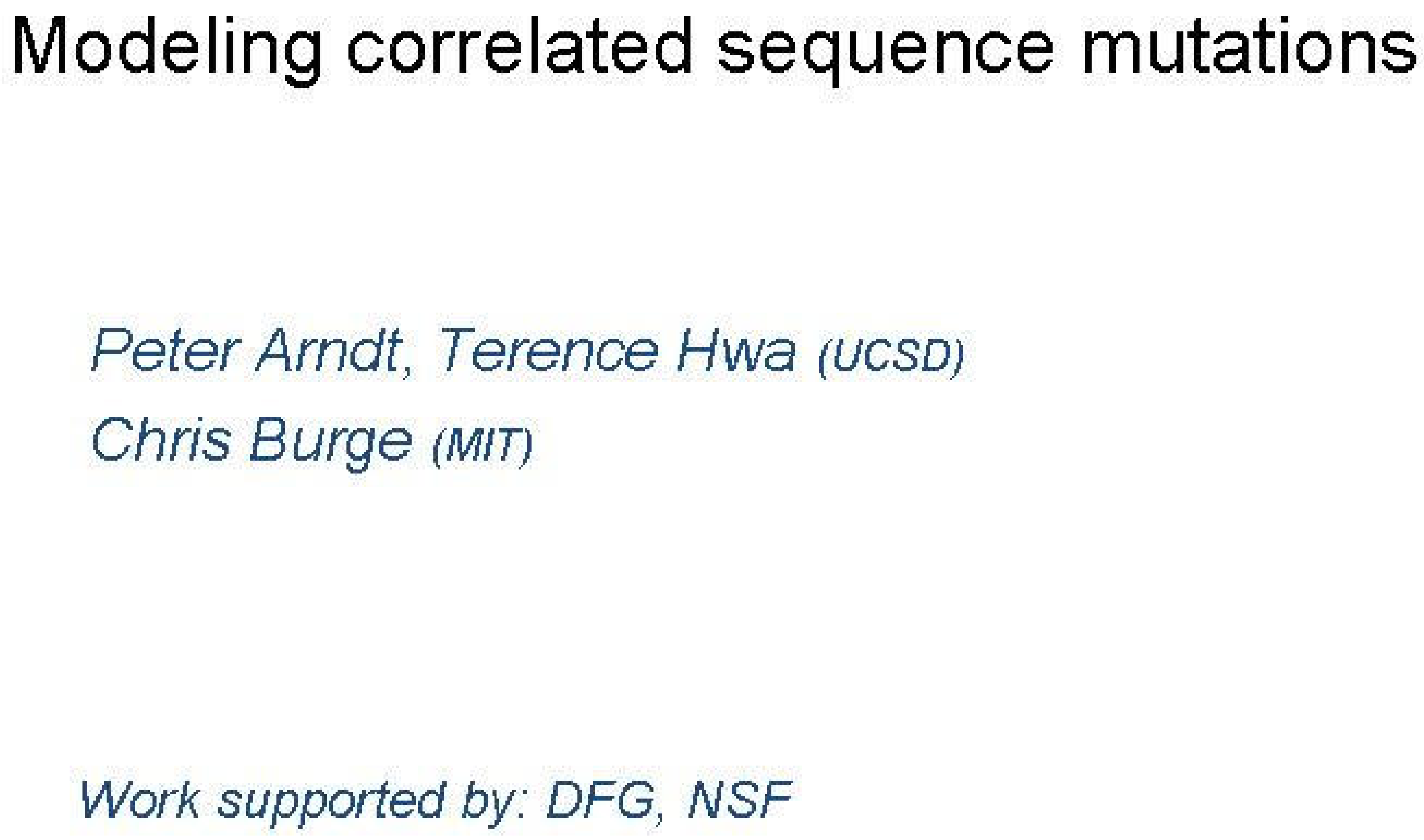


Modeling correlated sequence mutations

Peter Arndt, Terence Hwa (UCSD)
Chris Burge (MIT)

