‘Realism’ and ‘Instrumentalism’ in models of molecular evolution

David Penny

Montpellier, June 08

Galileo

Overview

sites free to vary

summing sources of error

‘rates’ of molecular evolution

estimates of time intervals

do we know anything? (flat priors)

ALLAN
WILSON
CENTRE

Human/chimp divergence

1) *Ramapithecus* = 12Ma → HC = 5±1Ma

But *Ramapithecus* in Asia, HCG in Africa.

Is 18-20Ma a better estimate for divergence?

2) *Ramapithecus* = 18Ma → HC = 7.5±1.5Ma

Or should we combine uncertainties?

ALLAN WILSON CENTRE

In this case, I would rather not – leave it as a conditional estimate – need both.

sites free to vary

	rate $k_{aa} \times 10^9$/yr
- fibrinopeptides	8.3
- lysozyme	2.0
- hemoglobin α	1.2
- cytochrome c	0.3
- histone H4	0.01

Dickerson, 1971

explained the differences by the proportion of sites ‘free to vary’.

change of function should show a rate change

ALLAN WILSON CENTRE

realism

we use a tiny fraction of the information in the data

Alignment original sequence order	Reordered Alignment shuffled/reordered
AIIFLNSALGPSPELFPIILATKVL	ASAGPSPPATPLLIIIIILLFFNEKV
AIMFLNSALGPPTELFPVILATKVL	ASAGPPTPATPLLIMVILLFFNEKV
SIMFLNHTLNPTPELFPIILATETL	SHTNPTPPATPLLIMIILLFFNEET
TILFLNSSLGLQPEVTPTVLATKTL	TSSGLQPPATPLLILTVLVTFNEKT
TLLFLNSMLKPPSELFPIILATKTL	TSMKPPSPATPLLLLIILLFFNEKT
ALLFLNSTLNPPTELFPLILATKTL	ASTNPPTPATPLLLLLILLFFNEKT
AILFLNSFLNPPKEFFPIILATKIL	ASFNPPKPATPLLILIILFFFNEKI

c columns

c! alignments

If c = 1000, we use ≈ 1/ 1000! of the information

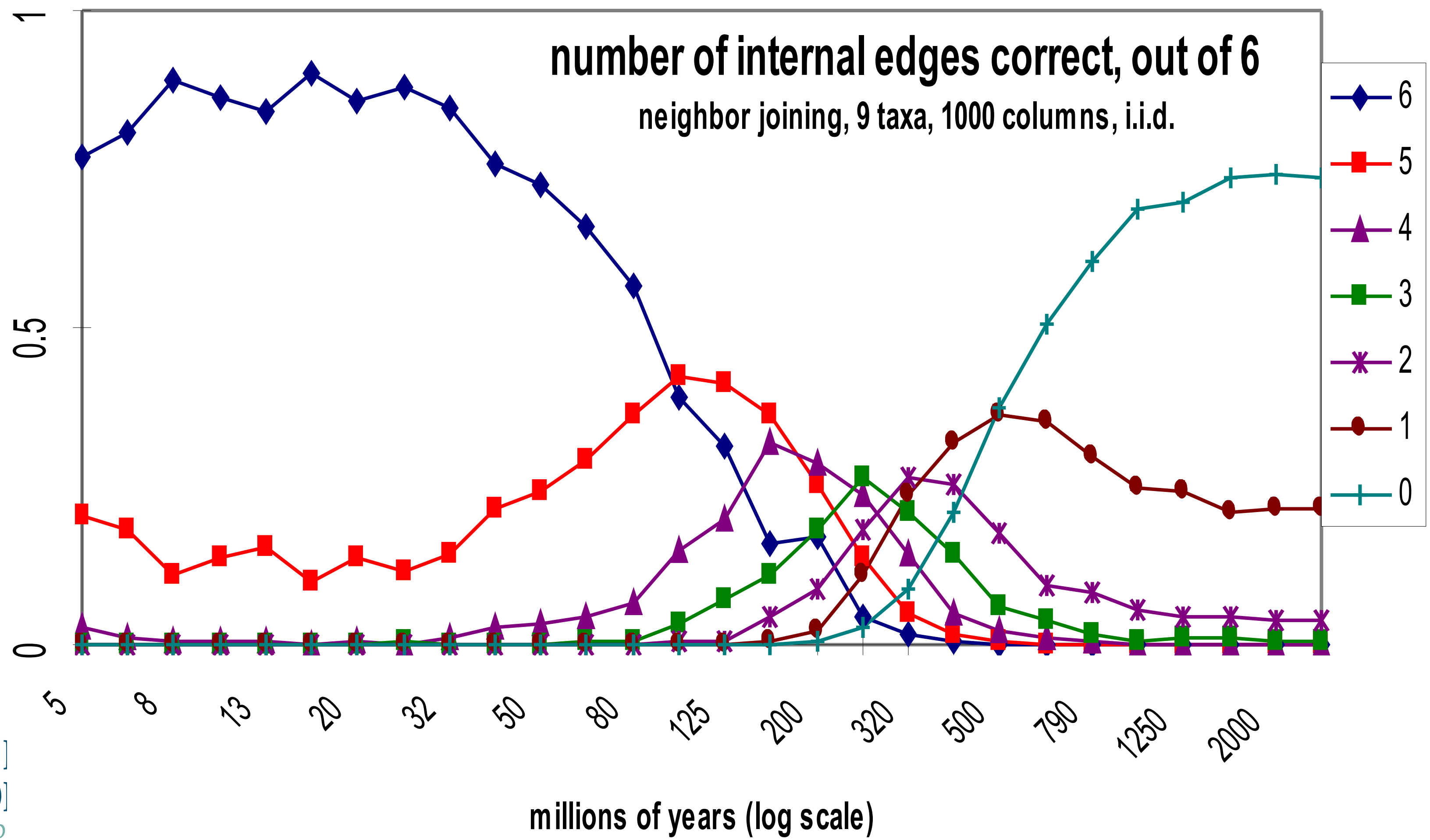
sites change

X-ray crystallographers: the strongest conclusion we have is that the same sites in different species may be fixed, in others they are variable.

