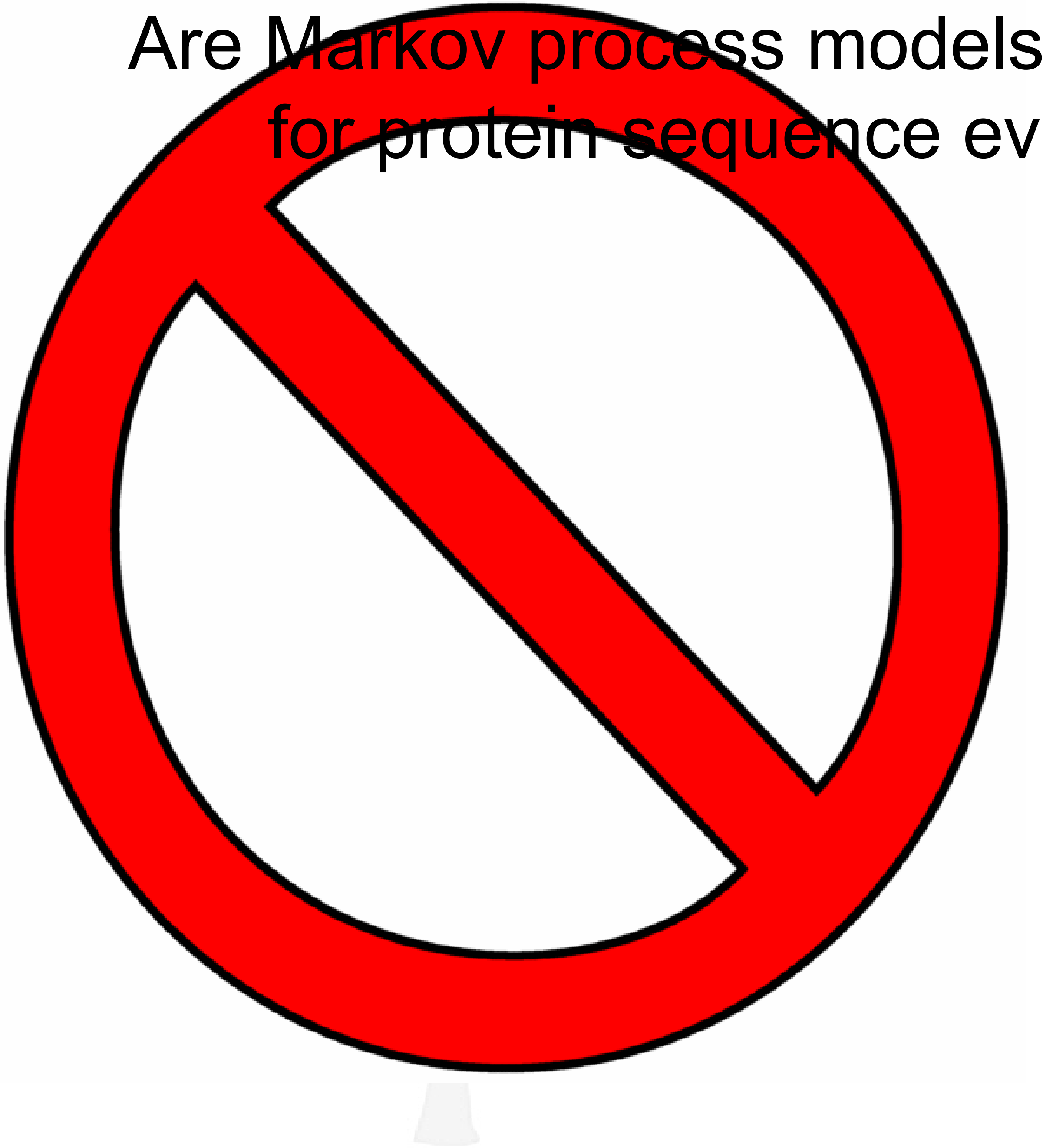
Is protein sequence evolution constant over time?

Carolin Kosiol &
Nick Goldman

goldman@ebi.ac.uk
http:/www.ebi.ac.uk/goldman

MIEP08
12 June 2008

Are Markov process models appropriate for protein sequence evolution?



NO

?

Evidence of non-Markov evolution of amino acid sequences

Amino Acid Substitution Matrices From Protein Blocks

S. Henikoff and J.G. Henikoff
Proceedings of the National Academy of Sciences
of the United States of America 89:10915–10919. 1992

Tree-based Maximal Likelihood Substitution Matrices and Hidden Markov Models

G. Mitchison and R. Durbin
Journal of Molecular Evolution 41:1139–1151. 1995

Amino Acid Substitution During Functionally Constrained Divergent Evolution of Protein Sequences

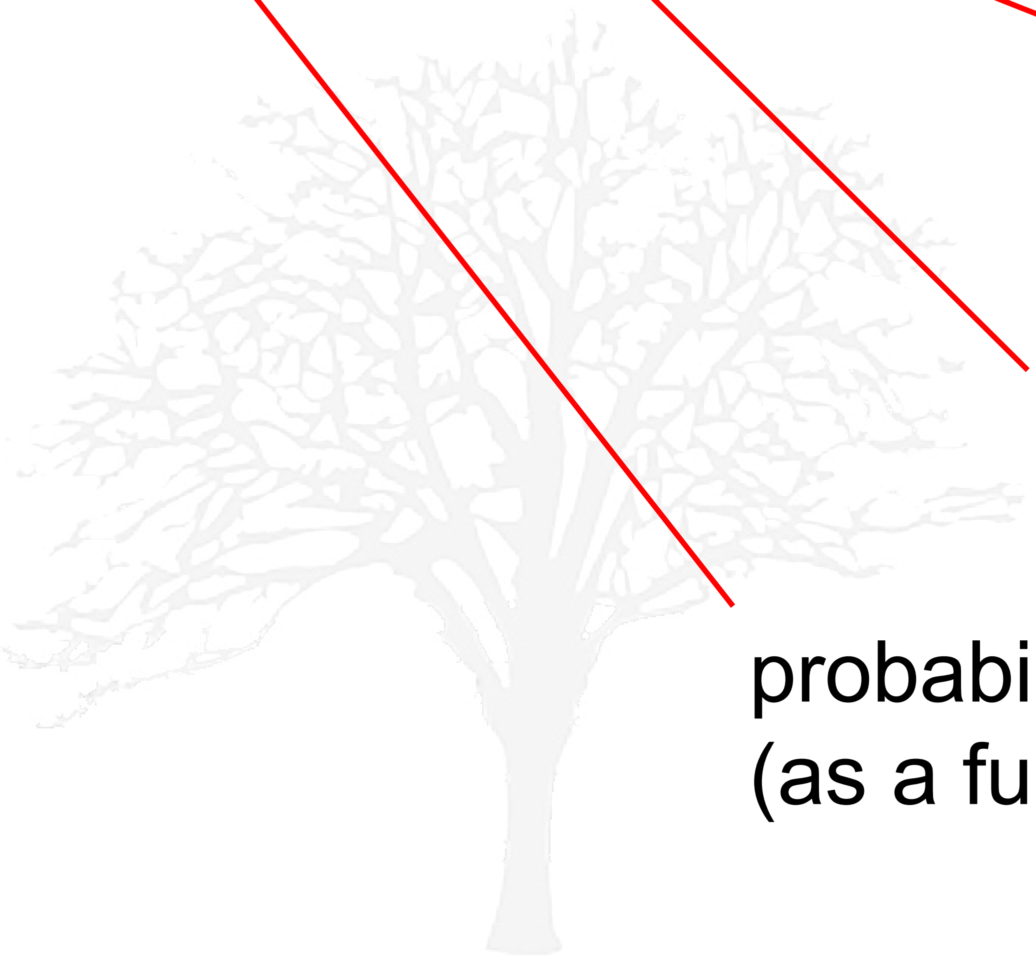
S.A. Benner, M.A. Cohen and G.H. Gonnet
Protein Engineering 7:1323–1332. 1994

