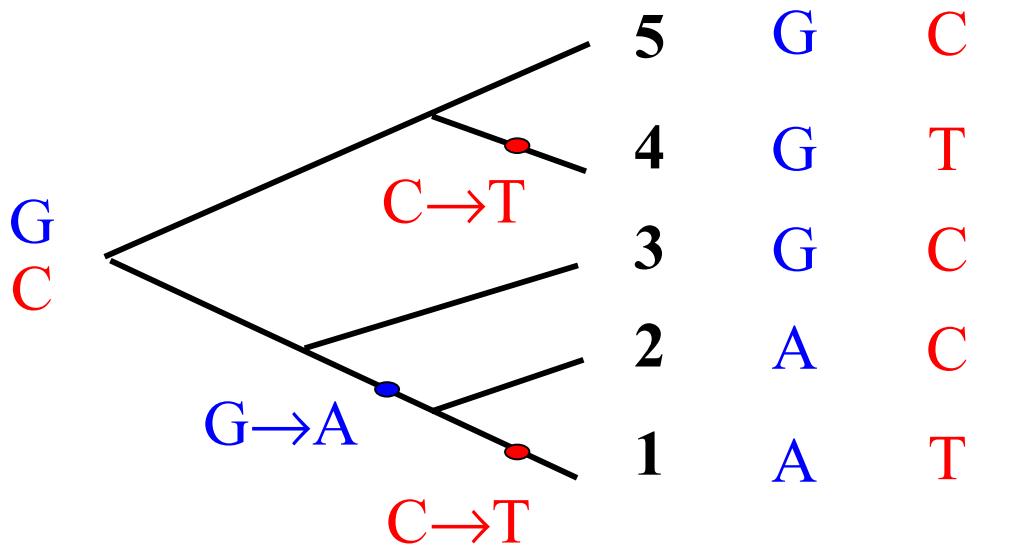
The statistical approach to molecular phylogeny: improved models

N. Galtier

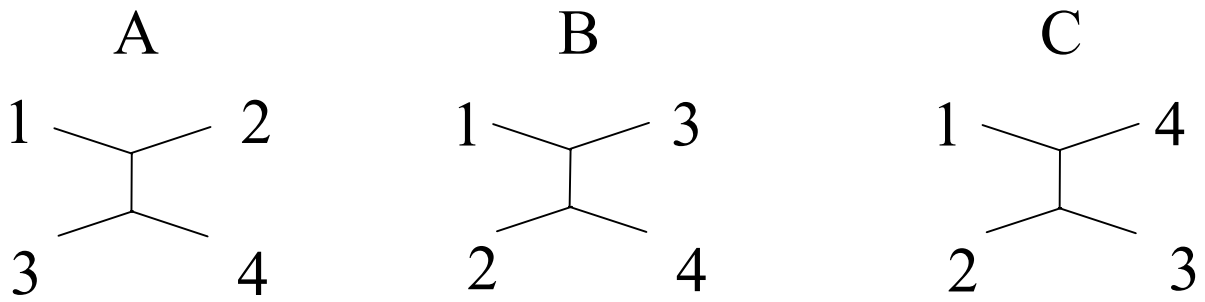
C.N.R.S. UMR 5000 – "Génome, Populations, Interactions"
Université Montpellier 2

galtier@univ-montp2.fr

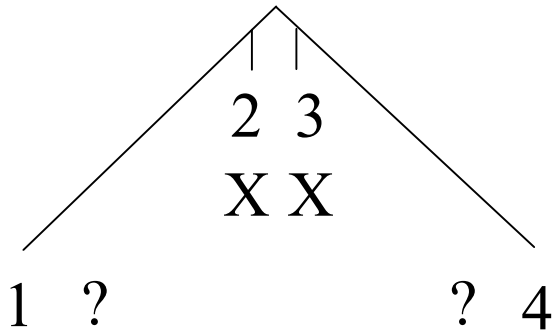
MOLECULAR PHYLOGENY AND SATURATION



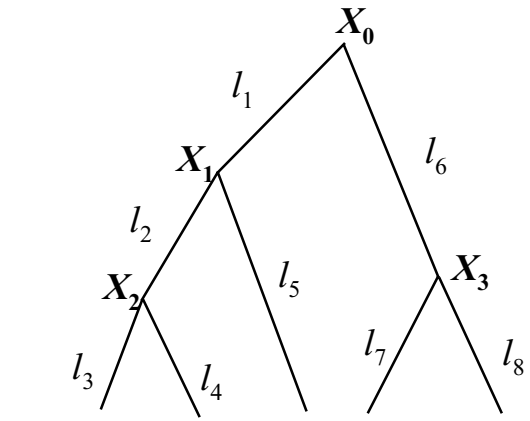
MAXIMUM PARSIMONY AND LONG BRANCH ATTRACTION



	a	b	c	d	e
1:	X	X	X	X	Y
2:	X	Y	Y	Y	X
3:	Y	X	Y	Z	X
4:	Y	Y	X	X	Z



LIKELIHOOD CALCULATION IN MOLECULAR PHYLOGENY



Data set : **Y**

	A	C	G	T
A		β	α	β
C	β		β	α
G	α	β		β
T	β	α	β	

Substitution rate matrix : **M**

$$\Pr(\mathbf{Y} \mid l_i, \mathbf{M}) = \prod_i \Pr(y_i \mid l_i, \mathbf{M})$$

$$\Pr(y_1 \mid l_i, \mathbf{M}) = \sum_{x_0}\sum_{x_1}\sum_{x_2}\sum_{x_3} \Pr(X_0=x_0).\Pr(X_1=x_1 \mid X_0=x_0).\ \Pr(X_2=x_2 \mid X_1=x_1).\ \Pr(y_{11}=\mathrm{A} \mid X_2=x_2).\ \Pr(y_{12}=\mathrm{A} \mid X_2=x_2).$$
$$\Pr(y_{13}=\mathrm{C} \mid X_1=x_1).\ \Pr(X_3=x_3 \mid X_0=x_0).\ \Pr(y_{14}=\mathrm{A} \mid X_3=x_3).\ \Pr(y_{15}=\mathrm{G} \mid X_3=x_3)$$

MODELING MARKOVIAN EVOLUTION ALONG A BRANCH

Let $\mathbf{M}=(m_{ij})$ be the rate matrix.
Let $\mathbf{F}(t)$ be the vector of state probabilities at time t.