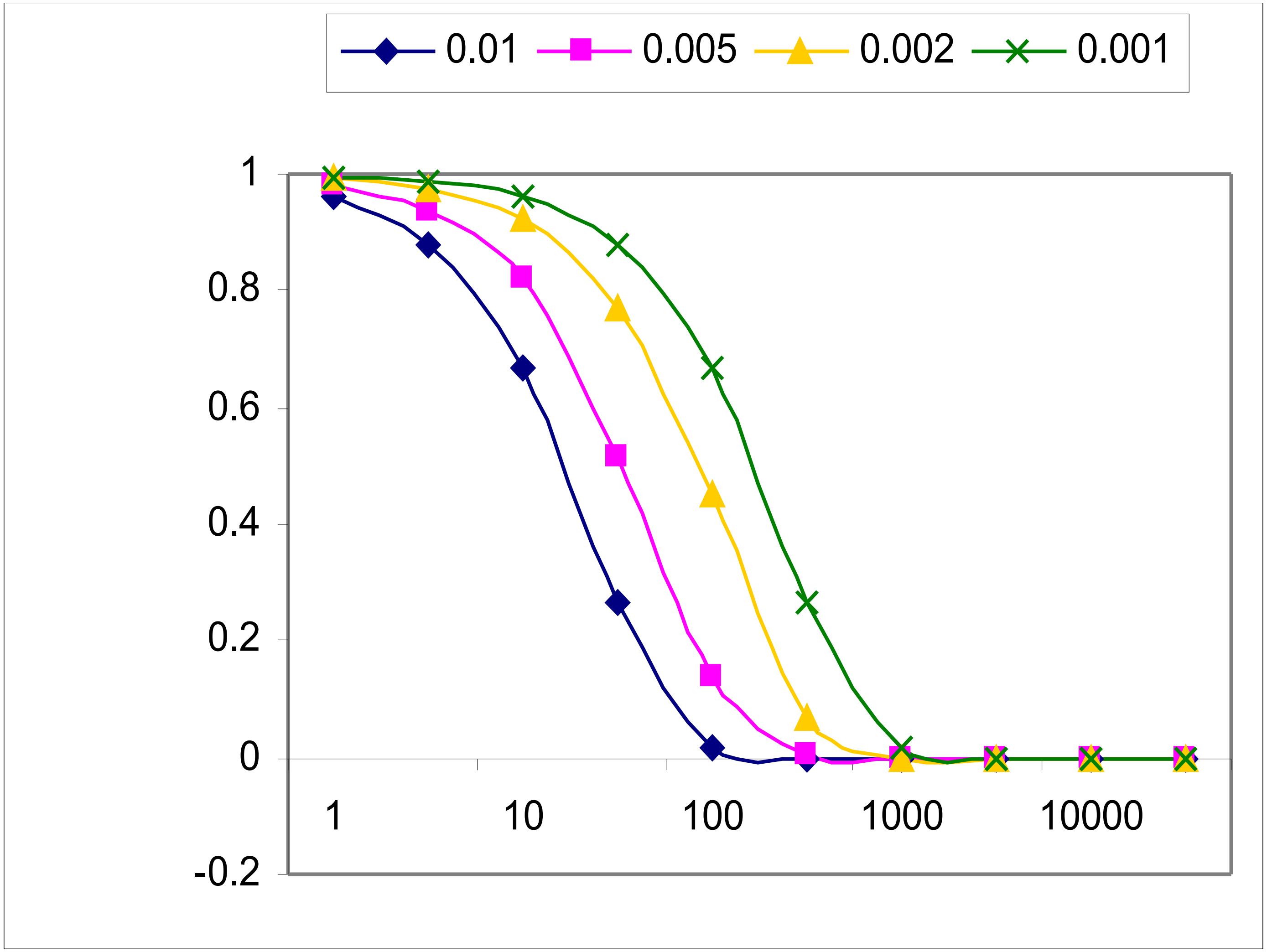
Molecular Phylogeneticists: Our methods (such as the Gamma distribution) assume sites are in the SAME rate class across the entire tree (AND, we only need one parameter- so there).