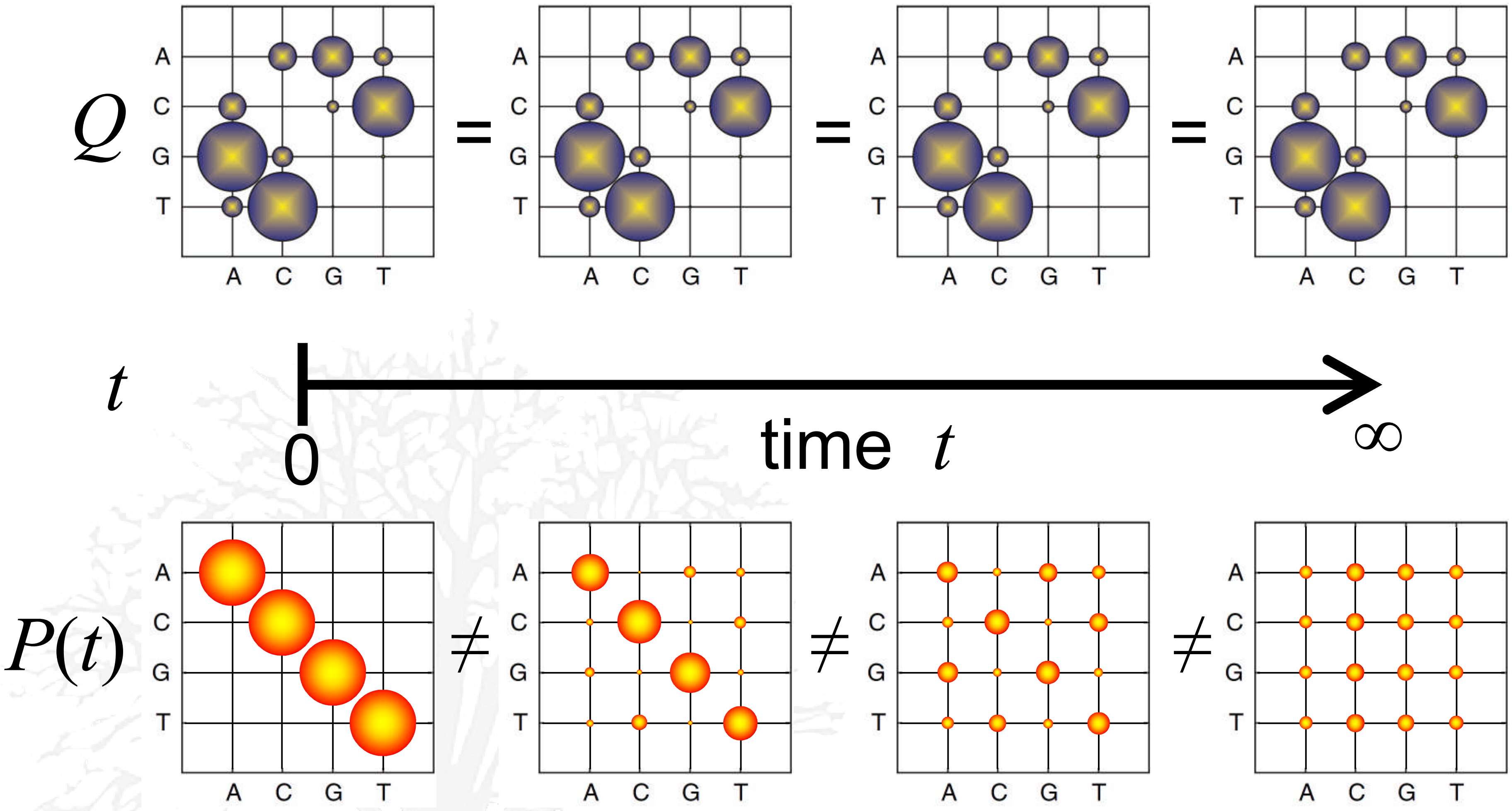
$$P(t) = \exp(tQ) = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \cdots$$

instantaneous rate matrix

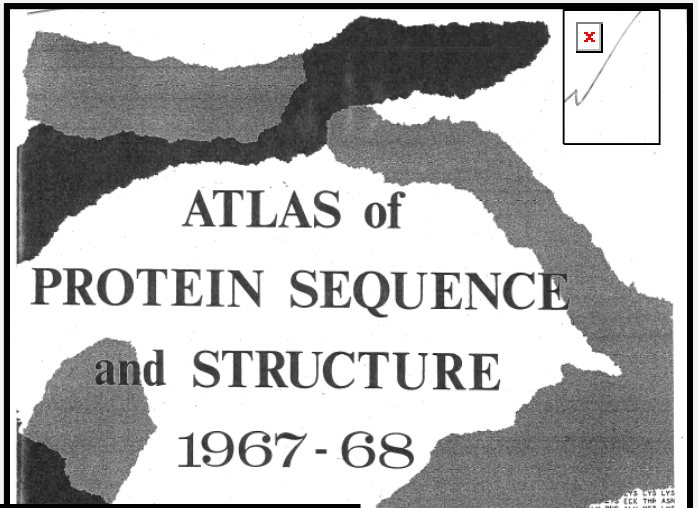
time

probability of change (as a function of time)