The dynamics is described by:

$$\mathbf{F}(t+dt)=\mathbf{F}(t)+\mathbf{M}.\mathbf{F}(t).dt$$

$$d\mathbf{F}(t)/dt=\mathbf{M}.\mathbf{F}(t)$$

This linear system solves as:

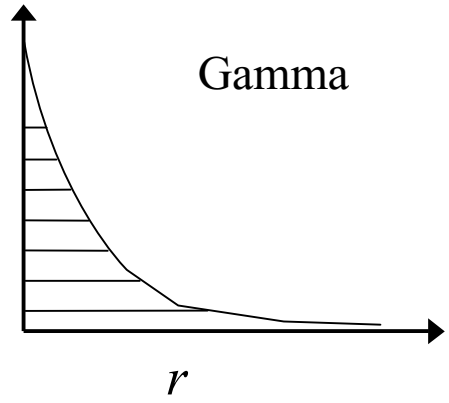
$$\mathbf{F}(t)=e^{\mathbf{M}.t}.\mathbf{F}(0)$$

Matrix $\mathbf{P}(t)=e^{\mathbf{M}t}$ gives the probabilities of state j after evolution during time (branch length) t according to model $\mathbf{M}$ given initial state i (required for likelihood calculation).
Taking the exponent of a matrix is best achieved by diagonalizing it:

$$\exp(\mathbf{Q}^{-1}.\mathrm{diag}(\lambda_i).\mathbf{Q}) = \mathbf{Q}^{-1}.\mathrm{diag}(\exp(\lambda_i)).\mathbf{Q}$$

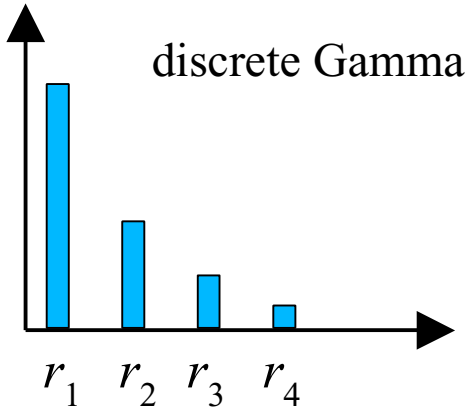
(where $\mathbf{Q}$ is any invertible matrix)

AMONG SITE RATE VARIATION



$$L(y) = \int f_{\Gamma}(r).L(y/r).dr$$

f_{Γ} : Gamma probability density function



$$L(y) = \sum_{i=1}^{g} \Pr(r = r_i).L(y/r_i)$$

g: assumed number of classes

The likelihood conditional on *r* is obtained by first multiplying branch lengths by *r*

EXAMPLE MODELS OF NUCLEOTIDE SUBSTITUTION

Jukes & Cantor 1969

	A	C	G	T
A	X	α	α	α
C	α	X	α	α
G	α	α	X	α
T	α	α	α	X

1 parameter
equiprobable changes

Kimura 1980

	A	C	G	T
A	X	α	$\kappa.\alpha$	α
C	α	X	α	$\kappa.\alpha$
G	$\kappa.\alpha$	α	X	α
T	α	$\kappa.\alpha$	α	X

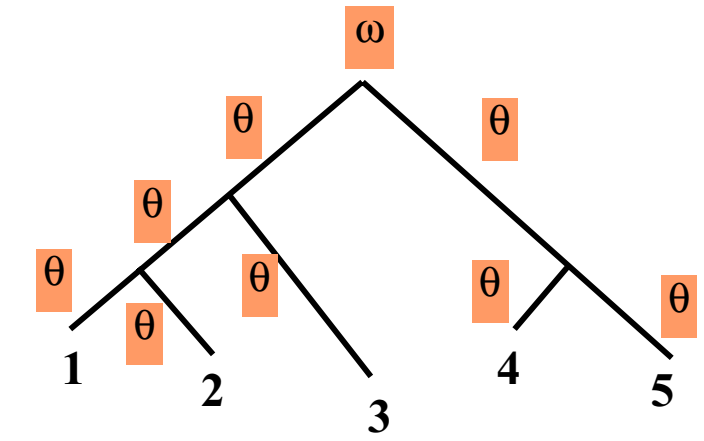
2 parameters
transition rate ≠
transversion rate

Tamura 1992

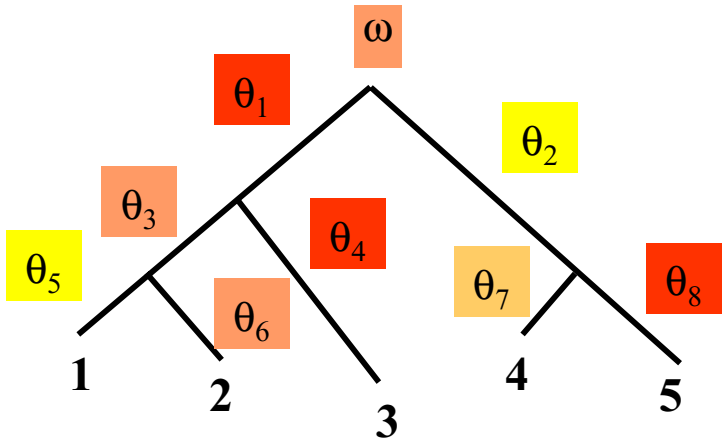
	A	C	G	T
A	X	$\alpha\frac{1-\theta}{2}$	$\kappa\alpha\frac{1-\theta}{2}$	$\alpha\frac{1-\theta}{2}$
C	$\alpha\frac{\theta}{2}$	X	$\alpha\frac{\theta}{2}$	$\kappa\alpha\frac{\theta}{2}$
G	$\kappa\alpha\frac{\theta}{2}$	$\alpha\frac{\theta}{2}$	X	$\alpha\frac{\theta}{2}$
T	$\alpha\frac{1-\theta}{2}$	$\kappa\alpha\frac{1-\theta}{2}$	$\alpha\frac{1-\theta}{2}$	X