ALLAN WILSON CENTRE

simulation results with standard model

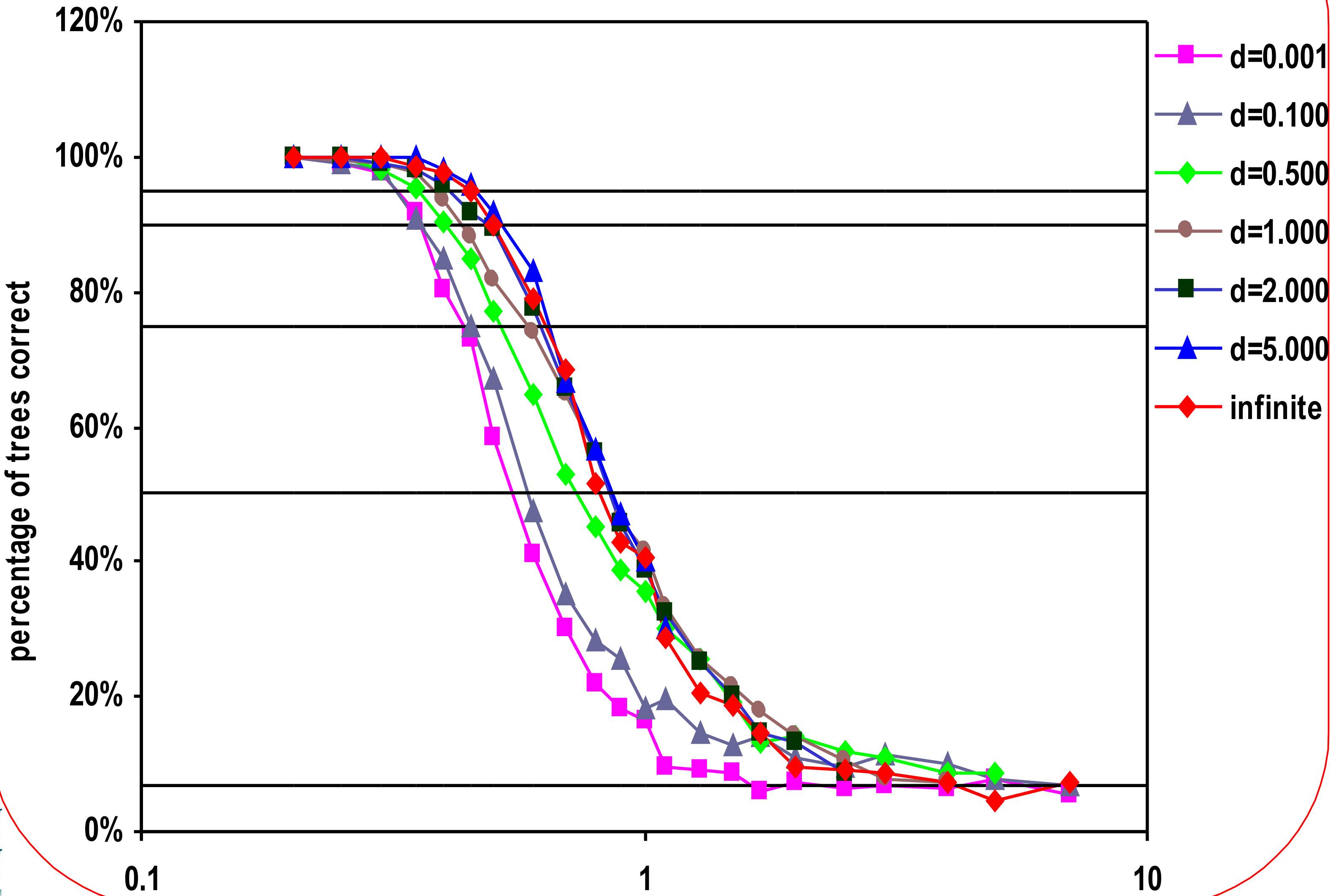


Calculated results, $\Delta \leq ¼ + ne^{-qt}$



ALLAN WILSON CENTRE

simulation results with covarion model



ALLAN
WILSON
CENTRE

do ‘rates’ exist !!!

We go ON

and ON

and ON

and ON

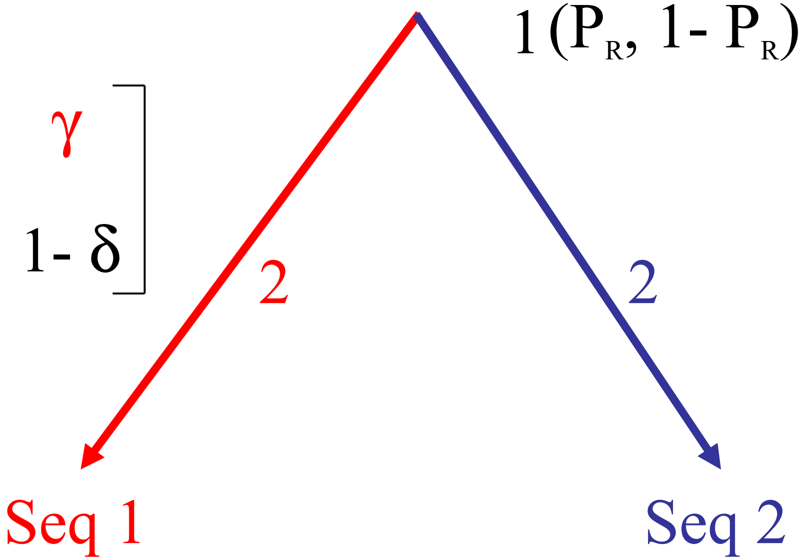
About ‘molecular clocks’.

Should we??

ALLAN WILSON CENTRE

not enough information to recover the full model

$$\begin{bmatrix} 1-\gamma & \gamma \\ \delta & 1-\delta \end{bmatrix}$$



$1\ (P_R,\ 1-P_R)$ *composition at root*

5 required,

3 available

two taxa, two codes

Seq 1 --

Seq 2 --

$$\text{Seq 1}\quad \begin{array}{c} R \\ Y \end{array}\begin{bmatrix} \alpha & \beta \\ \gamma & * \end{bmatrix}$$