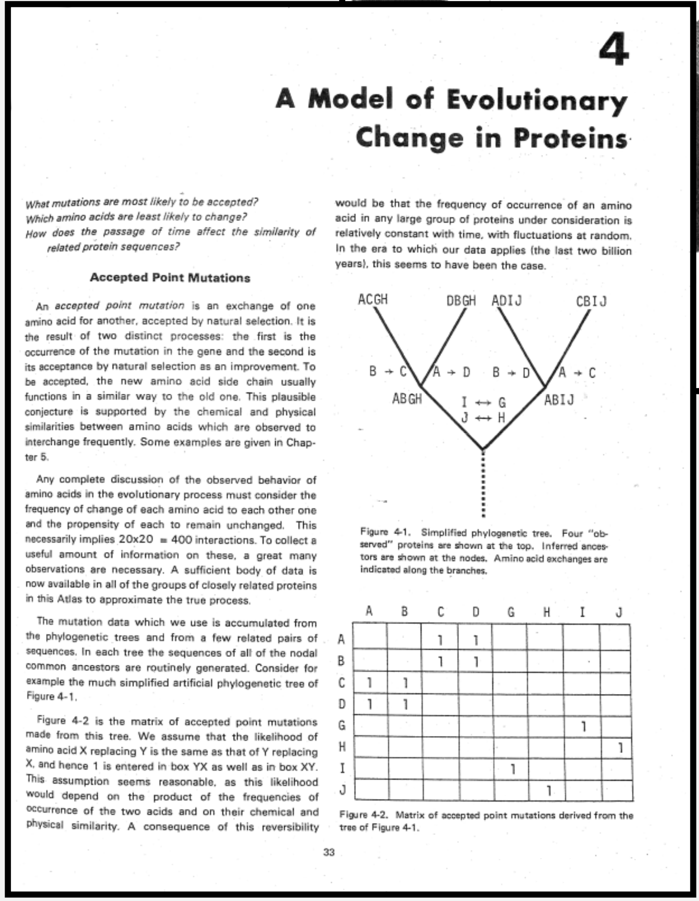

Q
=
=
=
A C G T
t
0
time t
∞
P(t)
≠
≠
≠

It is possible
to infer *P*(*t*)
from sequence data…



ATLAS of
PROTEIN SEQUENCE
and STRUCTURE
1967-68

Margaret O. Dayhoff
Richard V. Eck



4

A Model of Evolutionary Change in Proteins

What mutations are most likely to be accepted?
Which amino acids are least likely to change?
How does the passage of time affect the similarity of related protein sequences?

Accepted Point Mutations

An *accepted point mutation* is an exchange of one amino acid for another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of the mutation in the gene and the second is its acceptance by natural selection as an improvement. To be accepted, the new amino acid side chain usually functions in a similar way to the old one. This plausible conjecture is supported by the chemical and physical similarities between amino acids which are observed to interchange frequently. Some examples are given in Chapter 5.

Any complete discussion of the observed behavior of amino acids in the evolutionary process must consider the frequency of change of each amino acid to each other one and the propensity of each to remain unchanged. This necessarily implies 20x20 = 400 interactions. To collect a useful amount of information on these, a great many observations are necessary. A sufficient body of data is now available in all of the groups of closely related proteins in this Atlas to approximate the true process.

The mutation data which we use is accumulated from the phylogenetic trees and from a few related pairs of sequences. In each tree the sequences of all of the nodal common ancestors are routinely generated. Consider for example the much simplified artificial phylogenetic tree of Figure 4-1.

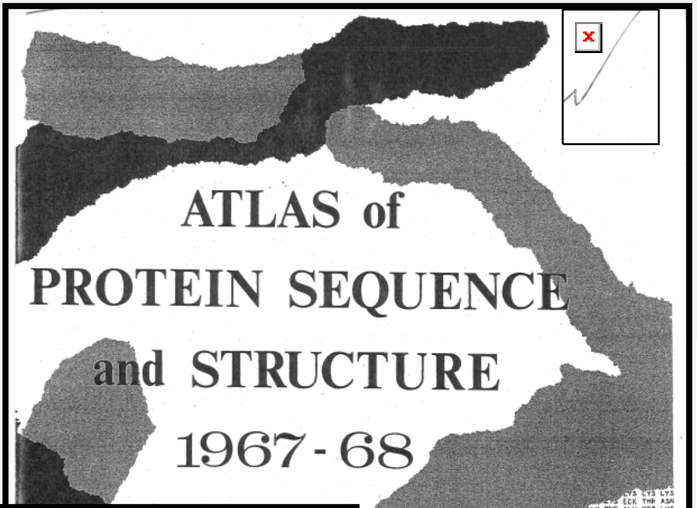
Figure 4-2 is the matrix of accepted point mutations made from this tree. We assume that the likelihood of amino acid X replacing Y is the same as that of Y replacing X, and hence 1 is entered in box YX as well as in box XY. This assumption seems reasonable, as this likelihood would depend on the product of the frequencies of occurrence of the two acids and on their chemical and physical similarity. A consequence of this reversibility would be that the frequency of occurrence of an amino acid in any large group of proteins under consideration is relatively constant with time, with fluctuations at random. In the era to which our data applies (the last two billion years), this seems to have been the case.

Figure 4-1. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 4-2. Matrix of accepted point mutations derived from the tree of Figure 4-1.

…and possible to infer *Q* from *P*(*t*)



ATLAS of PROTEIN SEQUENCE and STRUCTURE 1967-68

Margaret O. Dayhoff
Richard V. Eck

4

A Model of Evolutionary Change in Proteins

What mutations are most likely to be accepted?
Which amino acids are least likely to change?
How does the passage of time affect the similarity of related protein sequences?

Accepted Point Mutations

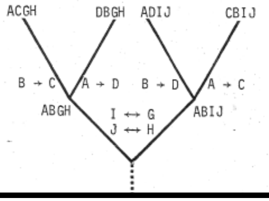
An *accepted point mutation* is an exchange of one amino acid for another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of the mutation in the gene and the second is its acceptance by natural selection as an improvement. To be accepted, the new amino acid side chain usually functions in a similar way to the old one. This plausible conjecture is supported by the chemical and physical similarities between amino acids which are observed to interchange frequently. Some examples are given in Chapter 5.

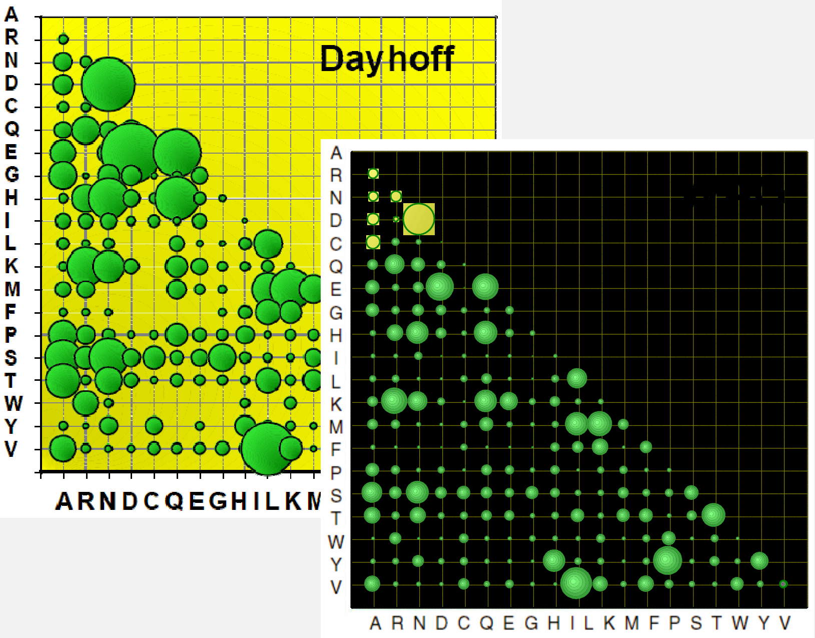
Any complete disc
amino acids in the ev
frequency of change
and the propensity
necessarily implies 2
useful amount of in
observations are nec
now available in all of
in this Atlas to appro

The mutation data
the phylogenetic tre
sequences. In each t
common ancestors
example the much si
Figure 4-1.