3 parameters
stationary GC% = $\theta \neq 50\%$

A NON-HOMOGENEOUS, NON-STATIONARY MODEL

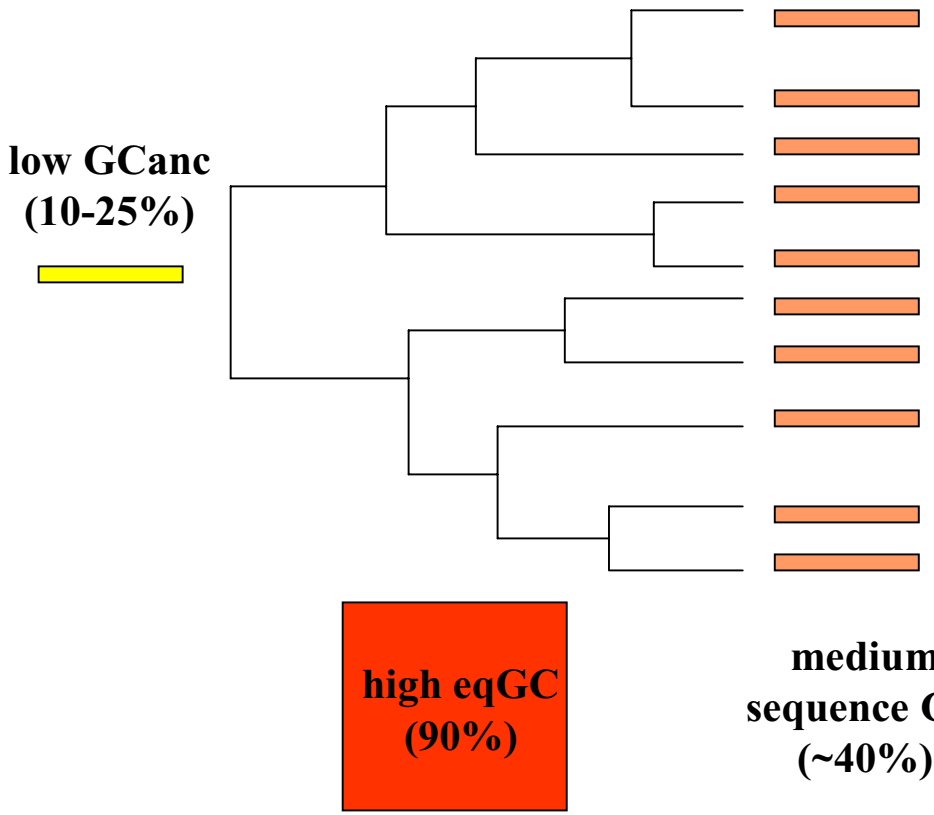


stationary,
homogeneous



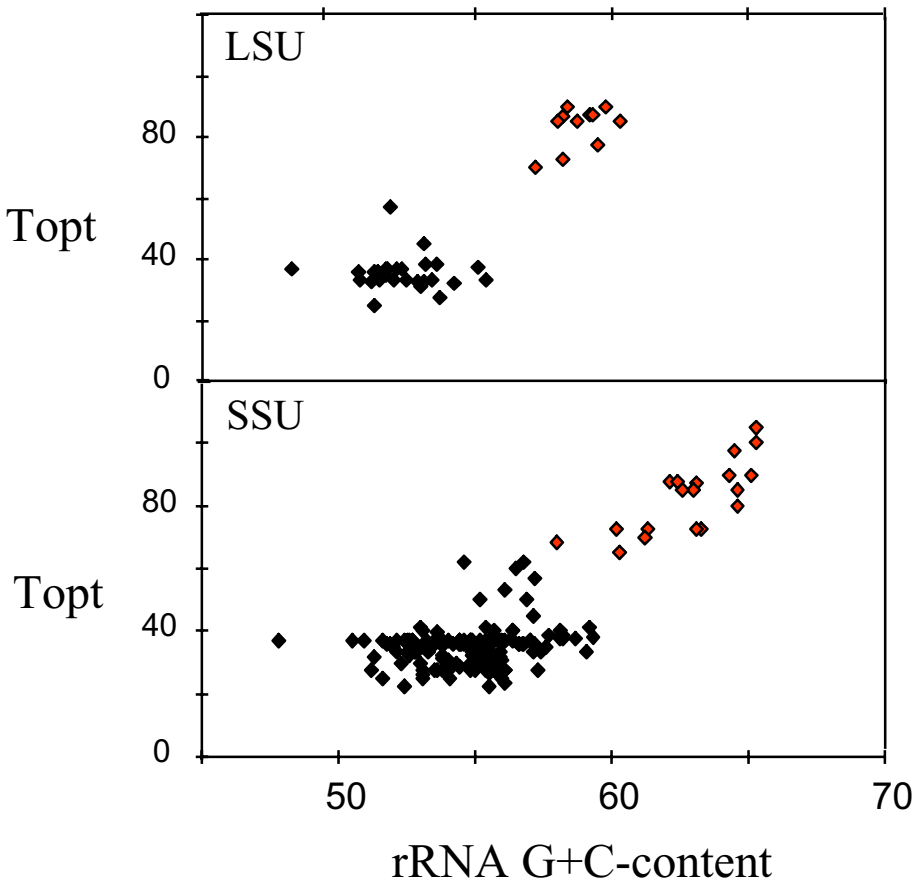
non-stationary,
non-homogeneous

ACCURACY OF ANCESTRAL GC% ESTIMATION (SIMULATIONS)

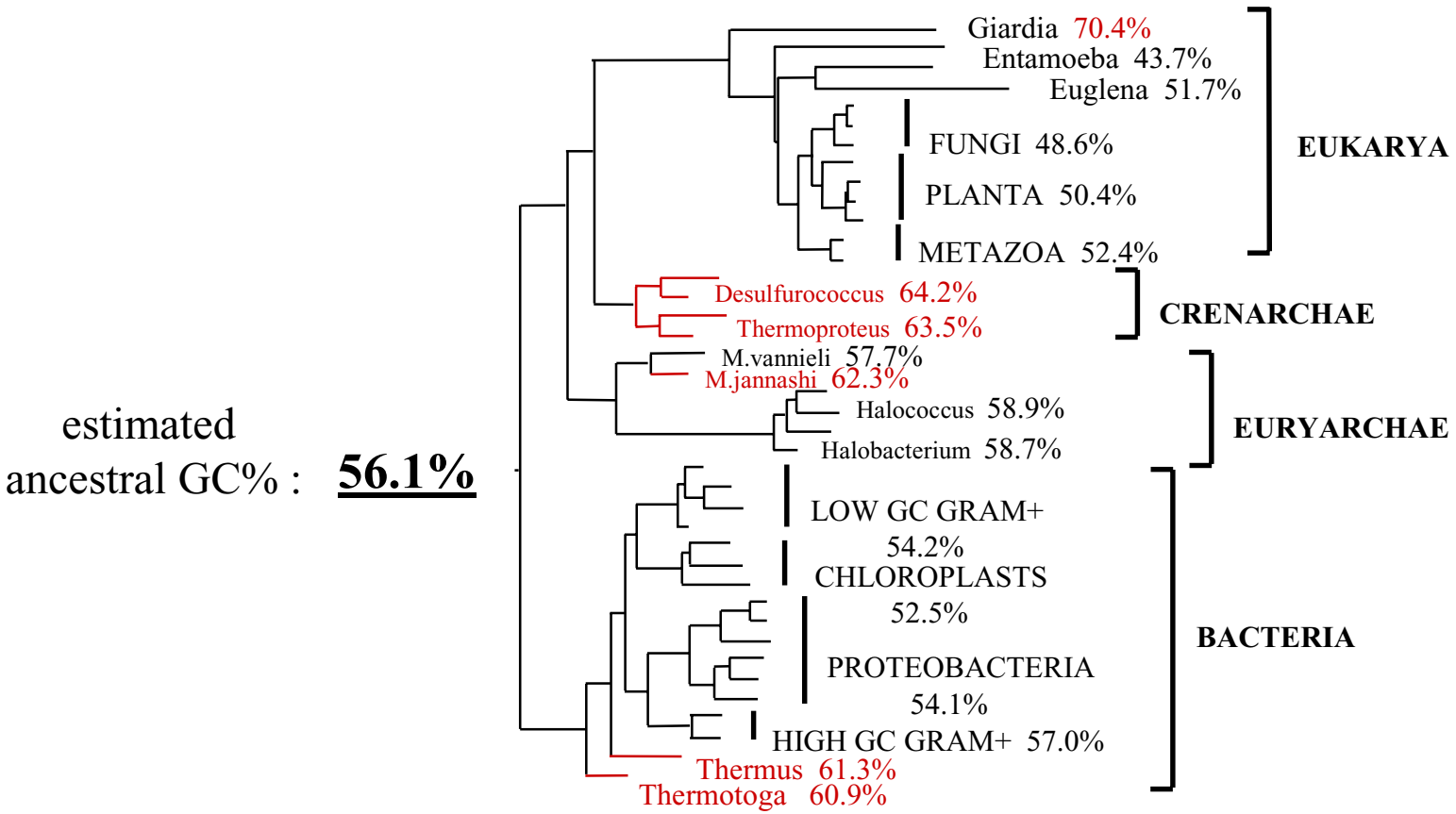


actual	MP	**NHML**
18%	32%	**19%**
10%	27%	**11%**
22%	40%	**21%**
14%	30%	**16%**
14%	28%	**15%**