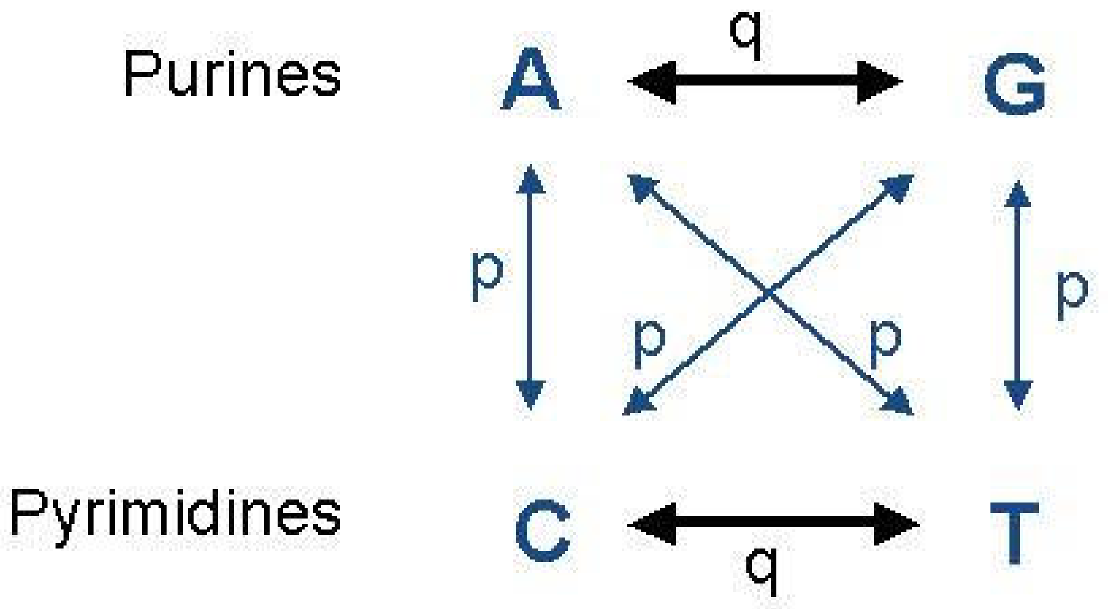
Work supported by: DFG, NSF

Outline

- Uncorrelated and correlated mutation processes
- Modeling
  - Computing the stationary state
  - Dynamic behavior
- Application
  - Mutations in Human Alu repeats

Point Mutations

One distinguishes the following point mutation processes:



Transversions:

Rate: p ~10^{-9} /(bp year)
(for Humans)

Transitions:

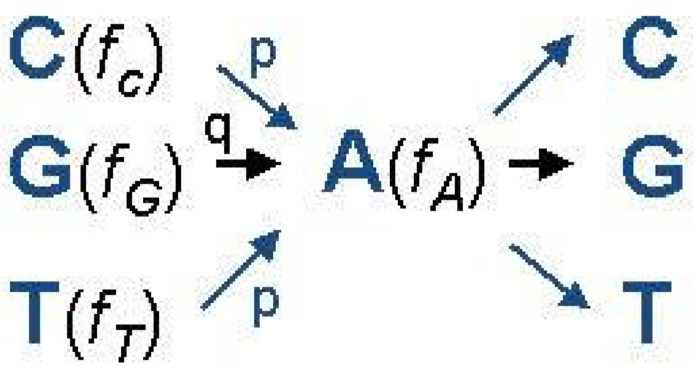
Rate: q ≈ 4 p

[Walker et al. 1999, Kapitonov 1995]