Figure 4-2 is the
made from this tree
amino acid X replacin
X, and hence 1 is en
This assumption se
would depend on
occurrence of the tv
physical similarity.

would be that the frequency of occurrence of an amino acid in any large group of proteins under consideration is relatively constant with time, with fluctuations at random. In the era to which our data applies (the last two billion years), this seems to have been the case.



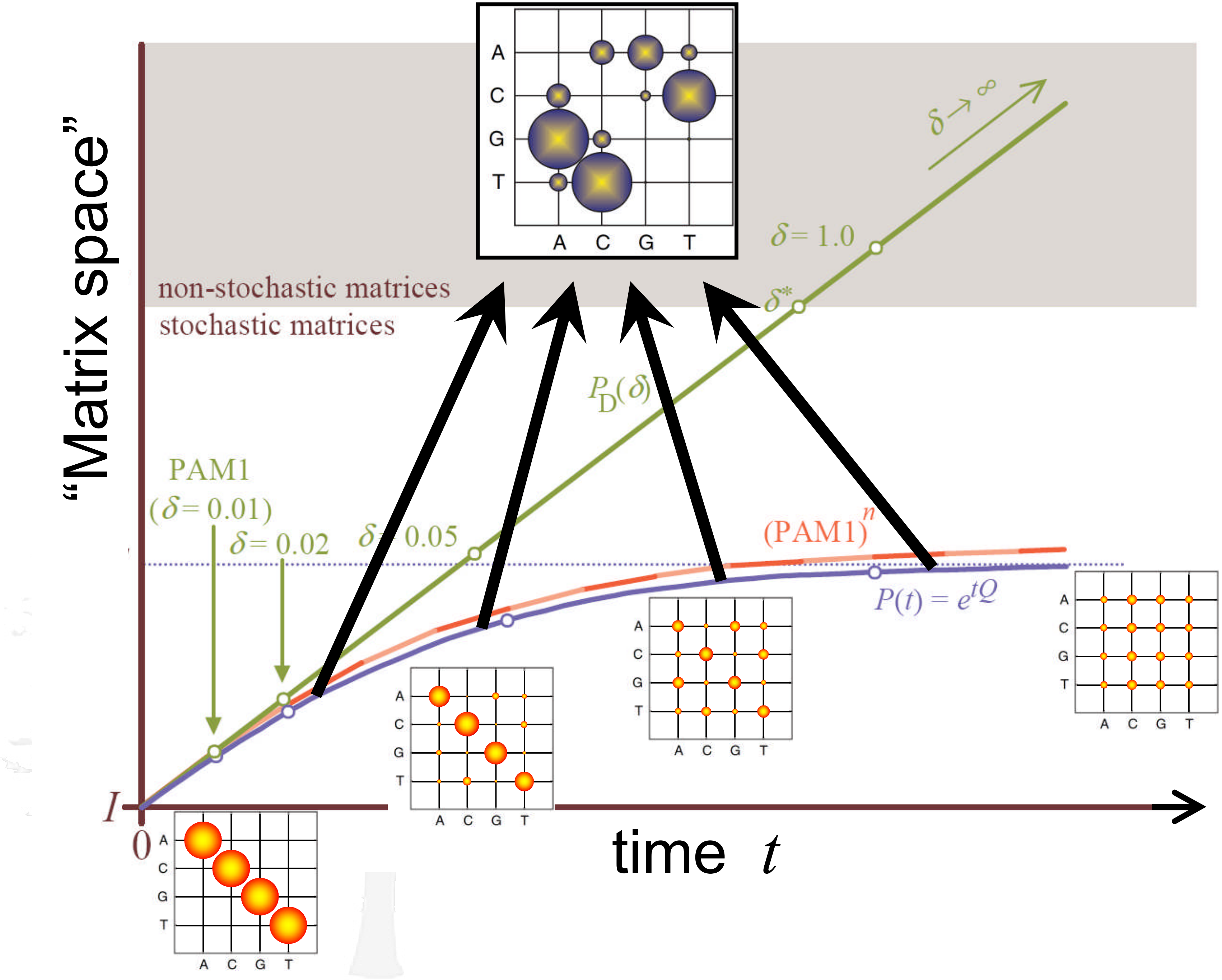


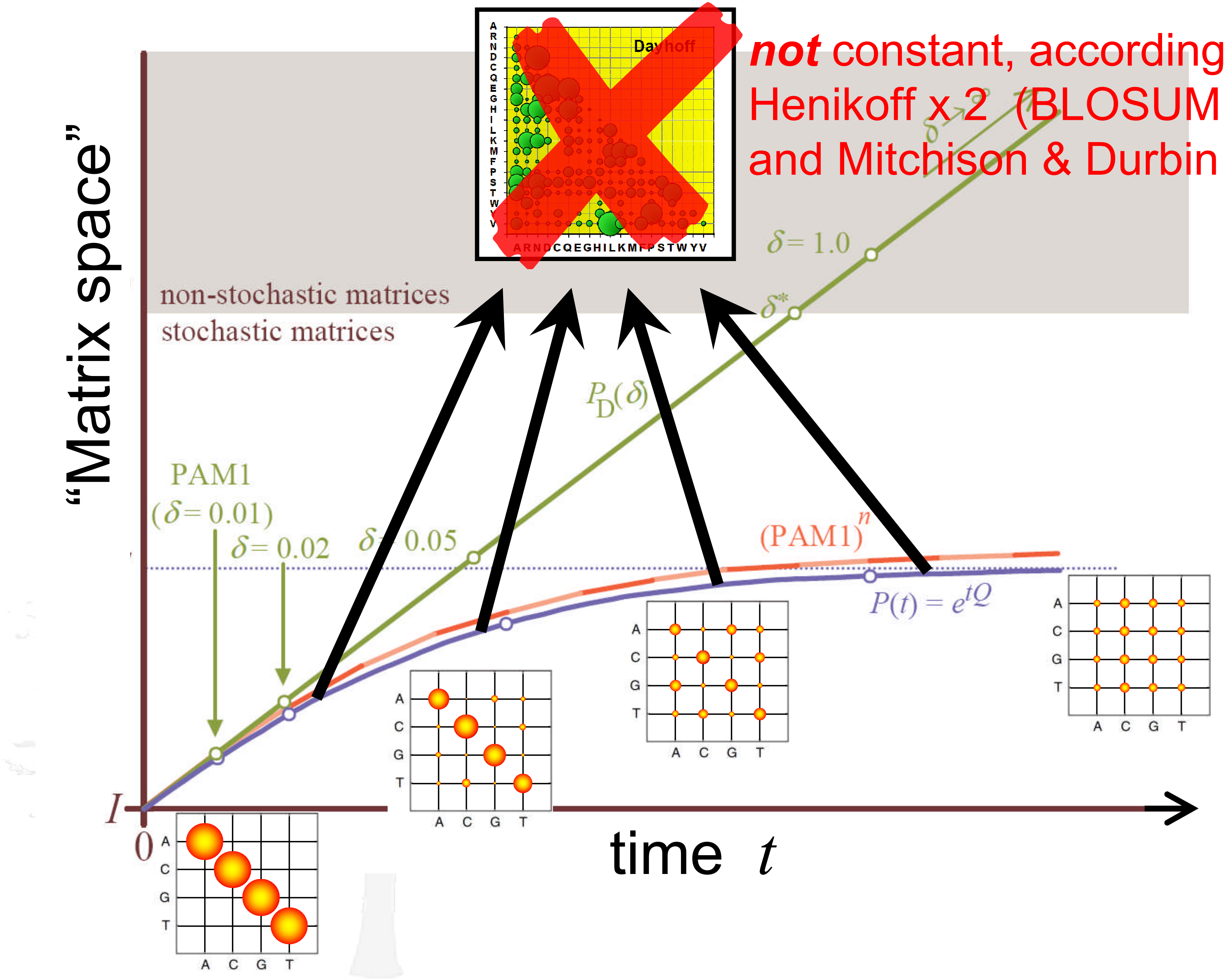
Different Versions of the Dayhoff Rate Matrix