$$\begin{array}{cc} R & Y \end{array}$$

Seq 2

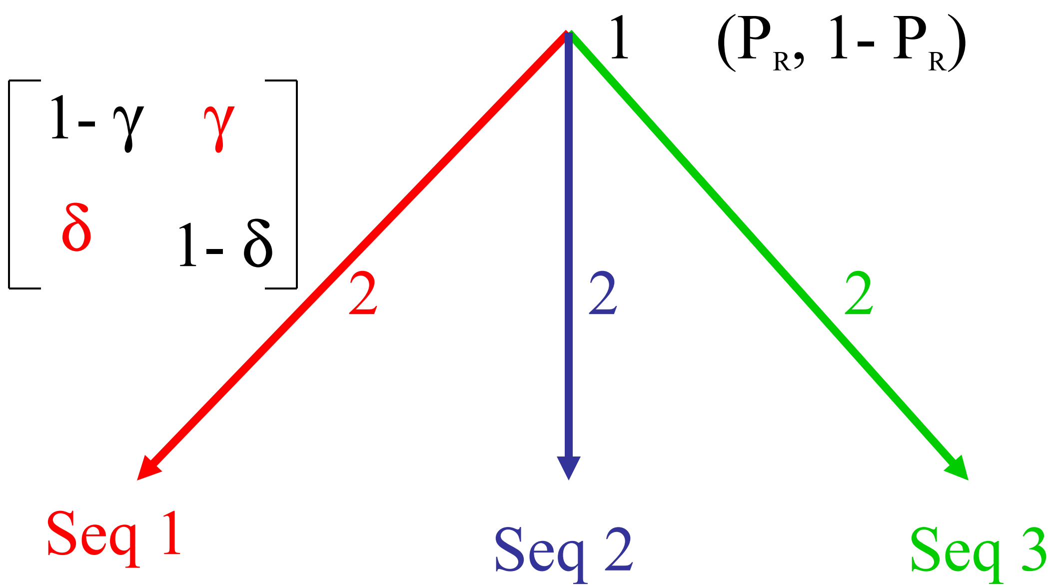
Divergence matrix, $F_{i,j}$

1	2	
R	R	α
R	Y	β
Y	R	γ
Y	Y	*

Three independent parameters estimated

ALLAN WILSON CENTRE

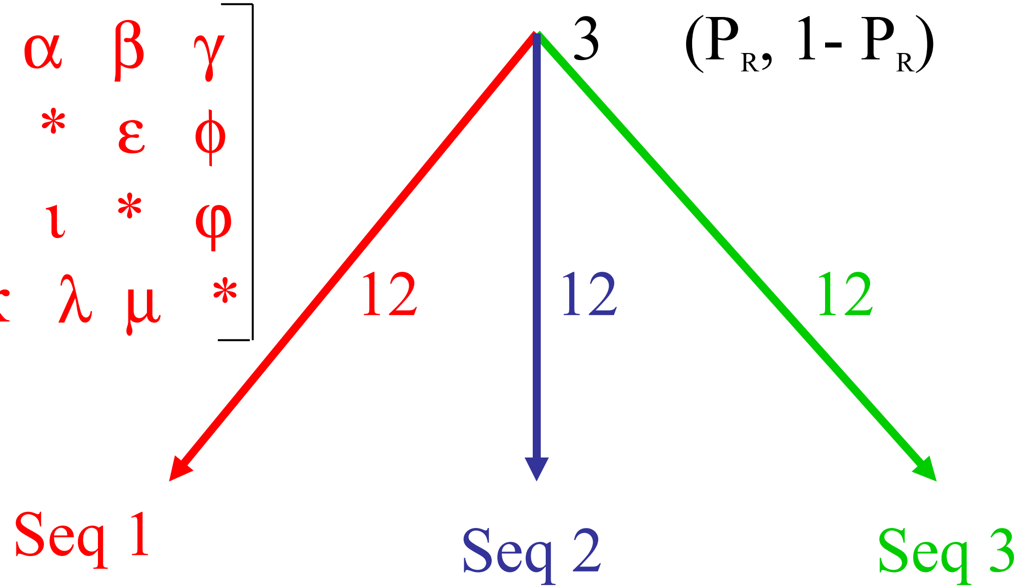
three taxa



7 required

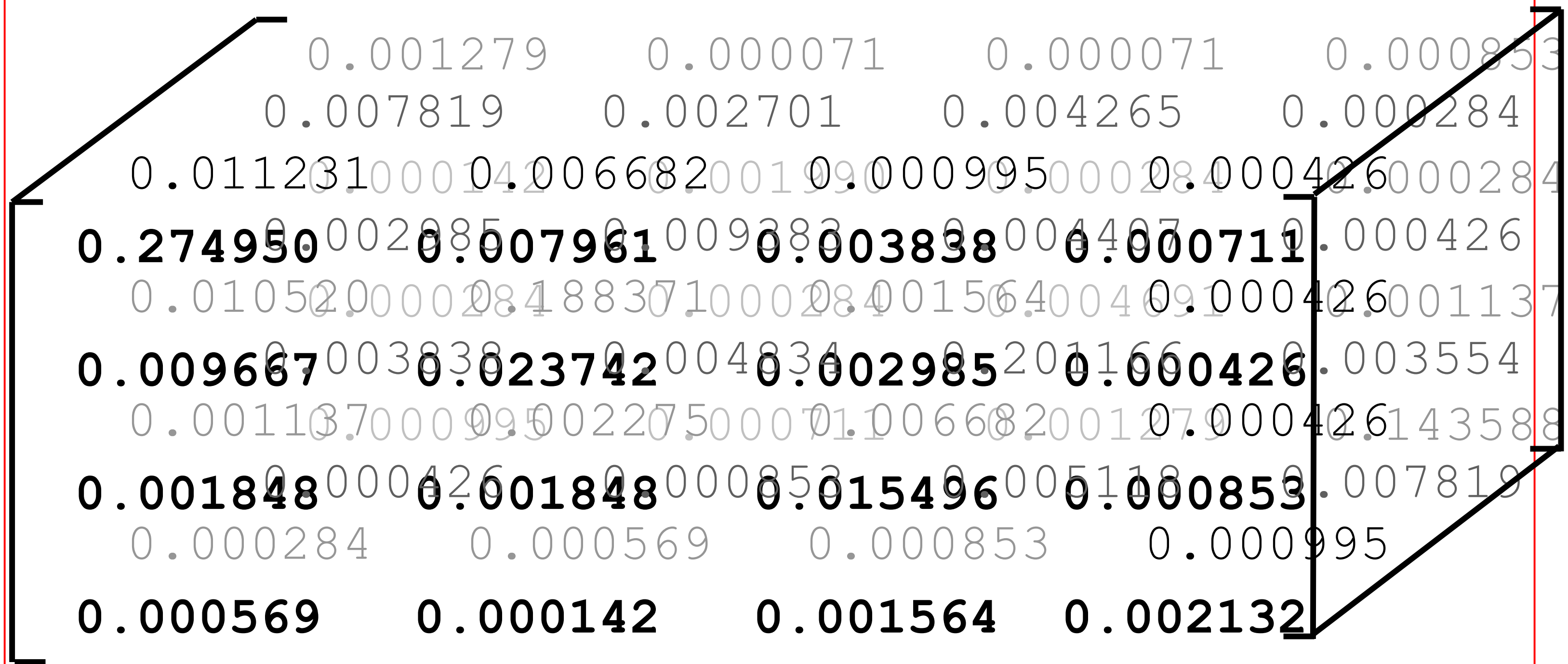
four character states

$$\begin{bmatrix} * & \alpha & \beta & \gamma \\ \delta & * & \varepsilon & \phi \\ \eta & \iota & * & \varphi \\ \kappa & \lambda & \mu & * \end{bmatrix}$$



39 required

tensor, 3D matrix



64 – 1 = 63 values, but a sparse matrix!

primary diagonal

Gymnure, Mole and Shrew

T	**T**	**0.274950**	0.007961	0.003838	0.000711
T	C	0.009667	0.023742	0.002985	0.000426
T	A	0.001848	0.001848	0.015496	0.000853
T	G	0.000569	0.000142	0.001564	0.002132
C	T	0.011231	0.006682	0.000995	0.000426
C	**C**	0.010520	**0.188371**	0.001564	0.000426
C	A	0.001137	0.002275	0.006682	0.000426
C	G	0.000284	0.000569	0.000853	0.000995
A	T	0.007819	0.002701	0.004265	0.000284
A	C	0.002985	0.009383	0.004407	0.000426
A	**A**	0.003838	0.004834	**0.201166**	0.003554
A	G	0.000426	0.000853	0.005118	0.007819
G	T	0.001279	0.000071	0.000071	0.000853
G	C	0.000142	0.001990	0.000284	0.000284
G	A	0.000284	0.000284	0.004691	0.001137
G	**G**	0.000995	0.000711	0.001279	**0.143588**
		T	**C**	**A**	**G**

ALLAN WILSON CENTRE