The stationary state

… is quite simple□□□□□□

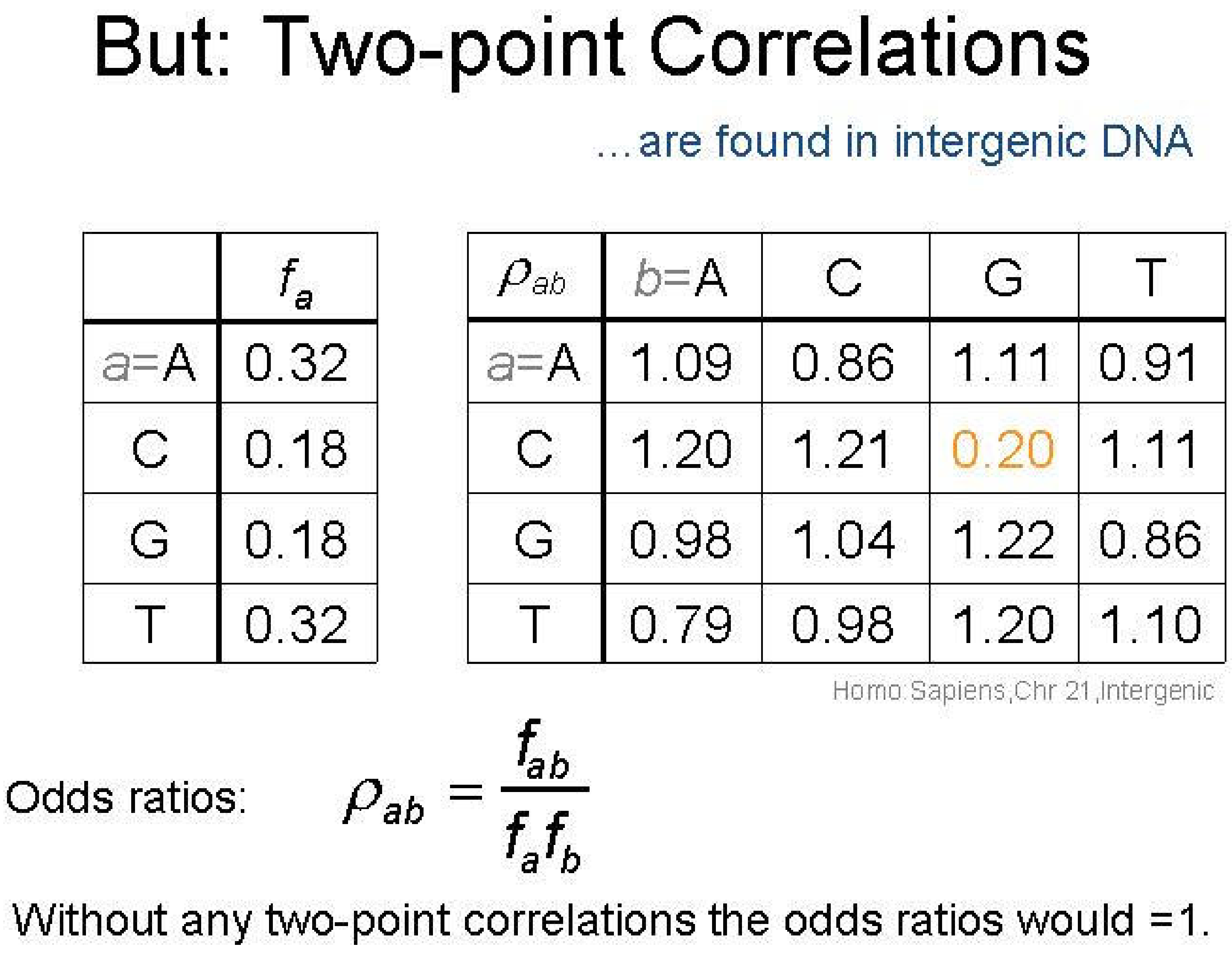
- no interactions between neighboring bases
- state space is 4-dimensional



$$0=\frac{\partial}{\partial t}\begin{pmatrix} f_A \\ f_G \\ f_C \\ f_T \end{pmatrix}=\begin{pmatrix} d & q & p & p \\ q & d & p & p \\ p & p & d & q \\ p & p & q & d \end{pmatrix}\begin{pmatrix} f_A \\ f_G \\ f_C \\ f_T \end{pmatrix}$$

$$d=-2p-q$$

- f_A=f_C=f_G=f_T=0.25 (for all p and q)
- CG content f_C+f_G= 50%
- Pair-correlations are simply: f_{ab}= f_a f_b