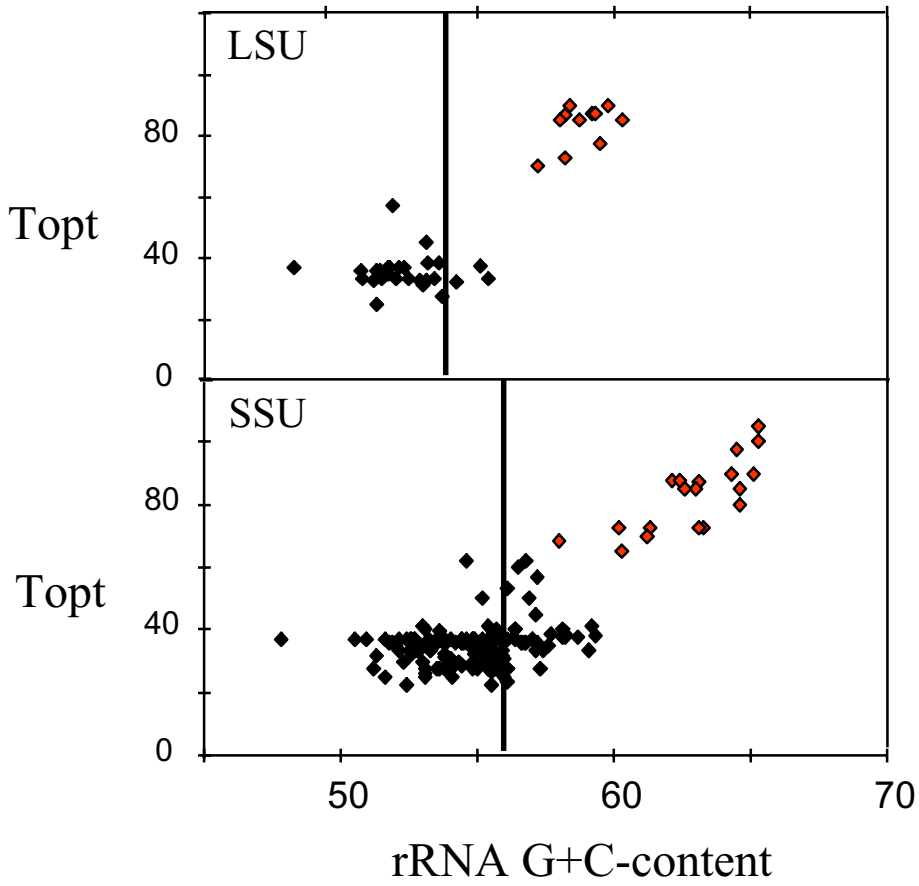
OPTIMAL GROWTH TEMPERATURE vs rRNA GC% IN PROKARYOTES



