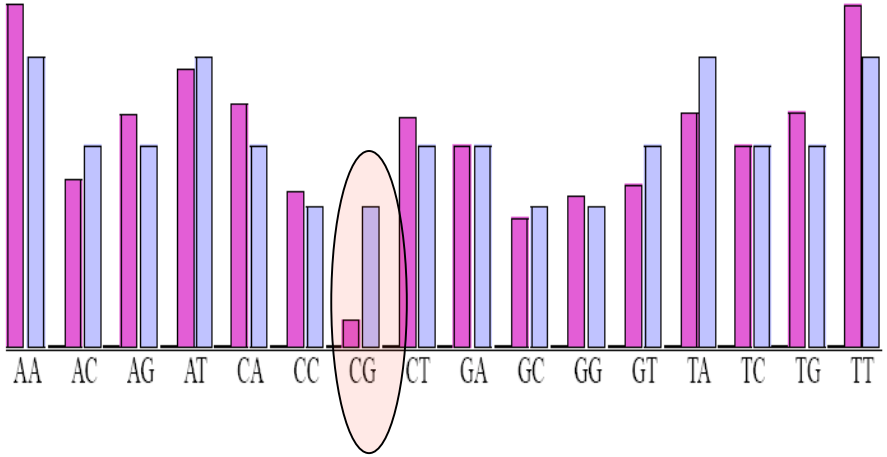
Context-dependent substitution models, molecular clocks, and the arrow of time

Gerton Lunter

Budapest, June 2008



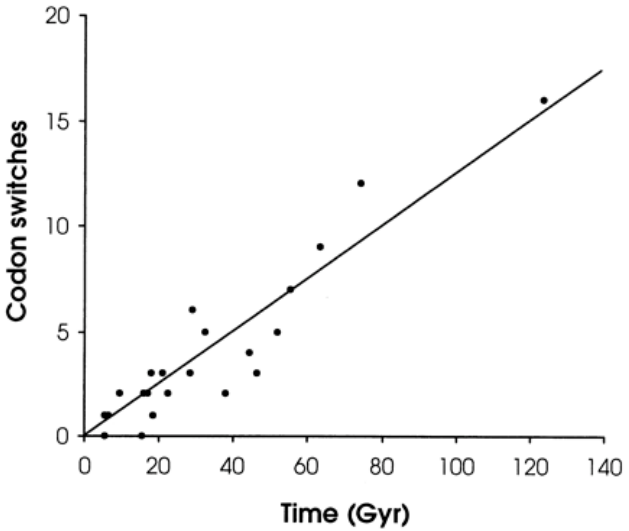
Observed
Expected

Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions

Michalis Averof,[1]* Antonis Rokas,[2] Kenneth H. Wolfe,[3] Paul M. Sharp[4]*

Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe, UUC Phe	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr	UGU Cys, UGC Cys	U
	UUA Leu, UUG Leu		UAA Stop, UAG Stop	UGA Stop, UGG Trp	C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
			CAA Gln, CAG Gln		
A	AUU Ile, AUC Ile, AUA Ile	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn	AGU Ser, AGC Ser	U C A G
	AUG Met		AAA Lys, AAG Lys	AGA Arg, AGG Arg	
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G
			GAA Glu, GAG Glu		



Adv. Appl. Prob. **32,** 499–517 (2000)
Printed in Northern Ireland
© *Applied Probability Trust* 2000

PROBABILISTIC MODELS OF DNA SEQUENCE EVOLUTION WITH CONTEXT DEPENDENT RATES OF SUBSTITUTION

JENS LEDET JENSEN,* *Aarhus University*
ANNE-METTE KRABBE PEDERSEN,** *Aarhus University*

Model:

nnnnn**abc**nnnnn → nnnnn**ab'c**nnnnn: Rate λ(a,b,b',c)

Proposition:

The process is reversible, with a stationary distribution that has a (first order) Markov structure, iff λ can be written as

$$\lambda(a,b,b',c) = \exp [\varphi(a,b',c) + \Psi(a,b,b',c)]$$

with Ψ symmetric in b and b'.

Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood

Adam Siepel and David Haussler*†*

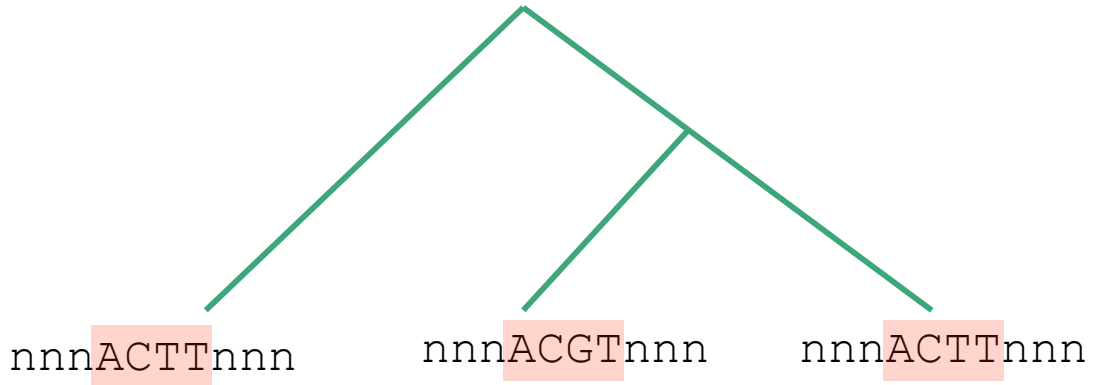
Mol. Biol. Evol. 21(3):468–488. 2004

Model: Rates depend on both neighbouring nucleotides
Double mutations forbidden

Inference: “sliding window” approximation

$$P(x_\bullet \mid \psi) = \prod_{i=N}^{L} P(x_{\bullet,i} \mid x_{\bullet,i-N+1} \ldots, x_{\bullet,i-1}, \psi)$$

$$P(x_{\bullet,i} \mid x_{\bullet,i-N+1}, \ldots, x_{\bullet,i-1}, \psi) = \frac{P(x_{\bullet,i-N+1}, \ldots, x_{\bullet,i-1}, x_{\bullet,i} \mid \psi)}{\sum_{\tilde{x}_{\bullet,i}} P(x_{\bullet,i-N+1}, \ldots, x_{\bullet,i-1}, \tilde{x}_{\bullet,i} \mid \psi)}$$

