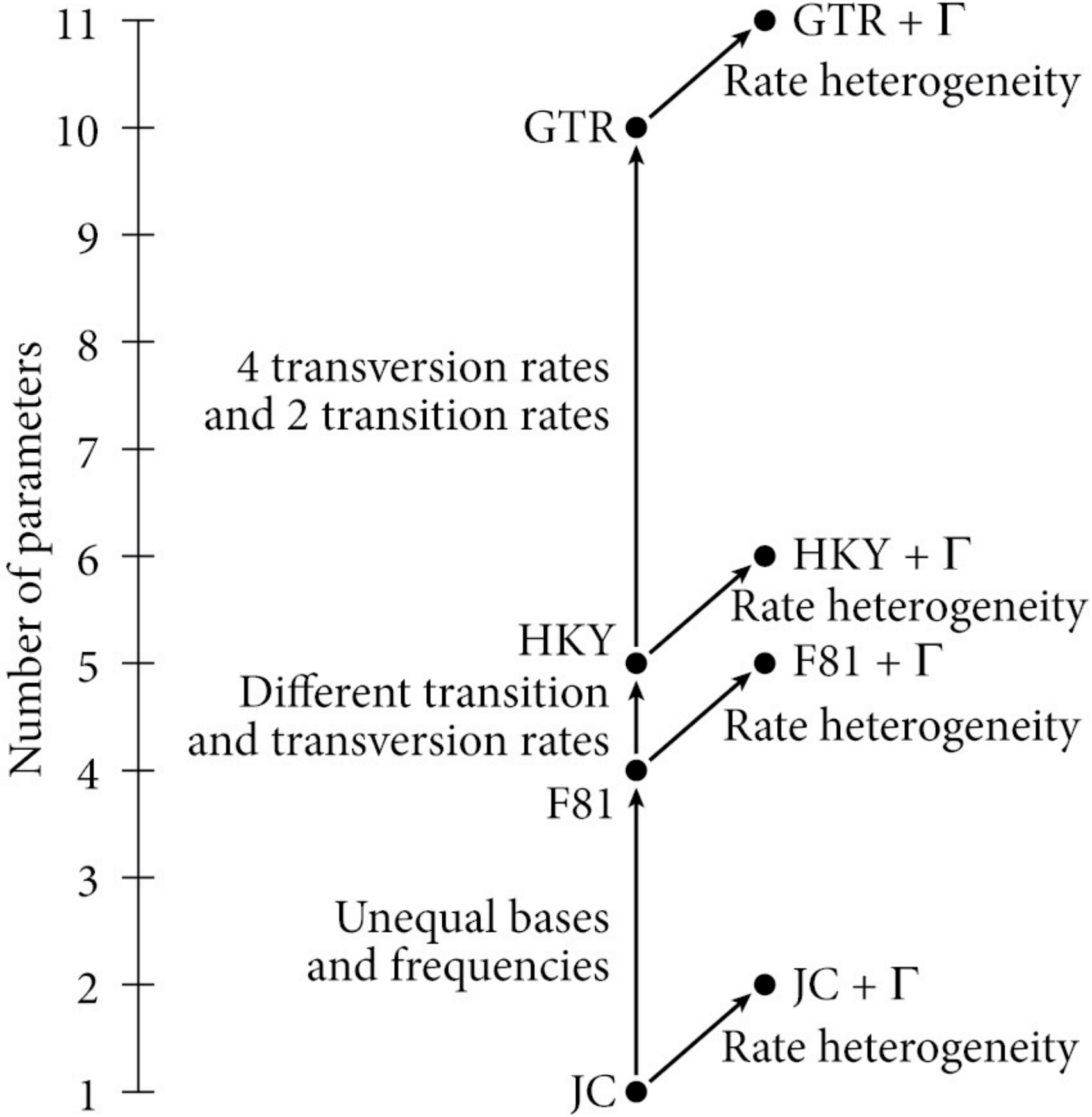
These slides supplement the following excellent presentations from the Wood's Hole Workshop in Molecular Evolution:

Paul Lewis - http://www.eeb.uconn.edu/people/plewis/downloads/wh2012/Likelihood_WoodsHole_24July2012_1-per-page.pdf