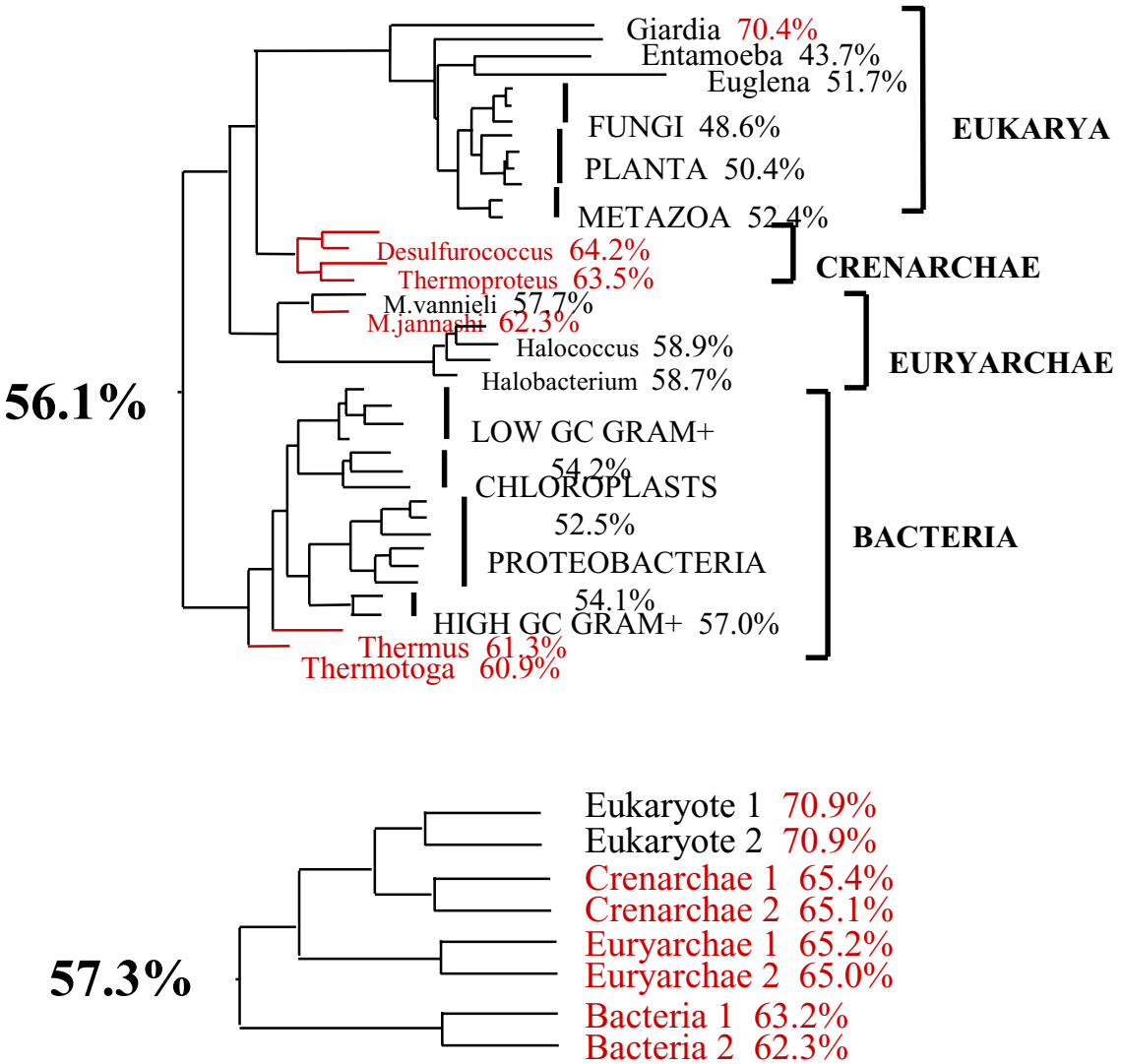
THE rRNA UNIVERSAL TREE OF LIFE



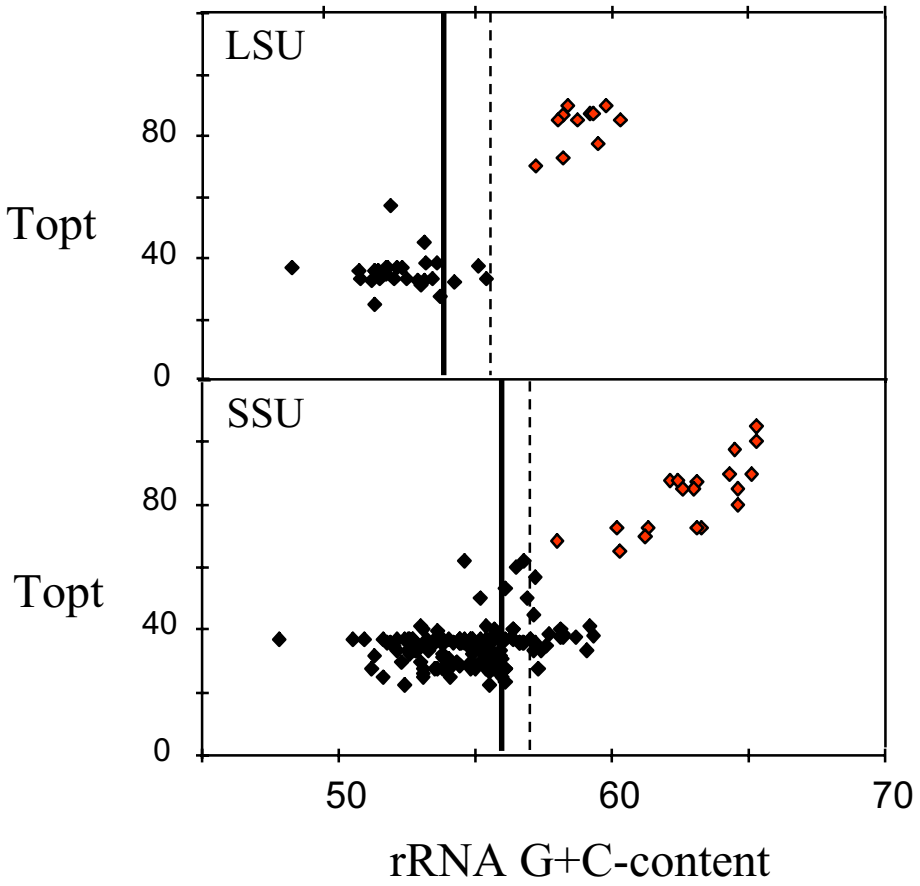
A NON-HYPERTHERMOPHILIC ANCESTOR ?



CONTROL FOR SPECIES SAMPLING

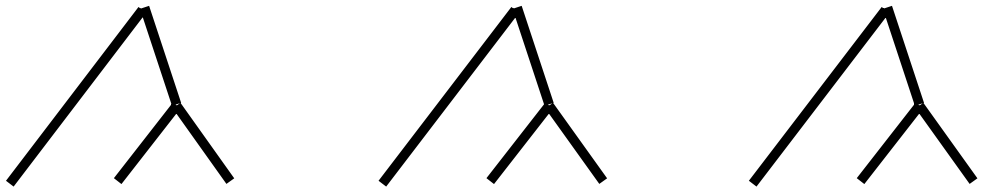


A NON-HYPERTHERMOPHILIC ANCESTOR ?

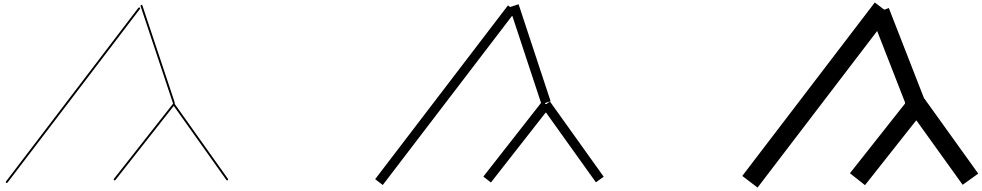


THREE MODELS OF RATE DISTRIBUTION AMONG SITES/LINEAGES

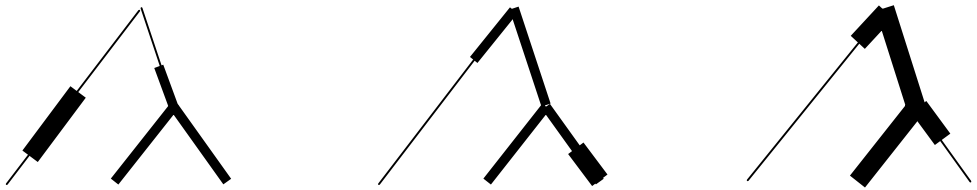
Constant rate among sites



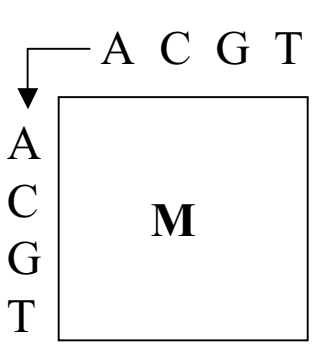
Variable rates between sites



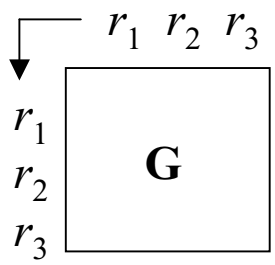
Site-specific rate variation = COVARIONS