But: Two-point Correlations

...are found in intergenic DNA

	f_a
a=A	0.32
C	0.18
G	0.18
T	0.32

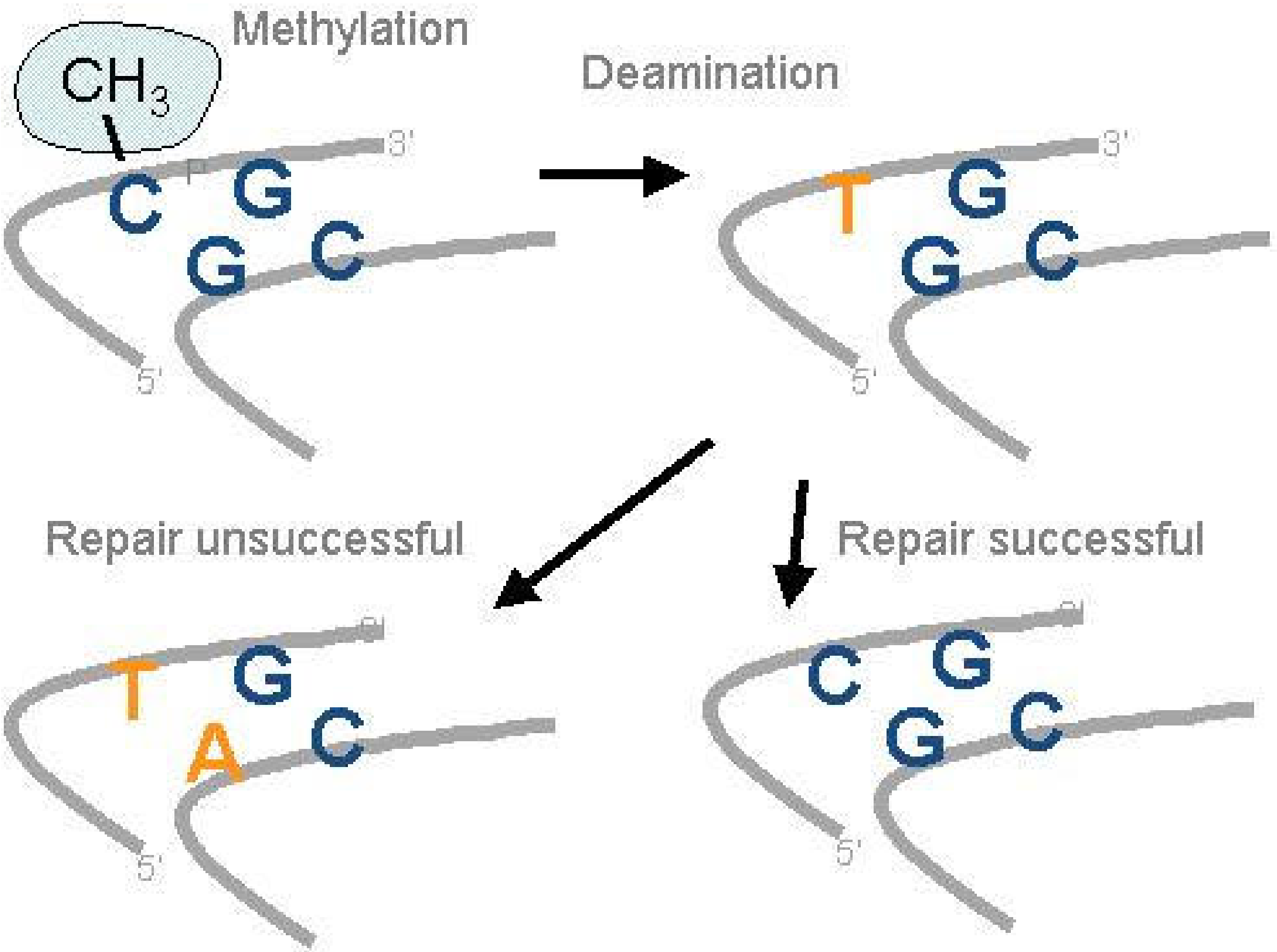
ρ_{ab}	b=A	C	G	T
a=A	1.09	0.86	1.11	0.91
C	1.20	1.21	0.20	1.11
G	0.98	1.04	1.22	0.86
T	0.79	0.98	1.20	1.10