secondary diagonals

		Gymnure (moon rat)	Mole,	Shrew	
T	T	**0.274950**	0.007961	0.003838	0.000711
T	C	0.009667	**0.023742**	0.002985	0.000426
T	A	0.001848	0.001848	**0.015496**	0.000853
T	G	0.000569	0.000142	0.001564	**0.002132**
C	T	**0.011231**	0.006682	0.000995	0.000426
C	C	0.010520	**0.188371**	0.001564	0.000426
C	A	0.001137	0.002275	**0.006682**	0.000426
C	G	0.000284	0.000569	0.000853	**0.000995**
A	T	**0.007819**	0.002701	0.004265	0.000284
A	C	0.002985	**0.009383**	0.004407	0.000426
A	A	0.003838	0.004834	**0.201166**	0.003554
A	G	0.000426	0.000853	0.005118	**0.007819**
G	T	**0.001279**	0.000071	0.000071	0.000853
G	C	0.000142	**0.001990**	0.000284	0.000284
G	A	0.000284	0.000284	**0.004691**	0.001137
G	G	0.000995	0.000711	0.001279	**0.143588**
		T	**C**	**A**	**G**

ALLAN WILSON CENTRE

moon rat, 1+2

T	0.955	0.148	0.087	0.028
C	0.025	0.803	0.025	0.009
A	0.018	0.043	0.876	0.076
G	0.002	0.006	0.012	0.887
	T	C	A	G

T	.955 ±.004	.150 ±.013	.087 ±.009	.029 ±.008
C	.025 ±.003	.800 ±.014	.025 ±.005	.009 ±.003
A	.018 ±.003	.044 ±.006	.877 ±.011	.077 ±.011
G	.002 ±.001	.006 ±.002	.012 ±.002	.886 ±.015
	T	C	A	G

therefore we believe in symmetric models

ALLAN WILSON CENTRE

mole, shrew and moon rat

mole

	T	C	A	G
T	0.976	**0.062**	0.021	0.013
C	0.017	0.931	0.020	0.007
A	**0.006**	0.006	0.948	**0.012**
G	0.001	0.001	0.010	0.968