Carolin Kosiol and Nick Goldman

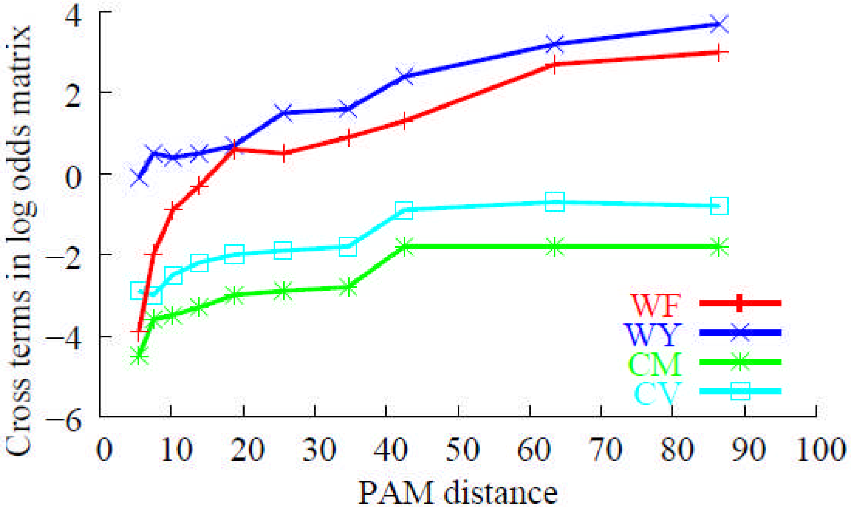
Mol. Biol. Evol. 22(2):193–199. 2005

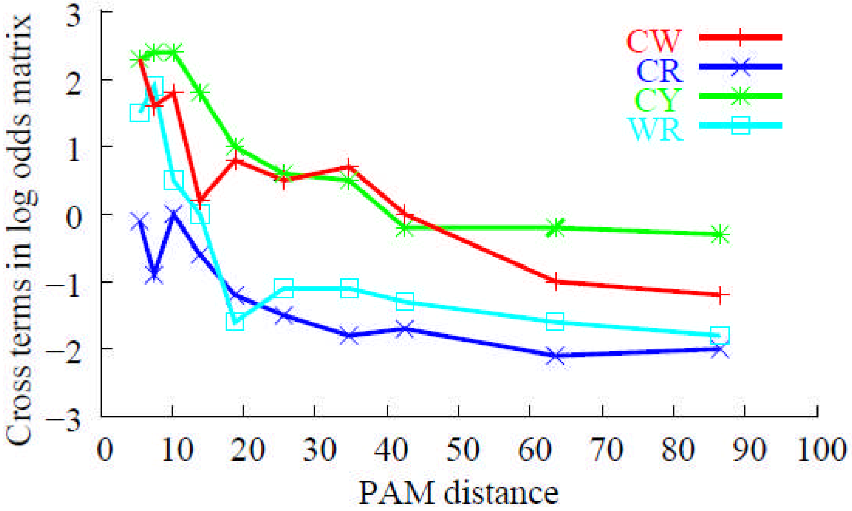
Many phylogenetic inference methods are based on Markov models of sequence evolution. These are usually expressed in terms of a matrix (Q) of instantaneous rates of change but some models of amino acid replacement, most notably the PAM model of Dayhoff and colleagues, were originally published only in terms of time-dependent probability matrices ($P(t)$). Previously published methods for deriving Q have used eigen-decomposition of an approximation to $P(t)$. We


"Matrix space"
A
C
G
T
A C G T
$\delta \to \infty$
$\delta = 1.0$
δ^*
non-stochastic matrices
stochastic matrices
$P_D(\delta)$
PAM1
($\delta = 0.01$)
$\delta = 0.02$
$\delta = 0.05$
$(\text{PAM1})^n$
$P(t) = e^{tQ}$
I
0
time t



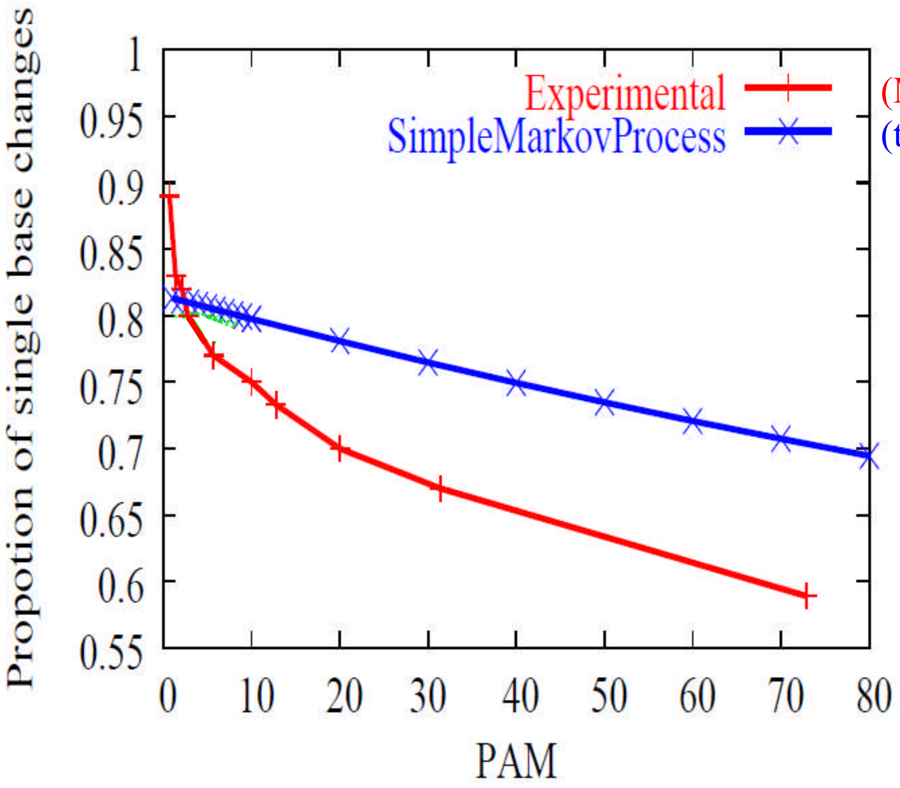
not **constant, according to Henikoff x 2 (BLOSUM) and Mitchison & Durbin**