John Huelsenbeck - https://molevol.mbl.edu/wiki/images/3/37/WoodsHoleHandout.pdf



http://upload.wikimedia.org/wikipedia/commons/3/35/Transitions-transversions-v3.png



Baum and Smith 2012, Figure 8.10

Rate matrix

The instantaneous rate of a given substitutions

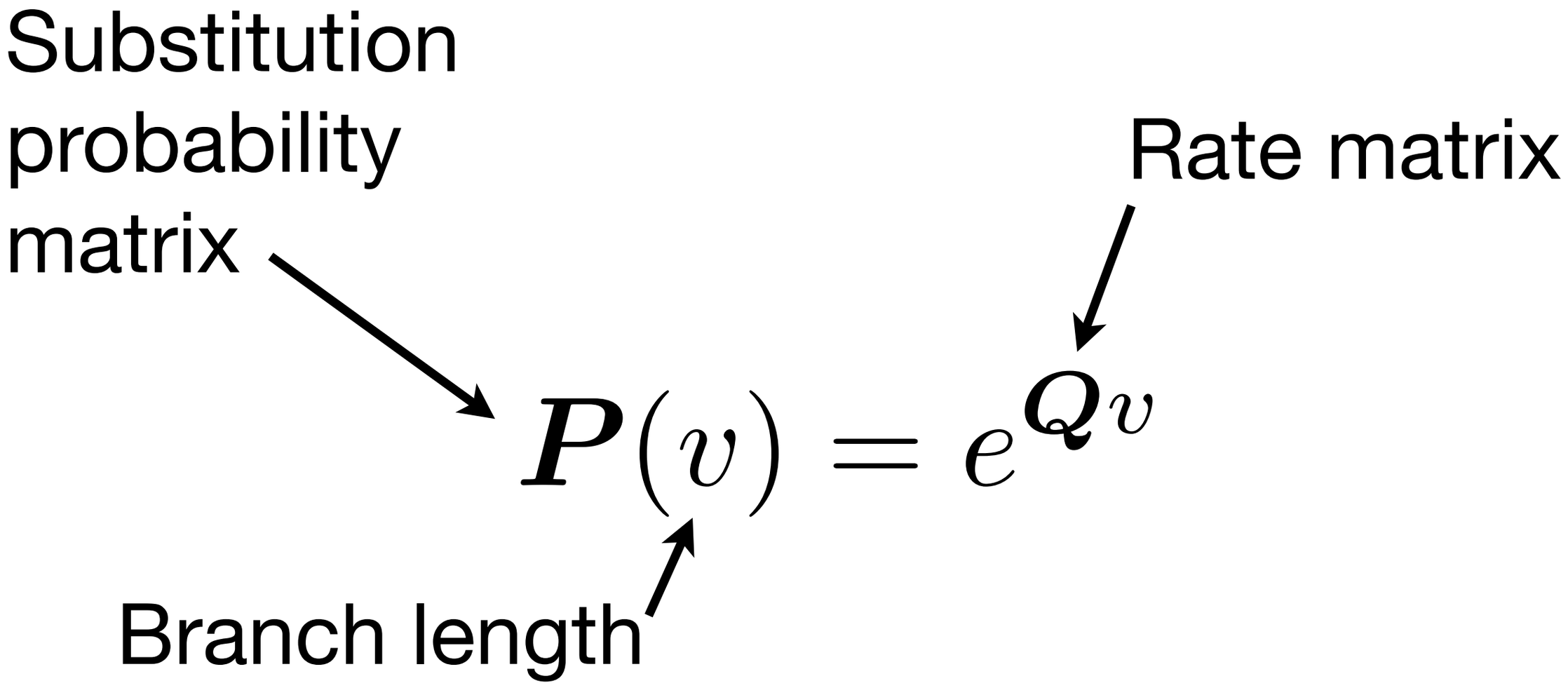
Q - Rate matrix

Substitution probability matrix

The probability of a given substitution occurring in a given interval (branch length). Because of reversals, there are an infinite number of histories that could have given rise to the particular substitution. Can be derived from the rate matrix.