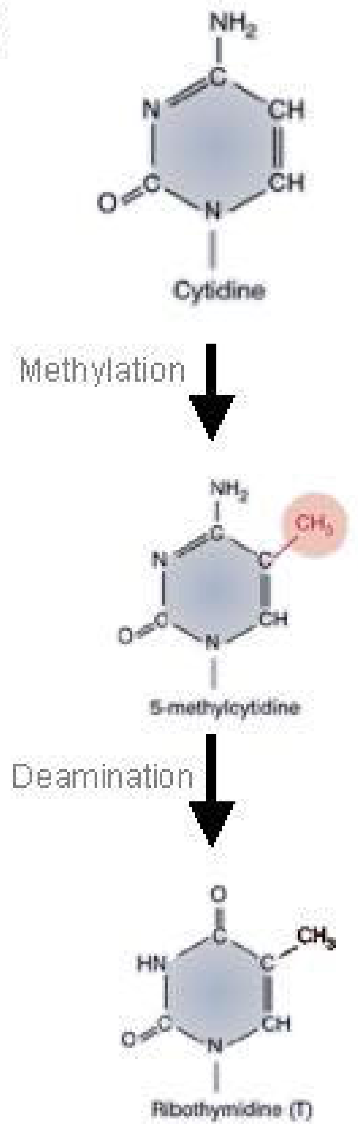
Homo Sapiens,Chr 21,Intergenic

Odds ratios: $\rho_{ab} = \frac{f_{ab}}{f_a f_b}$

Without any two-point correlations the odds ratios would =1.