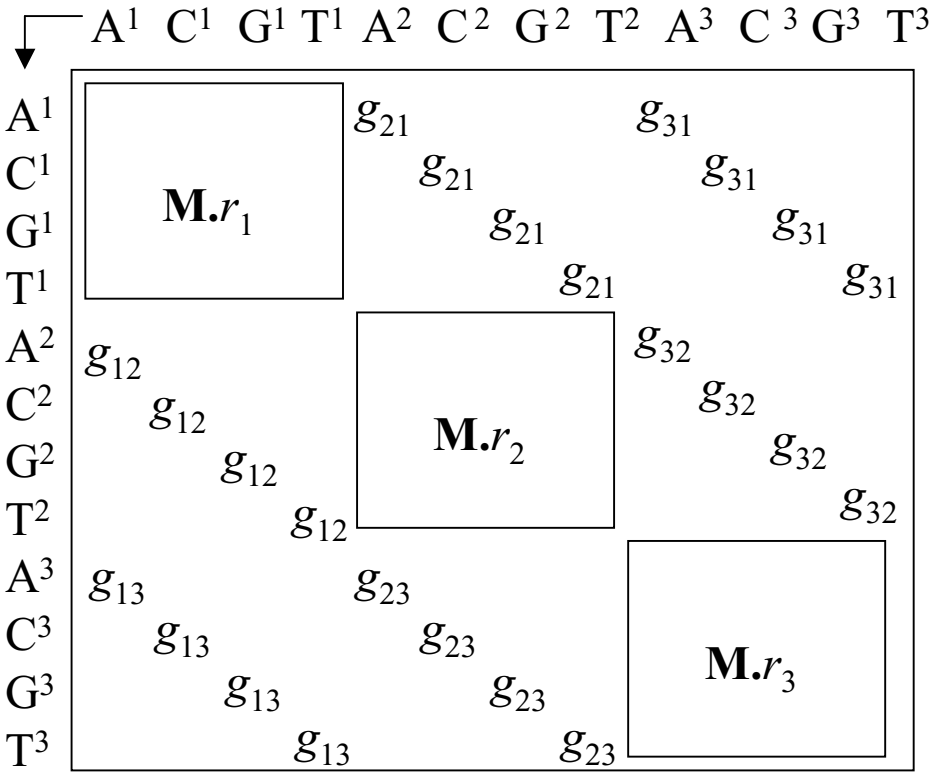
MARKOV-MODULATED MARKOV CHAINS AND COVARION



Substitution rate matrix



Rate-change rate matrix



Compound process rate matrix

$$\mathbf{R} = (r_1, r_2, r_3)$$

Vector of rate categories

$$\mathbf{Q} = \mathrm{diag}(\mathbf{R}) \otimes \mathbf{M} + \mathbf{G} \otimes \mathbf{I}_4$$

MODELING COVARION : A BRIEF BIBLIOGRAPHY

- Fitch (1971, JME) introduces the concept of "covarion"

- Tuffley & Steel (1998, Math. Biosci.), reviewed by Penny et al (2001, JME) formalize Fitch's model (2 rate classes, "on/off")

- Galtier (2001, MBE) extends it to an arbitrary number of Gamma-distributed rate classes

- Galtier & Jean-Marie (submitted, JCB) extend it to any **G** and **M** matrices, introduce the Kronecker formalism, and provide an efficient diagonalization algorithm.

DIAGONALIZING THE COMPOUND MATRIX

The compound matrix $\mathbf{Q} = \mathrm{diag}(\mathbf{R})\otimes\mathbf{M} + \mathbf{G}\otimes\mathbf{I}_4$ has size $g.m$, where m is the number of states (4, 20, 61) and g the number of rate classes (4-10).

Standard diagonalization algorithms have complexity $o(n^4)$, n being the matrix size. Diagonalizing **Q** can therefore become limiting for likelihood calculation.

Let $\mathbf{B}\in\Re^m$ be a left eigenvector of **M** for eigenvalue $\lambda\in\Re$.

Define matrix **H** as: $\mathbf{H} = \mathbf{G} + \lambda.\mathrm{diag}(\mathbf{R})$

Let $\mathbf{A}\in\Re^g$ be an eigenvector of **H** for eigenvalue $\mu\in\Re$.

We show that $\mathbf{V}=\mathbf{A}\otimes\mathbf{B}$ is an eigenvector of **Q** for eigenvalue μ.

DIAGONALIZING THE COMPOUND MATRIX (2)

This leaves us with an algorithm for diagonalizing $\mathbf{Q}$ (size $m.g$) using the spectral decomposition of $\mathbf{M}$ (size m) and $\mathbf{G}$ (size g):

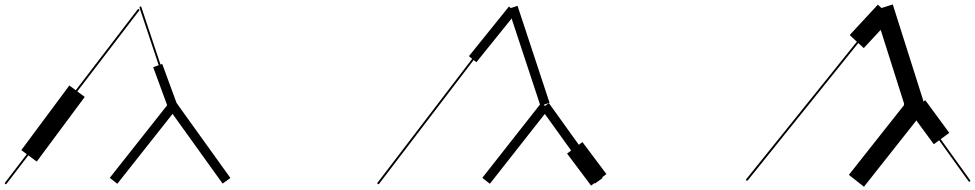
1. diagonalize $\mathbf{M}$, and record pairs $(\lambda^i, \mathbf{B}^i)$ of eigen-elements $(0<i\leq m)$

2. for every i,
   2.1. diagonalize $\mathbf{G}+\lambda^i.\mathbf{D_R}$ and record pairs $(\mu^{ij}, \mathbf{A}^{ij})$ of eigen-elements $(0<j\leq g)$.
   2.2. form g pairs of eigen-elements for $\mathbf{Q}$: $(\mu^{ij}, \mathbf{A}^{ij}\otimes \mathbf{B}^i)$.

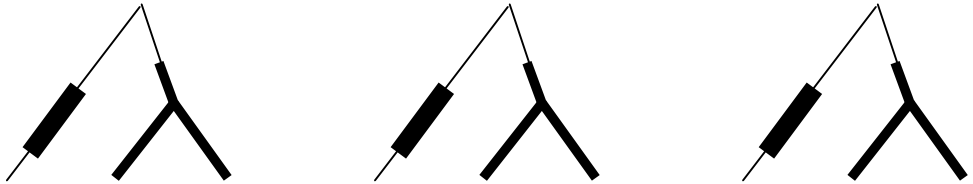
This is done in $o(m^4)+o(m)o(g^4)$, to be compared to the standard $o(m^4)o(g^4)$

COVARION vs MOLECULAR CLOCK

"Covarion" models involve site-specific changes of evolutionary rates.