Benner *et al.* found rate matrix elements varied with observed divergence

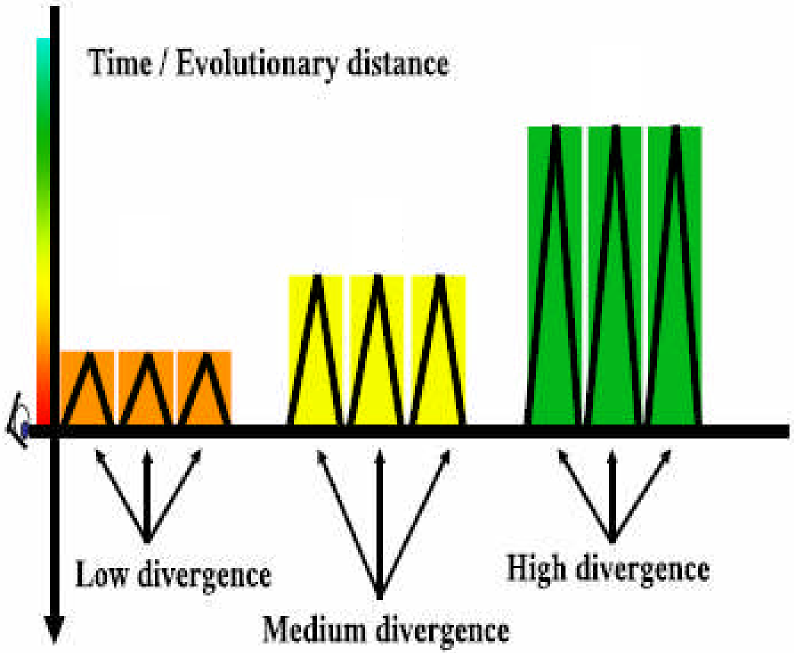
They argued that the genetic code influences the matrix strongly at early stages of divergence, while physicochemical properties are dominant at later stages

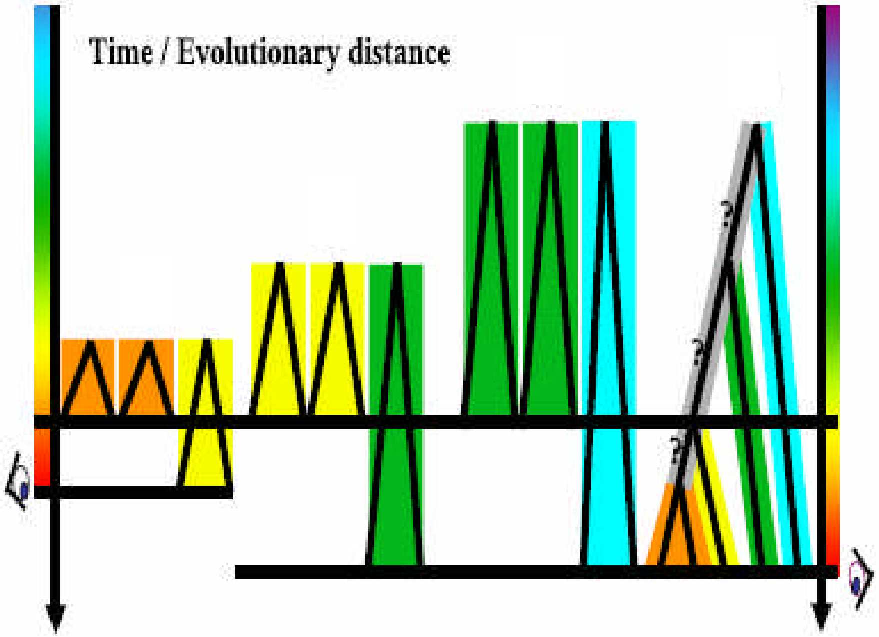


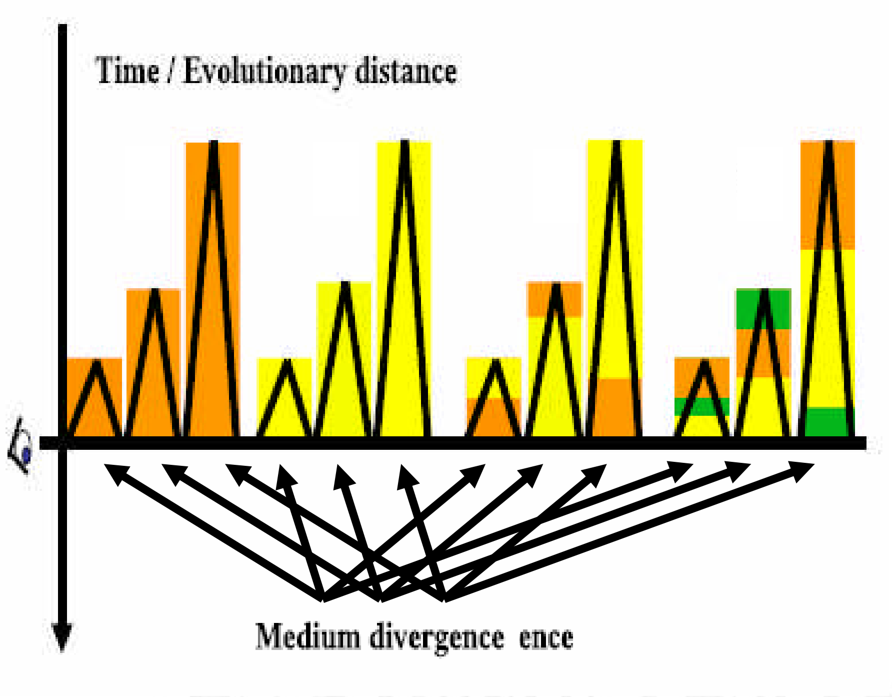
(Mitchison & Durbin)

(this study)

Mitchison & Durbin found the accumulation of amino acid replacements that could be generated by a single nucleotide change was inconsistent with a simple Markov process


Time / Evolutionary distance
Low divergence
Medium divergence
High divergence


Time / Evolutionary distance
?
?
?


Time / Evolutionary distance
Medium divergence ence

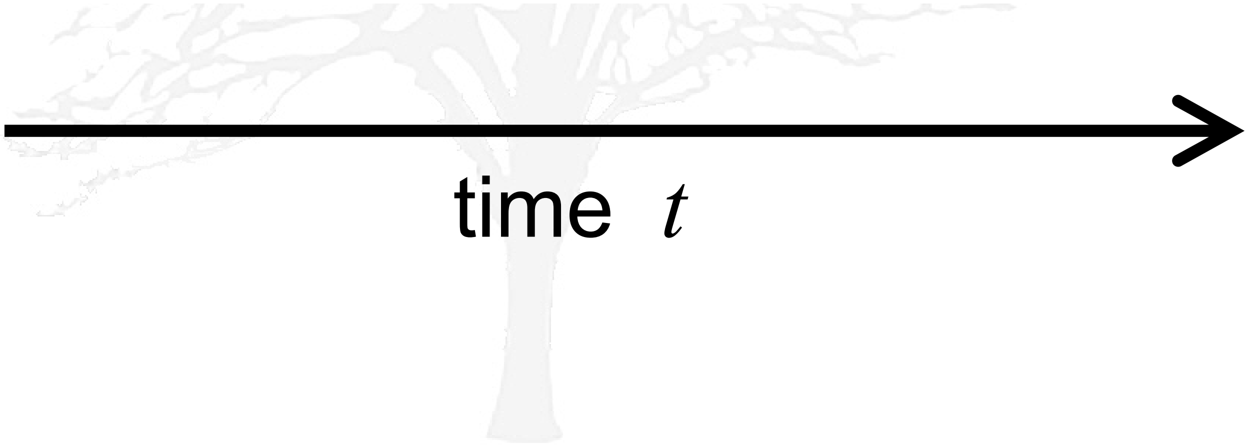
So, how ***will*** we explain the evidence of non-Markov behaviour? — the aggregated Markov process (AMP):

$$\begin{tikzcd}
\cdots \arrow[r] & X(t_k) = CTT \arrow[r] \arrow[d, "f"] & X(t_{k+1}) = CCT \arrow[r] \arrow[d, "f"] & \cdots \\
\cdots & Y(t_k) = L & Y(t_{k+1}) = P & \cdots
\end{tikzcd}$$