$\boldsymbol{P}$ - Substitution probability matrix

Substitution probability matrix



This is called matrix exponentiation

F81 model

Q - Rate matrix

		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	$-m(\pi_C + \pi_G + \pi_T)$	$\pi_C m$	$\pi_G m$	$\pi_T m$
	C (freq = π_C)	$\pi_A m$	$-m(\pi_A + \pi_G + \pi_T)$	$\pi_G m$	$\pi_T m$
	G (freq = π_G)	$\pi_A m$	$\pi_C m$	$-m(\pi_A + \pi_C + \pi_T)$	$\pi_T m$
	T (freq = π_T)	$\pi_A m$	$\pi_C m$	$\pi_G m$	$-m(\pi_A + \pi_C + \pi_G)$

P - Substitution probability matrix

		To:			
		A	C	G	T
From:	A	$\pi_A + (1 - \pi_A)e^{-mt}$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	C	$\pi_A(1 - e^{-mt})$	$\pi_C + (1 - \pi_C)e^{-mt}$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	G	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G + (1 - \pi_G)e^{-mt}$	$\pi_T(1 - e^{-mt})$
	T	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T + (1 - \pi_T)e^{-mt}$

Baum and Smith 2012, Figures 8.7, 8.8

F81 model

As the branch length goes to 0, **P** becomes a diagonal matrix

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

As the branch length goes to infinity, the rows become the equilibrium base frequencies

$\boldsymbol{P}$

		To: A	To: C	To: G	To: T
From:	A	$\pi_A + (1-\pi_A)e^{-mt}$	$\pi_C(1-e^{-mt})$	$\pi_G(1-e^{-mt})$	$\pi_T(1-e^{-mt})$
	C	$\pi_A(1-e^{-mt})$	$\pi_C + (1-\pi_C)e^{-mt}$	$\pi_G(1-e^{-mt})$	$\pi_T(1-e^{-mt})$
	G	$\pi_A(1-e^{-mt})$	$\pi_C(1-e^{-mt})$	$\pi_G + (1-\pi_G)e^{-mt}$	$\pi_T(1-e^{-mt})$
	T	$\pi_A(1-e^{-mt})$	$\pi_C(1-e^{-mt})$	$\pi_G(1-e^{-mt})$	$\pi_T + (1-\pi_T)e^{-mt}$

Baum and Smith 2012, Figure 8.8

Likelihood

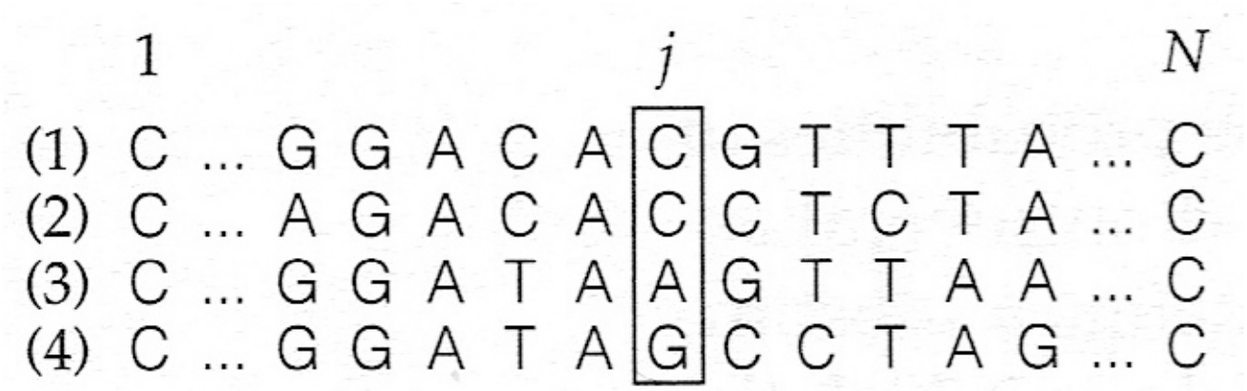
Likelihood is the probability of the data (D) given a hypothesis (H):