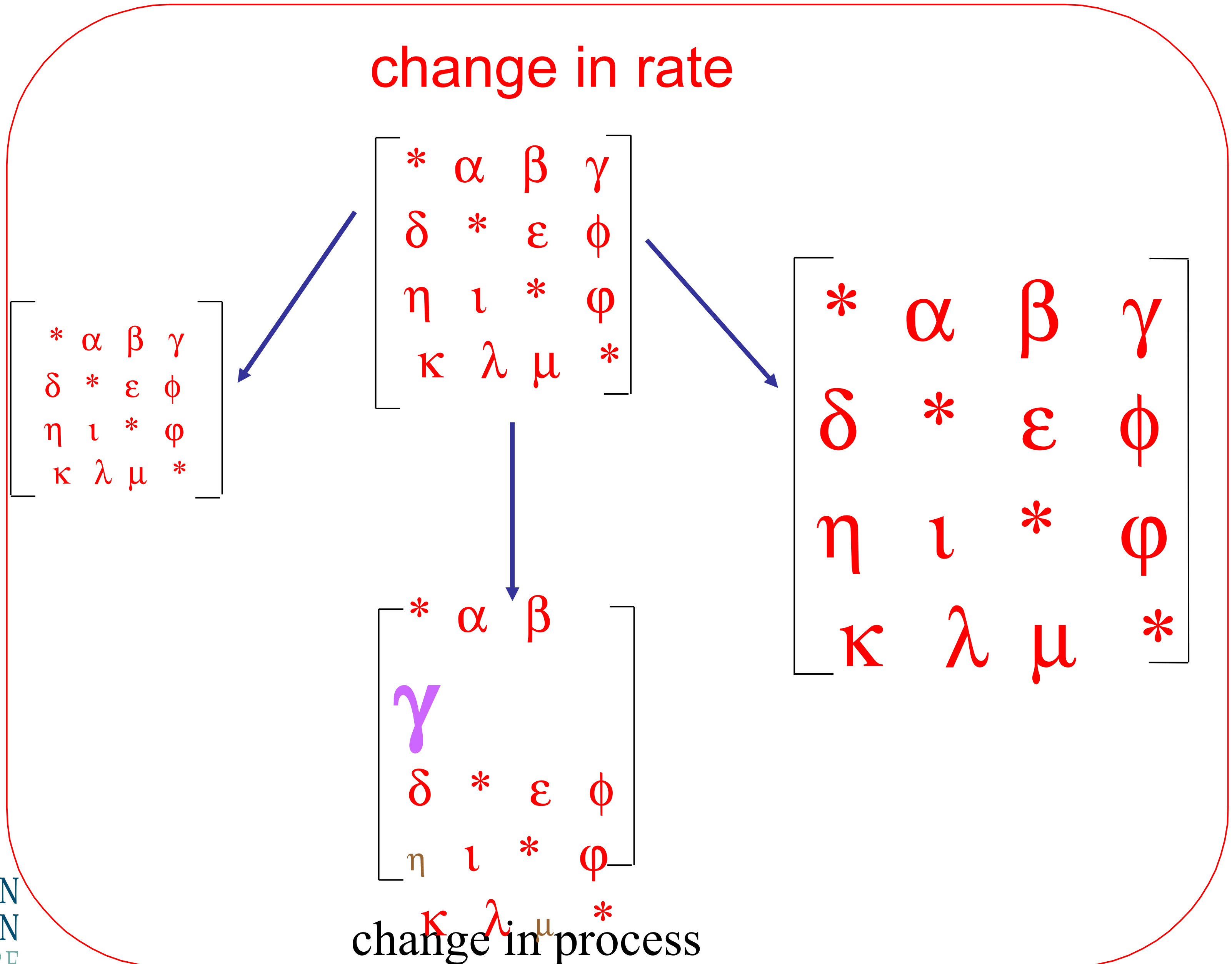
shrew

	T	C	A	G
T	0.977	**0.038**	0.024	0.011
C	0.020	0.951	0.020	0.003
A	**0.002**	0.009	0.942	**0.011**
G	0.001	0.001	0.015	0.976

moon rat

	T	C	A	G
T	0.955	**0.148**	0.087	0.028
C	0.025	0.803	0.025	0.009
A	**0.018**	0.043	0.876	**0.076**
G	0.002	0.006	0.012	0.887

ALLAN WILSON CENTRE

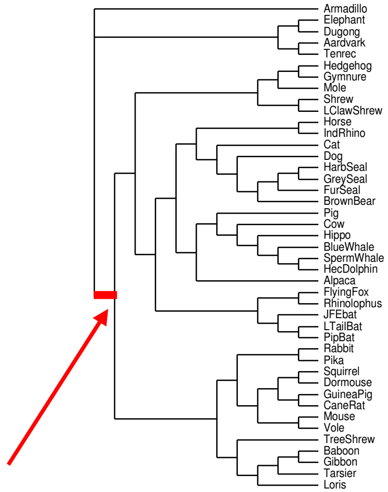

change in rate
* α β γ
δ * ε ϕ
η ι * φ
κ λ μ *
* α β γ
δ * ε ϕ
η ι * φ
κ λ μ *
* α β γ
δ * ε ϕ
η ι * φ
κ λ μ *
* α β
γ
δ * ε ϕ
η ι * φ
κ λ μ *
change in process

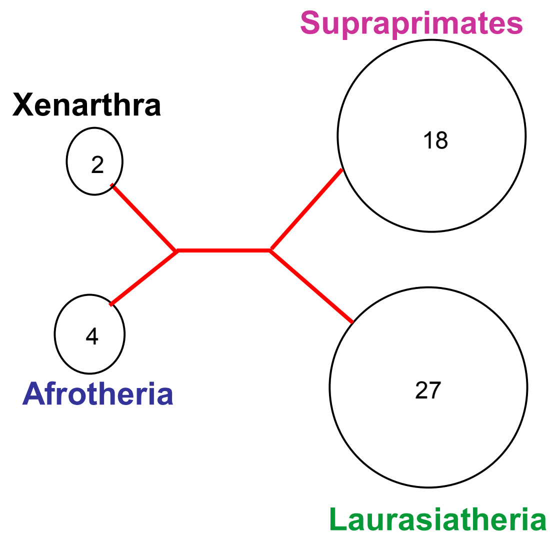
do we know anything?