Markov process
(codon evolution)

↓ Deterministic
function on states
(genetic code)

Non-Markov process
(protein evolution)



time t

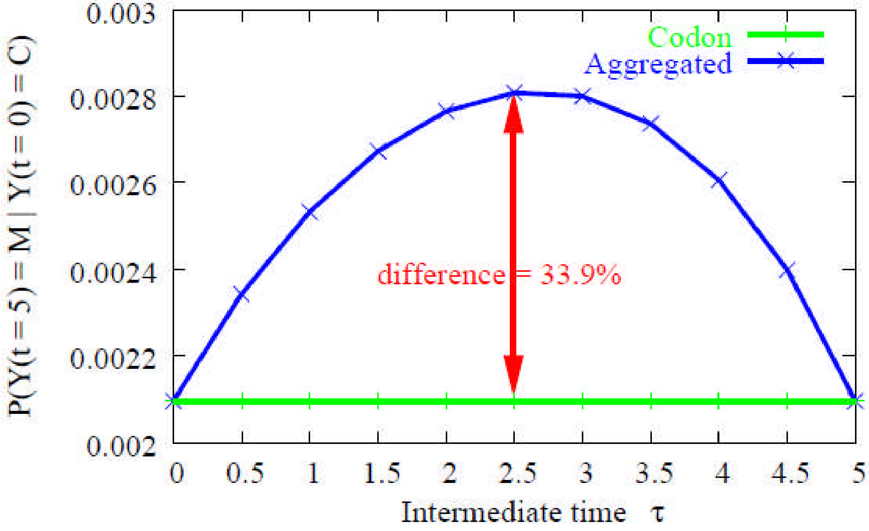
$$q_{ij,i\neq j} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon or requires} > 1 \text{ nucleotide substitution} \\ \pi_j & \text{if } i \longrightarrow j \text{ synonymous transversion} \\ \pi_j \kappa & \text{if } i \longrightarrow j \text{ synonymous transition} \\ \pi_j \omega & \text{if } i \longrightarrow j \text{ nonsynonymous transversion} \\ \pi_j \kappa \omega & \text{if } i \longrightarrow j \text{ nonsynonymous transition} \end{cases}$$

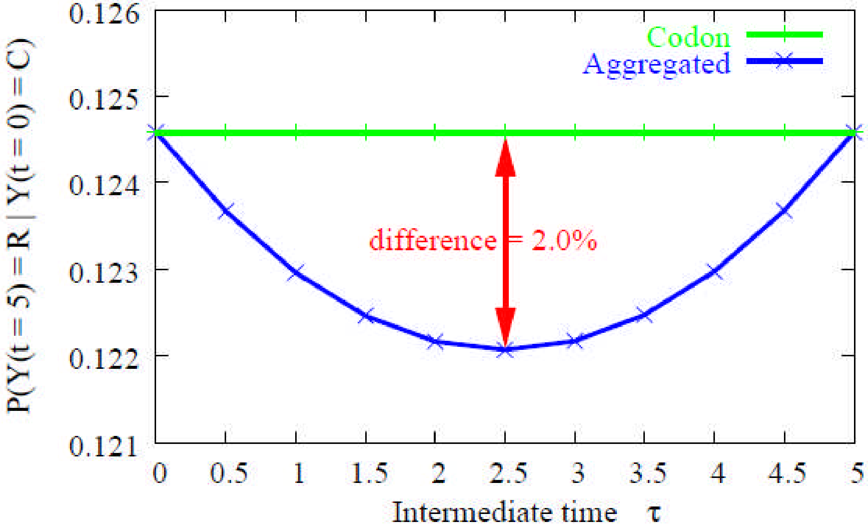
$$\kappa = 2.5 \qquad \omega = 0.2$$

$$\begin{array}{llll} r_1 = 0.00001, & r_2 = 0.0001, & r_3 = 0.0001, & r_4 = 0.001, \\ r_5 = 0.01, & r_6 = 0.1 & r_7 = 0.15, & r_8 = 0.2, \\ r_9 = 0.3, & r_{10} = 0.5, & r_{11} = 2.0, & r_{12} = 8.73889 \end{array}$$

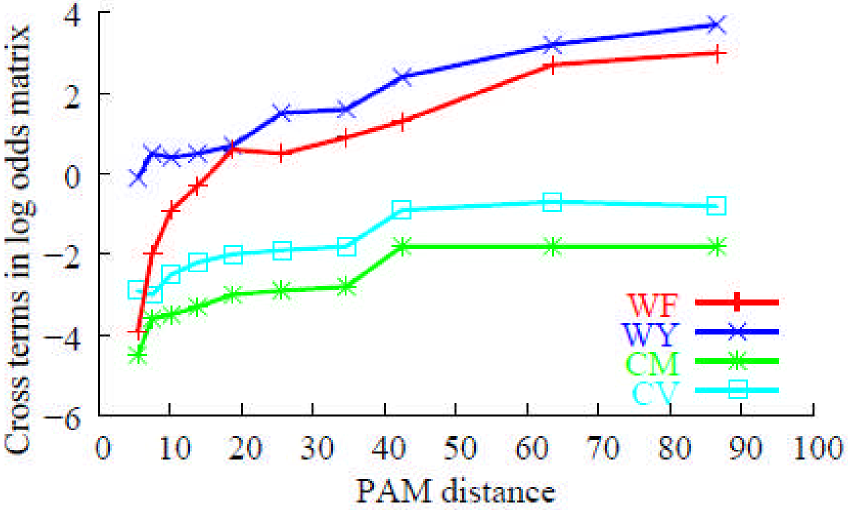
Aggregated Markov processes are not Markov:

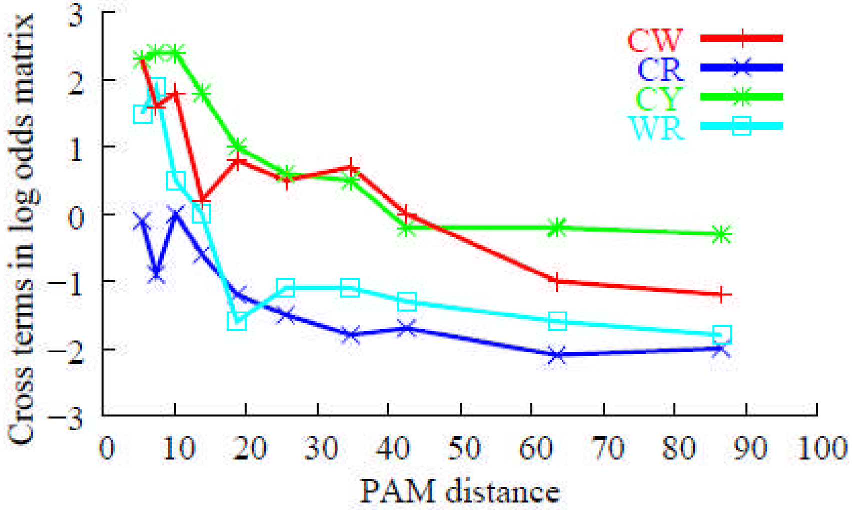
$$P(Y(t_1) = \mathrm{M} | Y(t_0) = \mathrm{C}) = \sum_{i=1}^{20} P(Y(t_1) = \mathrm{M} | Y(\tau) = \mathrm{A}_i) \times P(Y(\tau) = \mathrm{A}_i | Y(t_0) = \mathrm{C})$$



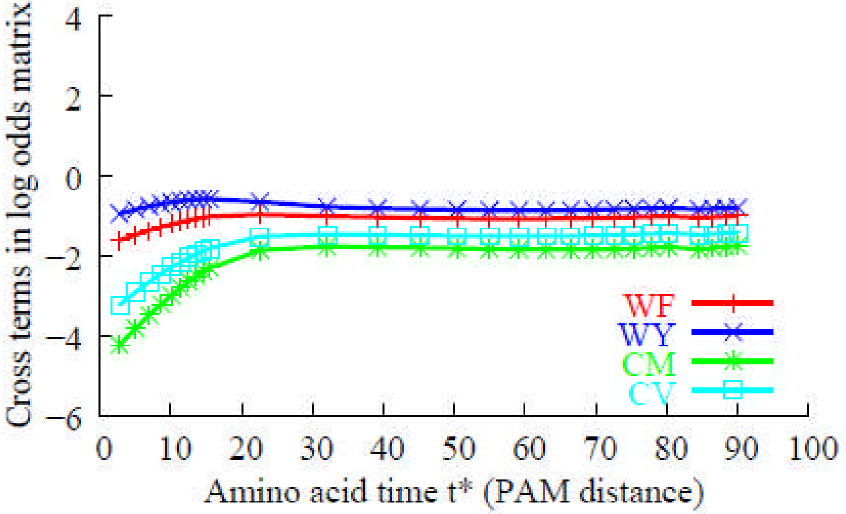


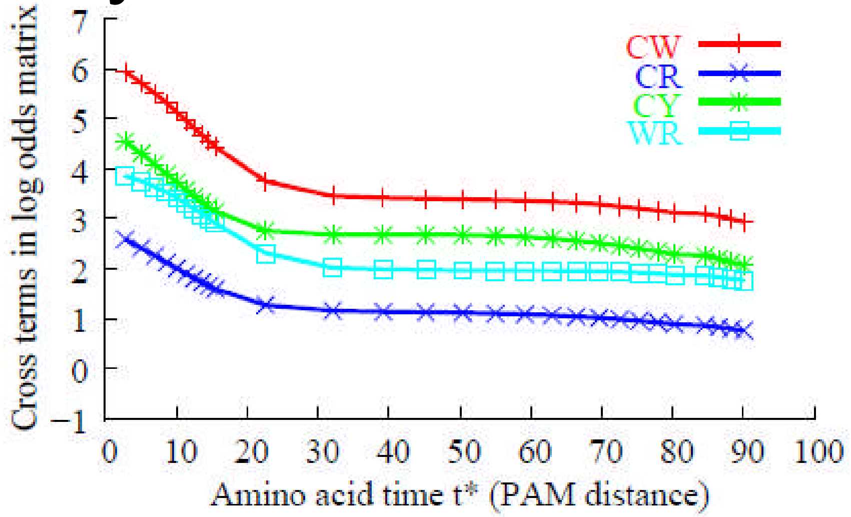
Benner *et al.* evidence:



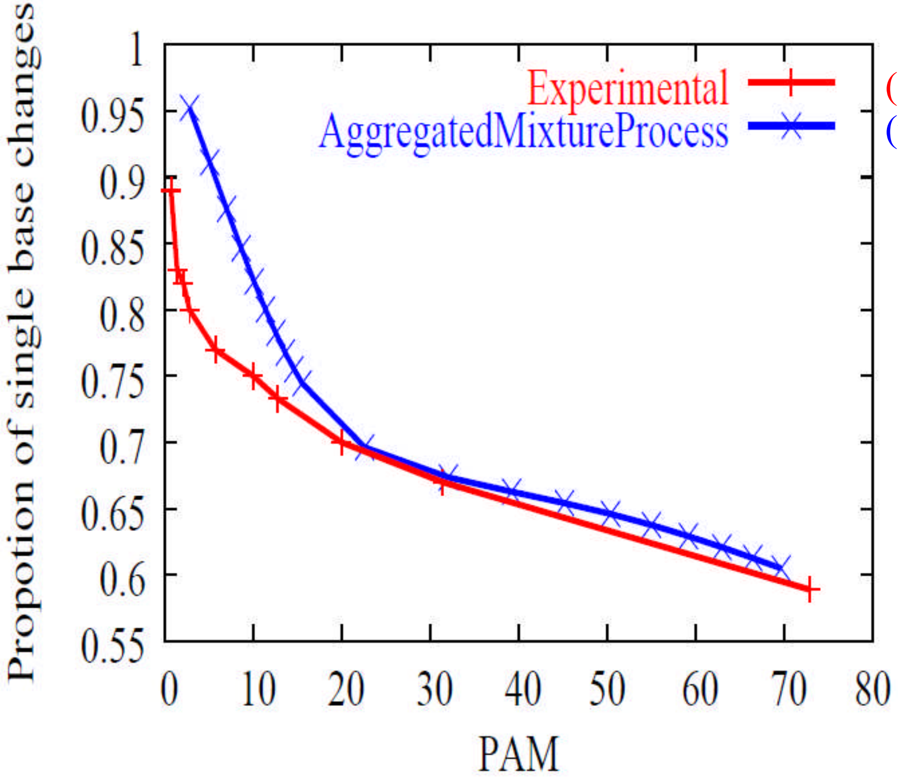


this study:





Mitchison & Durbin evidence:



(Mitchison & Durbin)

(this study)

Are Markov process models appropriate for protein sequence evolution?

PROCEED

WITH

CAUTION

Things to remember from Nick's talk:

- evolution should look the same whether we study it 100MYA or 1MYA or 1YA or today or tomorrow or …

- published evidence of non-Markov protein evolution can be explained by a time-independent codon model-based AMP

- we may proceed with current approaches to sequence evolution based on Markov models!

- possible consequences: non-Markov evolution of:
  - protein sequences
  - purine/pyrimidine (R/Y) encoded DNA (nucleotide-based AMP)