$$P(D|H)$$

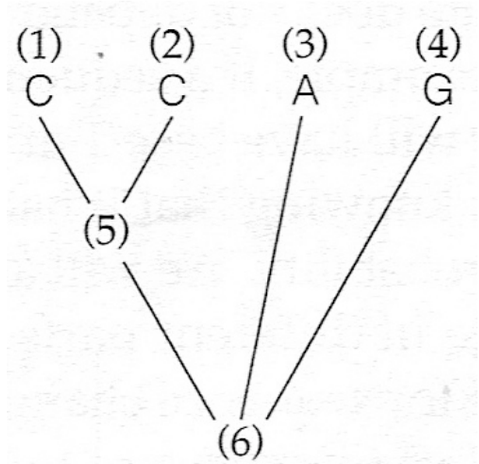
In our case, the data is our aligned matrix (the observed sequences and their inferred homologies) and the hypothesis is a particular tree and model of character evolution

Calculating likelihood

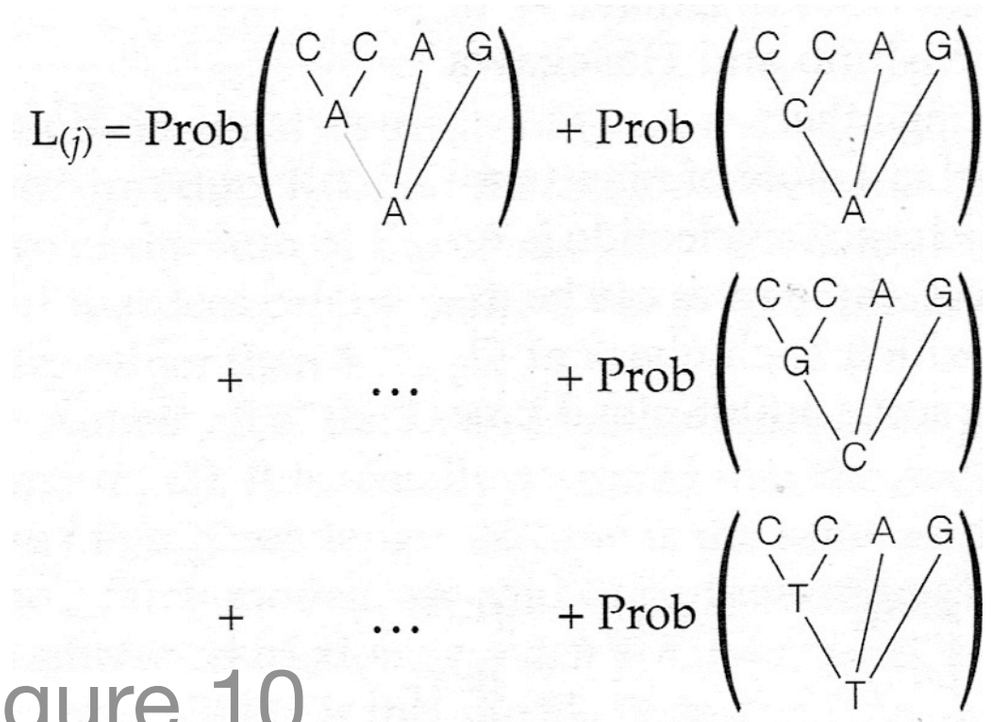
The data:



1 of 3 possible trees:



Likelihood for site j:



Swofford et al 1996, Figure 10

Calculating likelihood

The data:

	1							j						N
(1)	C	...	G	G	A	C	A	C	G	T	T	T	A	... C
(2)	C	...	A	G	A	C	A	C	C	T	C	T	A	... C
(3)	C	...	G	G	A	T	A	A	G	T	T	A	A	... C
(4)	C	...	G	G	A	T	A	G	C	C	T	A	G	... C

Likelihood of all sites:

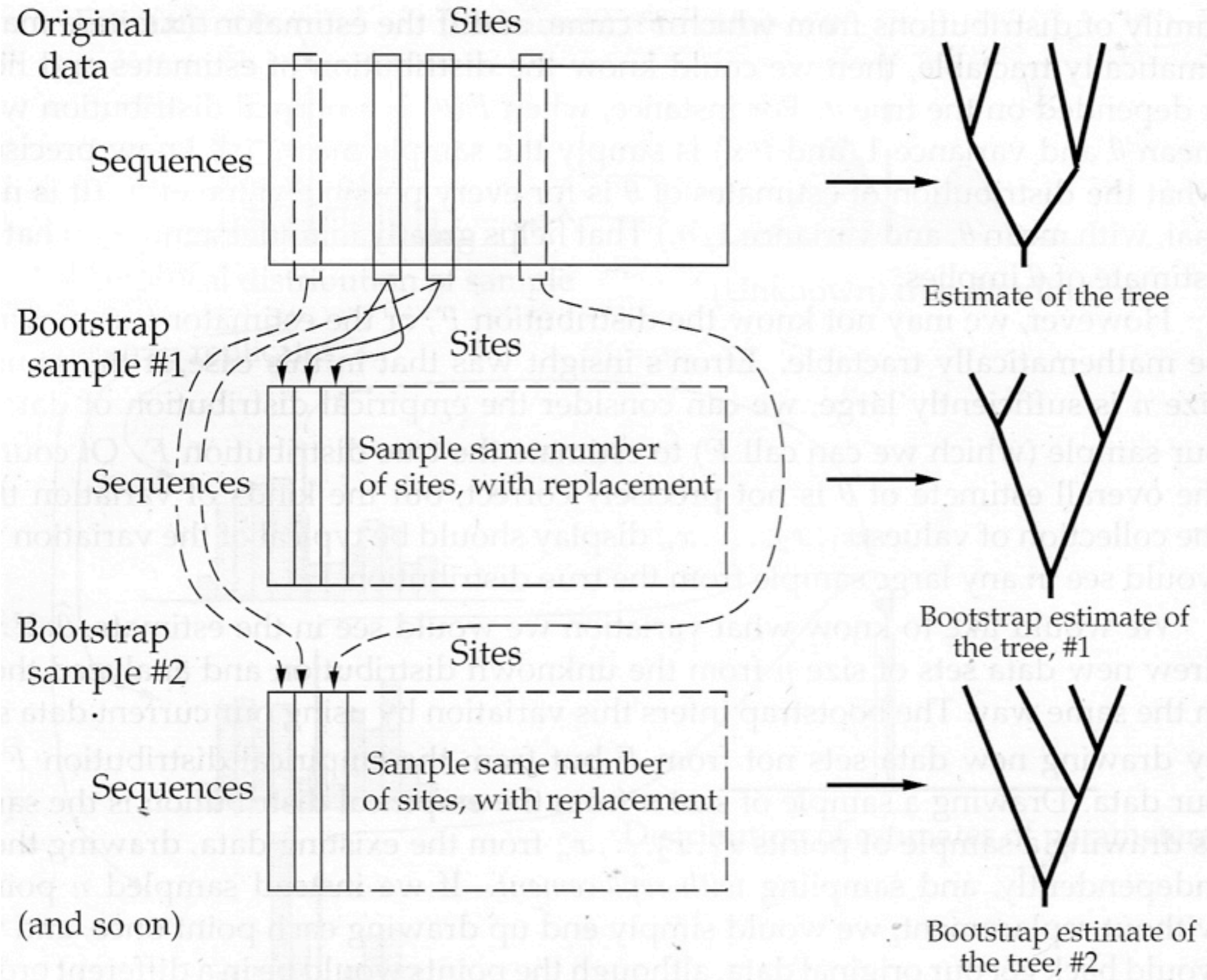
$$L = L_{(1)} \bullet L_{(2)} \bullet \ldots \bullet L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

Log likelihood of all sites:

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$$

Swofford et al 1996, Figure 10

Bootstraps



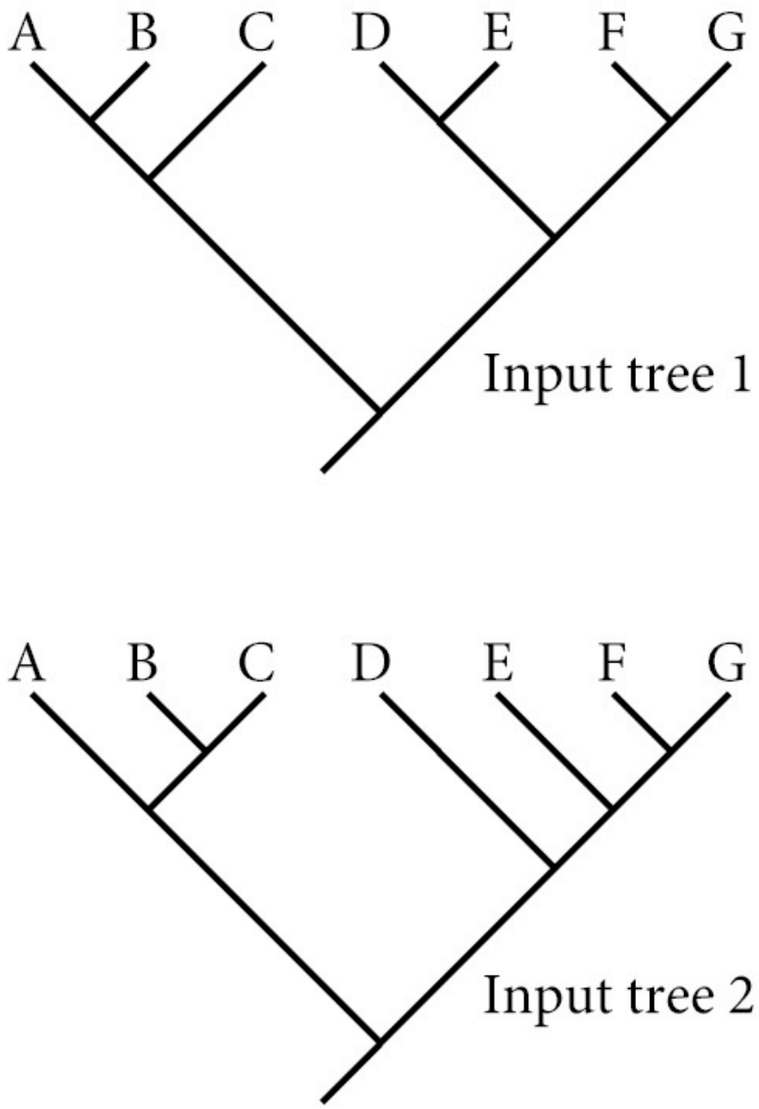
Felsenstein 2003, Figure 20.2

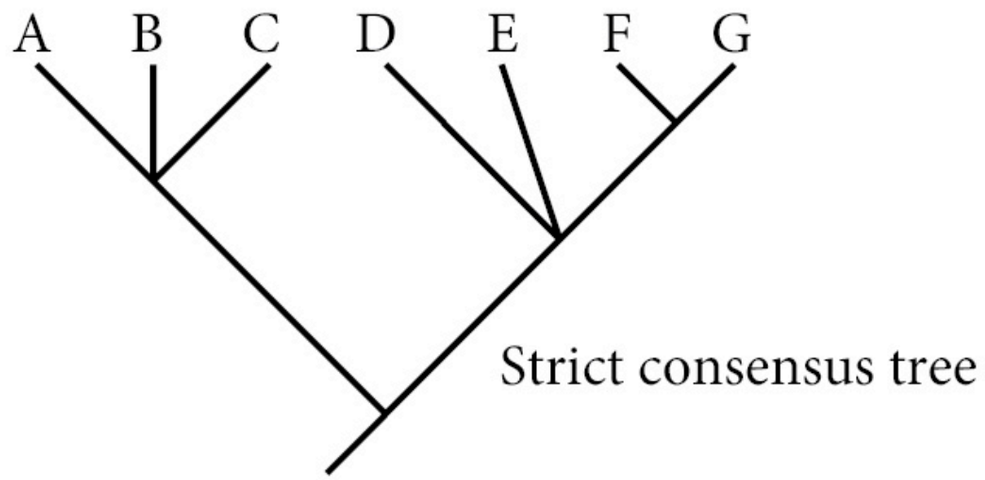
Consensus tree

A consensus tree summarizes a set of trees that have the same taxa