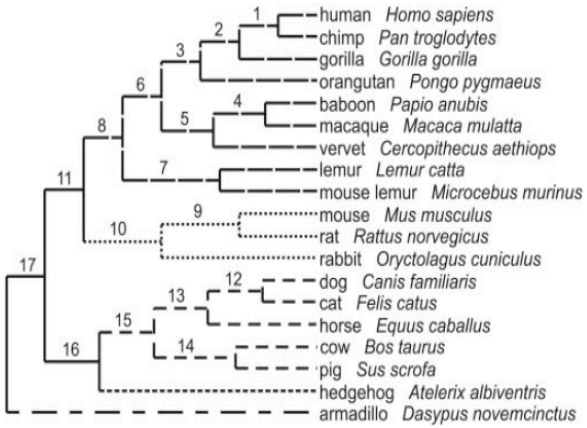


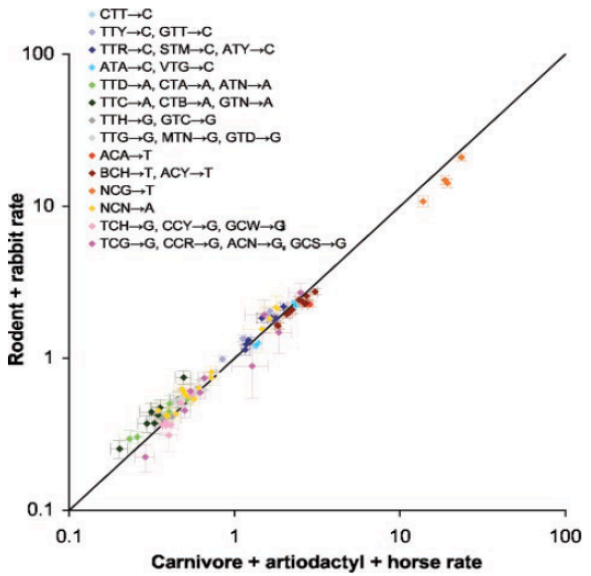
Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution

Dick G. Hwang*† and Phil Green*†‡

Model: Rates depend on both neighbouring nucleotides
Double mutations forbidden

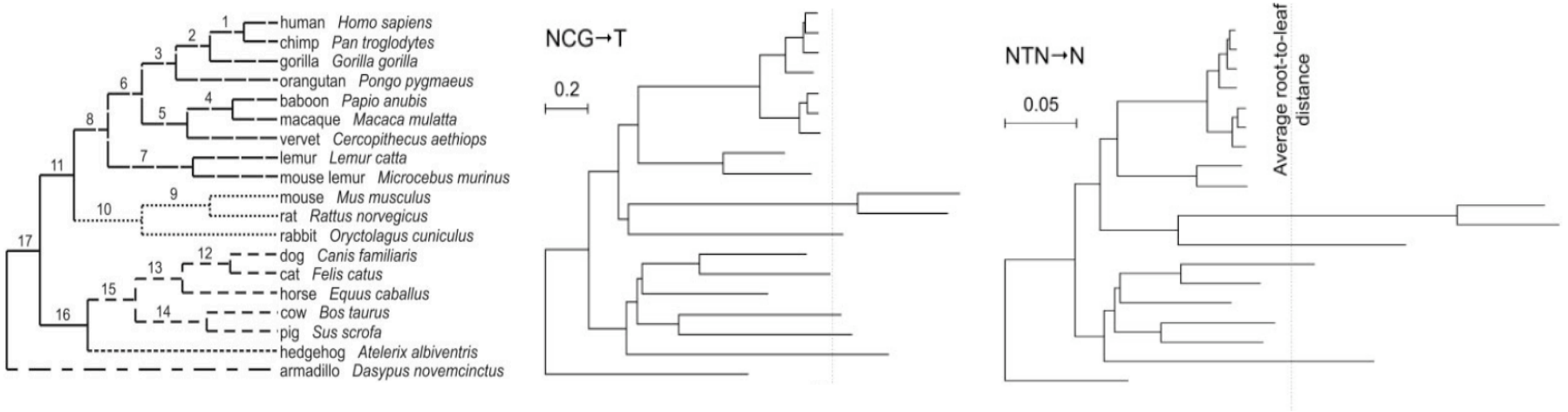
Inference: Bayesian MCMC
Discretized time; one mutation per time-slice



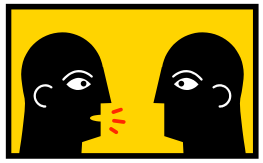


Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution

Dick G. Hwang*† and Phil Green*†‡



(switch to PDF)



Discussion

Context dependent effects: Should We Bother?

Yes:

- CpG effects are real, and strong (15 x mean background)
- CpG mutations are implicated in genetic disorders
- Cause of mutational inhomogeneity, particularly at short distances
- Possibly most clock-like signal we have for mammals?

No:

- Inclusion complicates model tremendously
- Makes very small difference in likelihood: don’t bother for alignments!
- Does it influence gene finding; motif evolution; phylogenetic inferences?