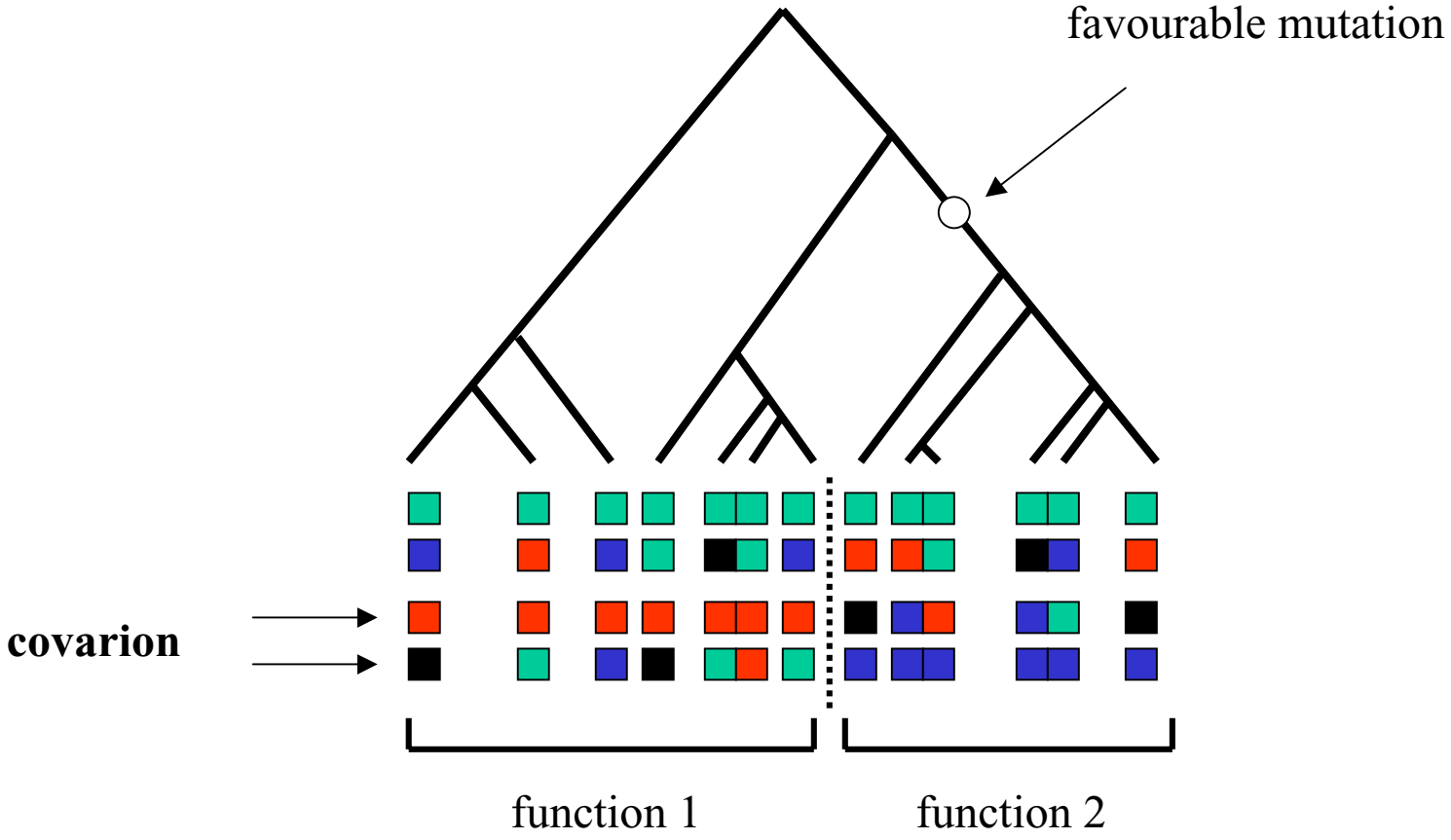
This is distinct from collective changes of evolutionary rates, that is, departure from the so-called **molecular clock**.

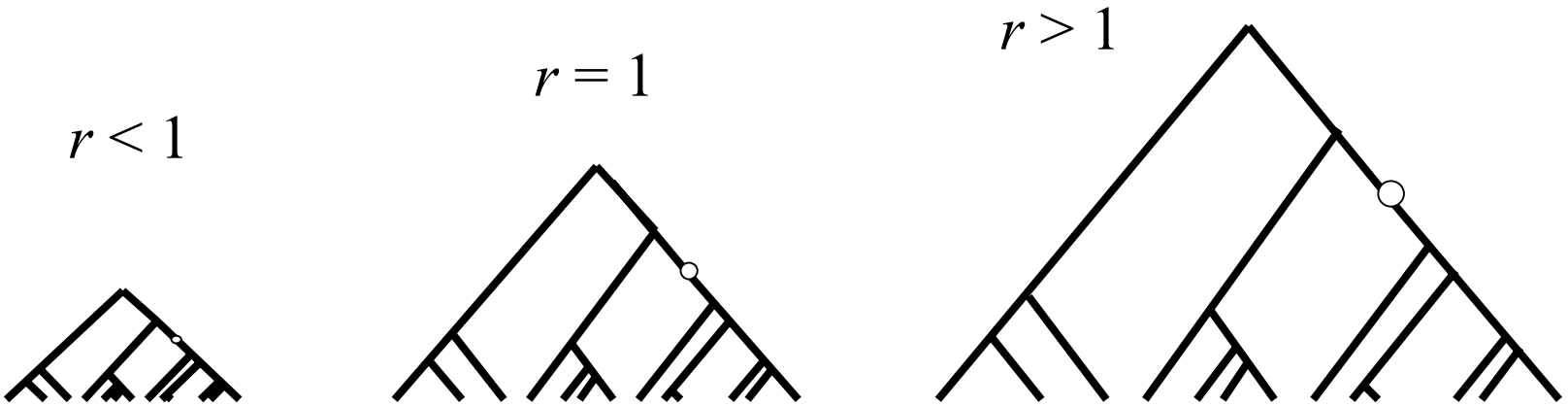


Departure from the clock is modelled very similarly to covarion (*e.g.* Huelsenbeck et al 2000 Genetics), but is hardly tractable in the likelihood framework (nonindependent sites).

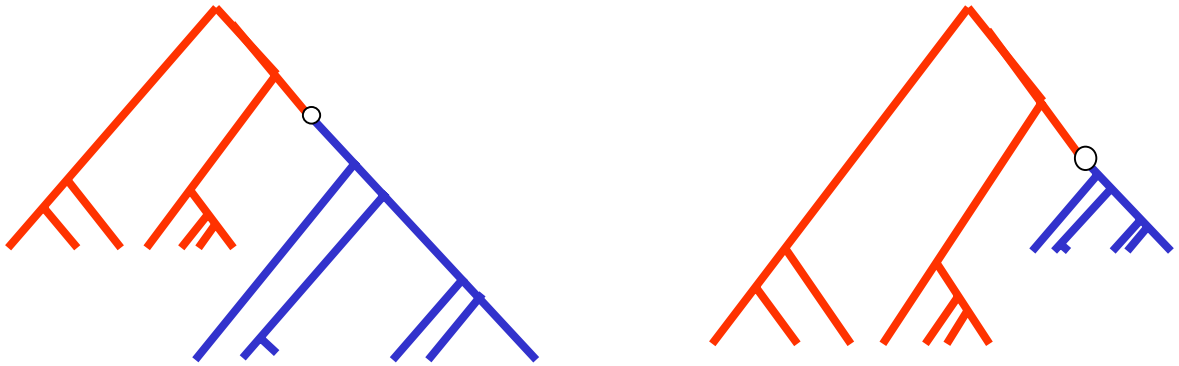
COVARION AND ANCESTRAL GC% ESTIMATION

	ER	ASRV	SSRV
params.	158	159	160
ln(L)	-21488.3	-20302.6	-20034.7
Ts/Tv	2.52	2.81	3.07
Gamma		0.650	0.247
covarion			1.825
tree lg	3.211	3.825	4.142
GCanc	**55.7%**	**53.8%**	**53.0%**