the curse of ‘flat priors’

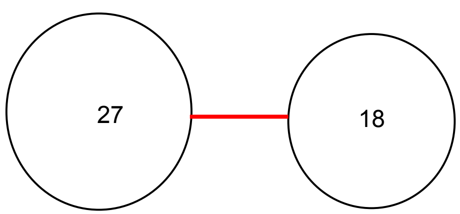
the ‘we know nothing syndrome’

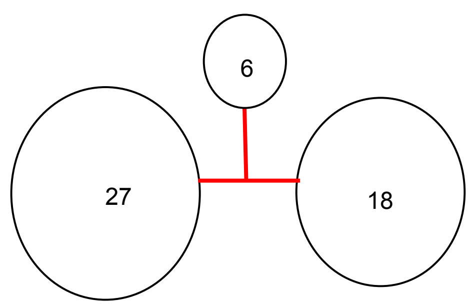
ALLAN WILSON CENTRE





binary trees, b(n) = (2n-5)!!
= 1 x 3 x 5 x 7 … 2n-5.

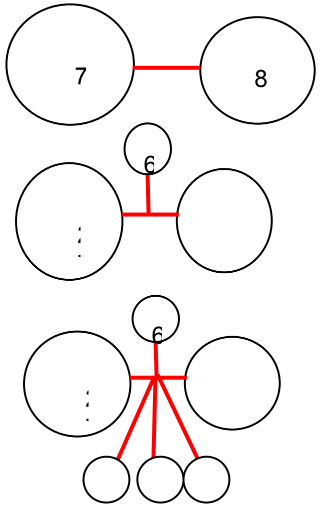




5.68×10^{-18}

binary trees, $b(n) = (2n-5)!!$

$= 1 \times 3 \times 5 \times 7 \ldots 2n-5.$

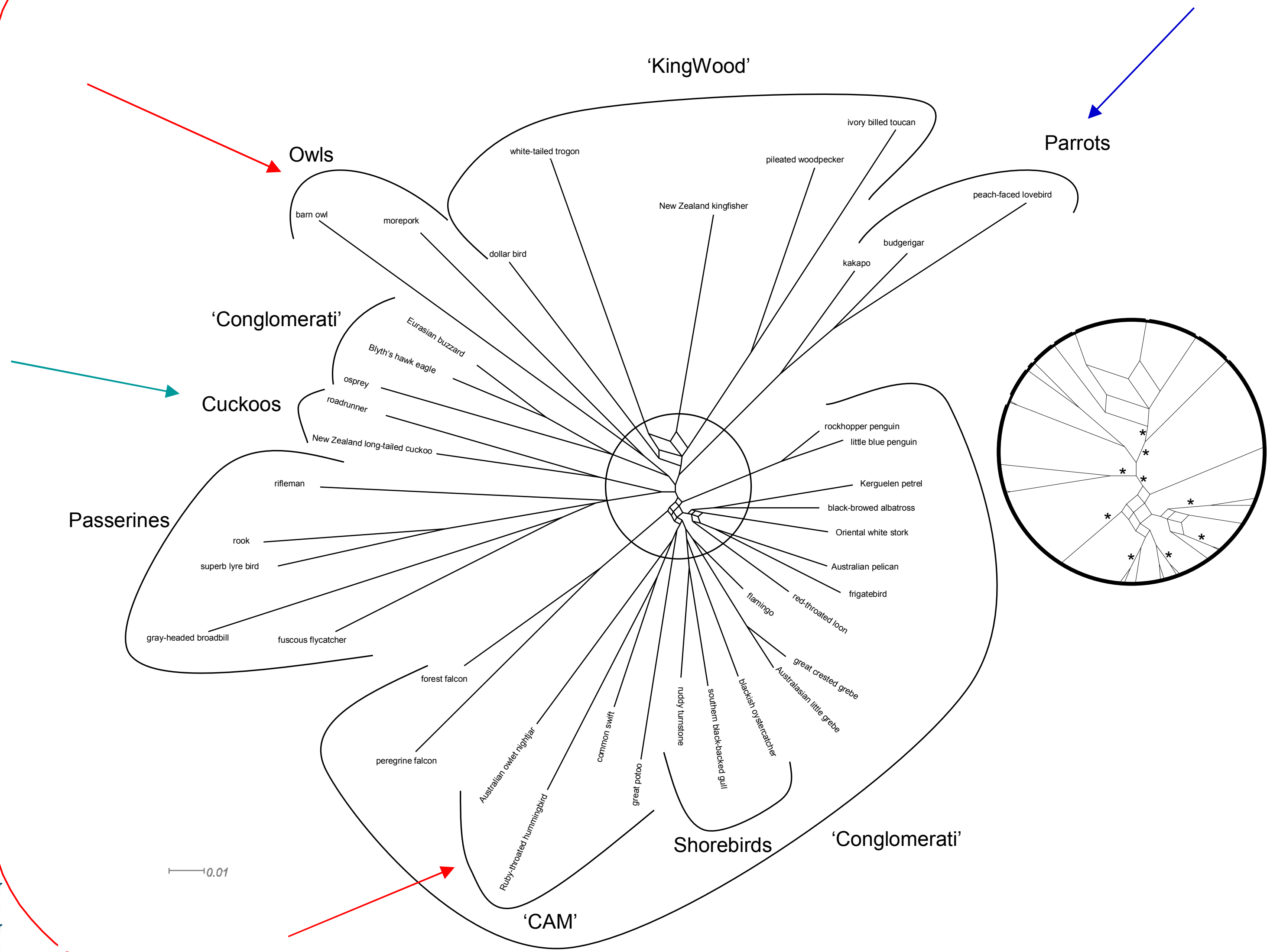


$b(n_1+1).b(n_2+1) / b(n_t)$

$b(n_1+1).b(n_2+1).b(n_3+1) / b(n_t)$

$b(n_1+1).b(n_2+1) \ldots b(n_i+1) / b(n_t)$

40 birds



'KingWood'
Owls
Parrots
'Conglomerati'
Cuckoos
Passerines
Shorebirds
'CAM'
'Conglomerati'
ivory billed toucan
white-tailed trogon
pileated woodpecker
New Zealand kingfisher
peach-faced lovebird
barn owl
morepork
dollar bird
budgerigar
kakapo
Eurasian buzzard
Blyth's hawk eagle
osprey
roadrunner
New Zealand long-tailed cuckoo
rifleman
rook
superb lyre bird
gray-headed broadbill
fuscous flycatcher
rockhopper penguin
little blue penguin
Kerguelen petrel
black-browed albatross
Oriental white stork
Australian pelican
frigatebird
flamingo
red-throated loon
great crested grebe