Consensus tree

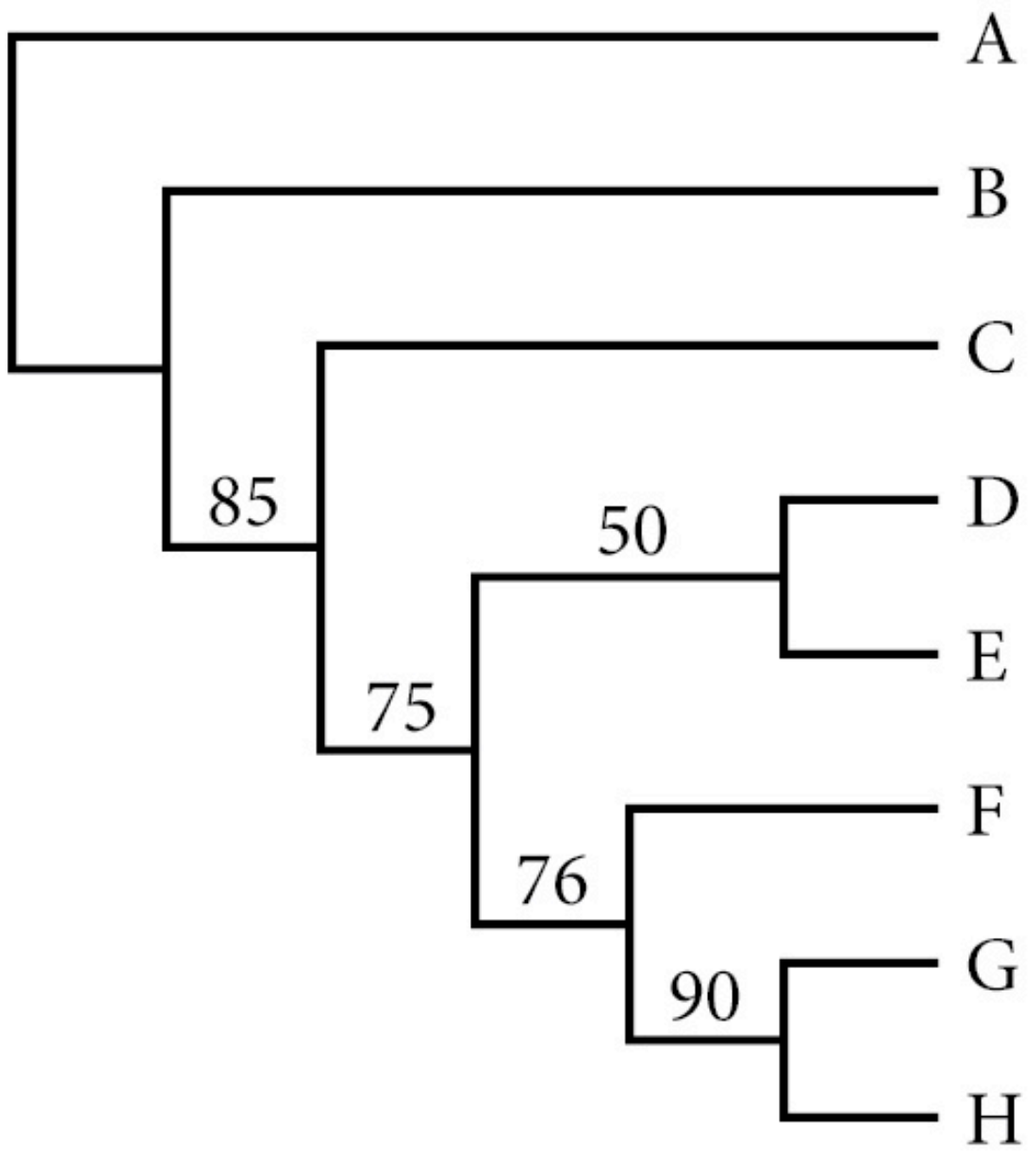
Strict





Baum and Smith 2012, Figure 3.26

Consensus tree



The number indicates the fraction of trees that include the corresponding edge (ie, branch).

Baum and Smith 2012, Figure 9.8