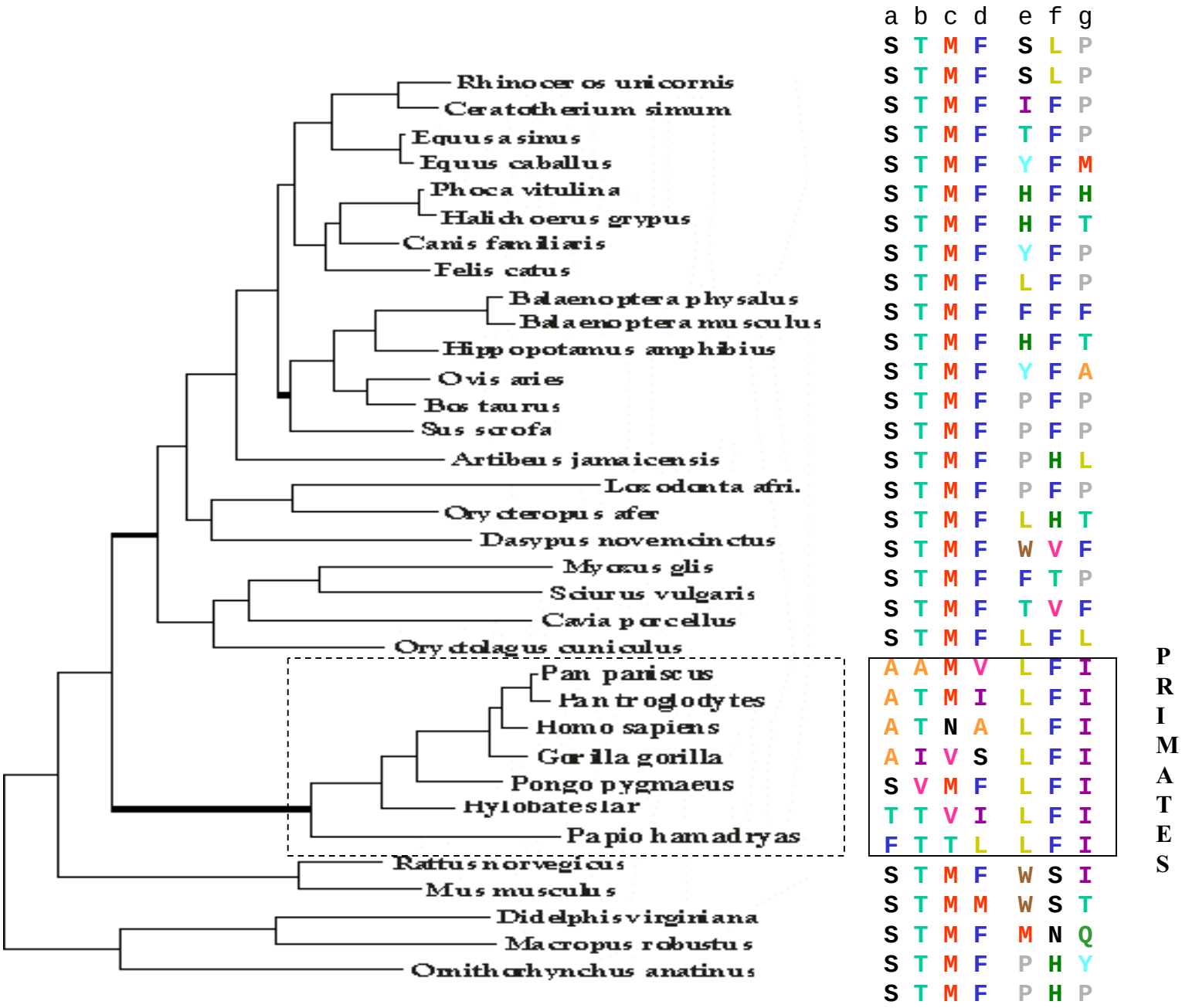

favourable mutation
covarion
function 1
function 2



$$LR = 2 \cdot [\ln(\max_{r_1, r_2}(L_1(r_1).L_2(r_2))) - \ln(\max_{r}(L_1(r).L_2(r)))] \sim \chi^2(1)$$



$r_2 > r_1$ $\quad$ $r_1 > r_2$


Rhinoceros unicornis
Ceratotherium simum
Equus asinus
Equus caballus
Phoca vitulina
Halichoerus grypus
Canis familiaris
Felis catus
Balaenoptera physalus
Balaenoptera musculus
Hippopotamus amphibius
Ovis aries
Bos taurus
Sus scrofa
Artibeus jamaicensis
Loxodonta afri.
Orycteropus afer
Dasypus novemcinctus
Myoxus glis
Sciurus vulgaris
Cavia porcellus
Oryctolagus cuniculus
Pan paniscus
Pan troglodytes
Homo sapiens
Gorilla gorilla
Pongo pygmaeus
Hylobates lar
Papio hamadryas
Rattus norvegicus
Mus musculus
Didelphis virginiana
Macropus robustus
Ornithorhynchus anatinus
a b c d e f g
S T M F S L P
S T M F S L P
S T M F I F P
S T M F T F P
S T M F Y F M
S T M F H F H
S T M F H F T
S T M F Y F P
S T M F L F P
S T M F F F F
S T M F H F T
S T M F Y F A
S T M F P F P
S T M F P F P
S T M F P H L
S T M F P F P
S T M F L H T
S T M F W V F
S T M F F T P
S T M F T V F
S T M F L F L
A A M V L F I
A T M I L F I
A T N A L F I
A I V S L F I
S V M F L F I
T T V I L F I
F T T L L F I
S T M F W S I
S T M M W S T
S T M F M N Q
S T M F P H Y
S T M F P H P
PRIMATES

MODEL SELECTION IN MOLECULAR PHYLOGENY

Which model should be used in a molecular phylogeny analysis?

The popular Akaike's criterion is defined as:

$$\mathbf{AIC=-2\ln(}\boldsymbol{L}\mathbf{) + 2}\boldsymbol{k}$$

where L is the (maximum) likelihood, and k the number of parameters of certain model of interest.

The model minimising AIC optimises the trade-off between fit and parameter number. Likelihood-ratio tests are a related approach.

But is this what we want?

MODEL SELECTION: AN EMPIRICAL APPROACH

Data: SSU + LSU rRNA from 26 bacterial species (3120 sites).

Two trees were used:

- "true tree", obtained from an NJ-analysis of rRNA data
- "wrong tree", obtained by disrupting a well-supported group of the "true tree"

Model	nb. param.	AIC	lnL(true)-lnL(wrong)
ER, homogeneous	126	85554.5	37.7
ASRV, homogeneous	127	80554.5	25.8
SSRV, homogeneous	128	**79800.2**	26.3
ER, non-homogeneous	254	86627.2	**39.5**