Australasian little grebe
blackish oystercatcher
southern black-backed gull
ruddy turnstone
great potoo
common swift
Ruby-throated hummingbird
Australian owlet nightjar
peregrine falcon
forest falcon
0.01

$$P(n,k) = \frac{R(k) \times B(n-k+1)}{B(n)}$$

probability with *n* taxa of observing a prespecified clade of size *k*.

with $n = 40$ and

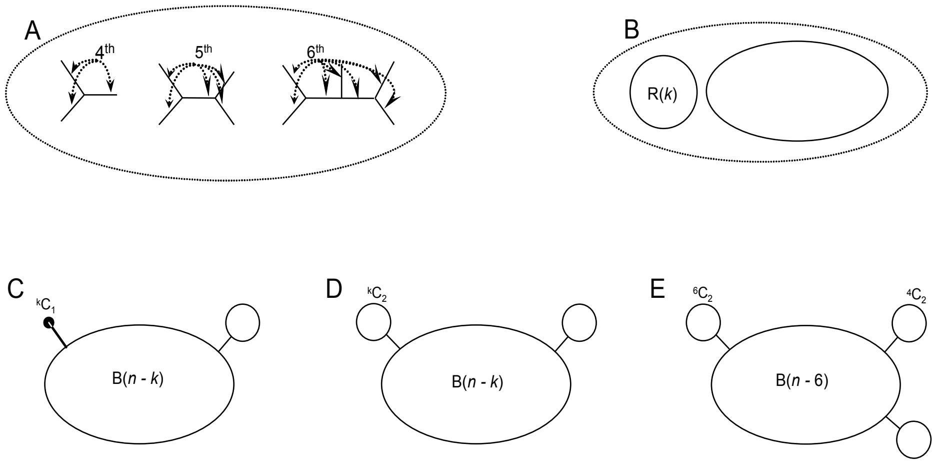
$k = 2$, $P \approx 0.013$ cuckoo,roadrunner

$k = 3$, $P \approx 0.0026$ parrots

$k = 4$, $P \approx 7.12 \times 10^{-6}$,

$k = 5$, $P \approx 5.84 \times 10^{-8}$.

ALLAN WILSON CENTRE



ALLAN WILSON CENTRE

potoo, owlet-nightjar, owl, barn owl, swift, hummingbird (6)

ALLAN WILSON CENTRE

Where next in Phylogeny?

allow realism in phylogeny

set the biological question

we have some bad failures

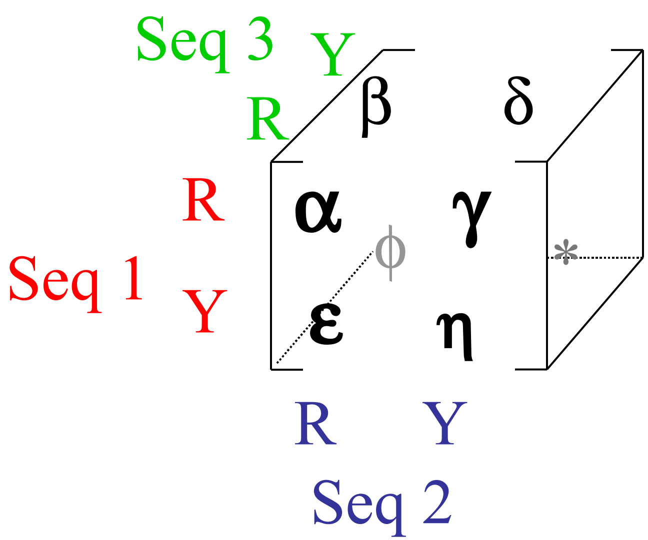
we need a range of alternatives

ALLAN WILSON CENTRE

Belief is the curse of the thinking class

tensor, 2-states

Seq 1 ----------------R--
Seq 2 ----------------R--
Seq 3 ----------------R--



R R R α
R R Y β
R Y R γ
R Y Y δ
Y R R ε
Y R Y ϕ
Y Y R η
Y Y Y *
1 2 3

7 available !