Correlated Mutations

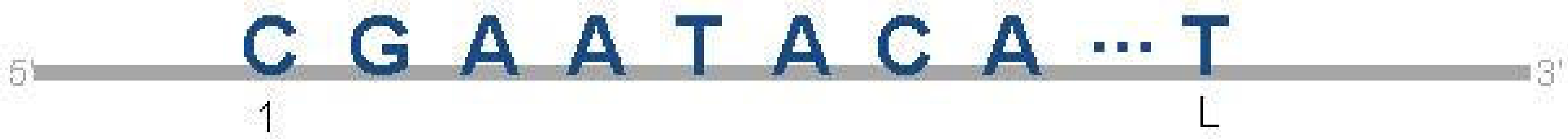


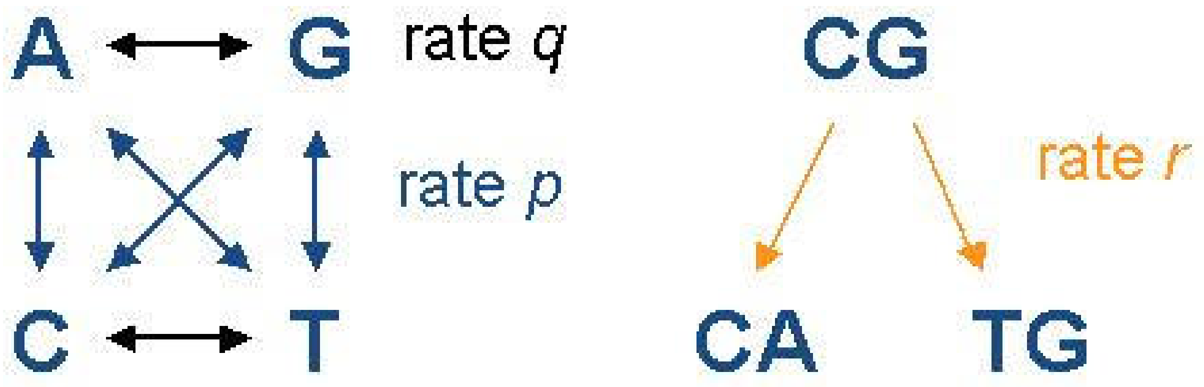


CpG ➡ TpG or CpA

rate ≈ 25 x Transversion rate

The Model

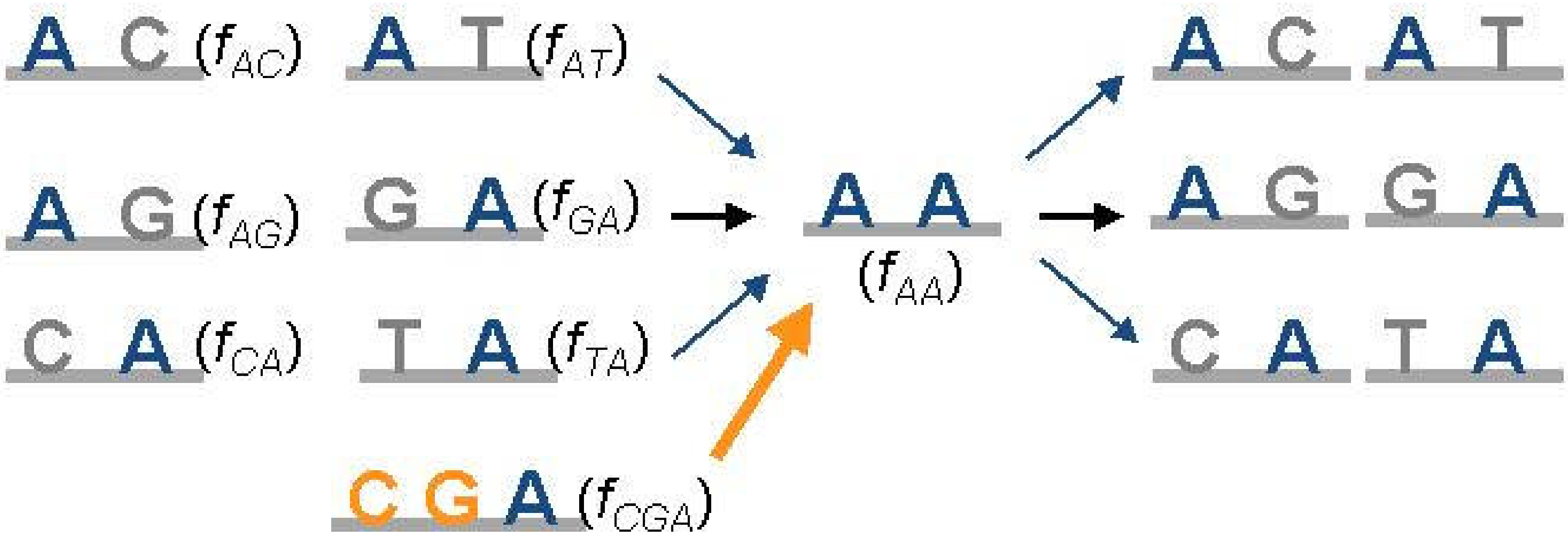




Transitions, Transversions, Correlated Pair Mutations (unidirectional)

Computing the stationary state

… is not equal to a enlargement of the alphabet



- no detailed balance, since there is back-reaction for the correlated mutation process
- two point correlations depend on three-point correlations
- non-equilibrium dynamics

Rate Equations

The corresponding rate equation for the above process:

$$\frac{\partial}{\partial t} f_{AA} = + p f_{\mathbf{C}A} + q f_{\mathbf{G}A} + p f_{\mathbf{T}A} - (2p+q) f_{\mathbf{A}A}$$

$$+ p f_{A\mathbf{C}} + q f_{A\mathbf{G}} + p f_{A\mathbf{T}} - (2p+q) f_{A\mathbf{A}}$$

$$\boxed{+ r f_{\mathbf{CG}A}}$$

Gives 16 Eq. for $f_{ab}(t)$ + 64 Eq. for $f_{abc}(t)$ +…

To truncate this hierarchy we use a Cluster Approximation:

$$\boxed{f_{abc} \approx f_{ab}\, f_{bc} \,/\, f_b}$$

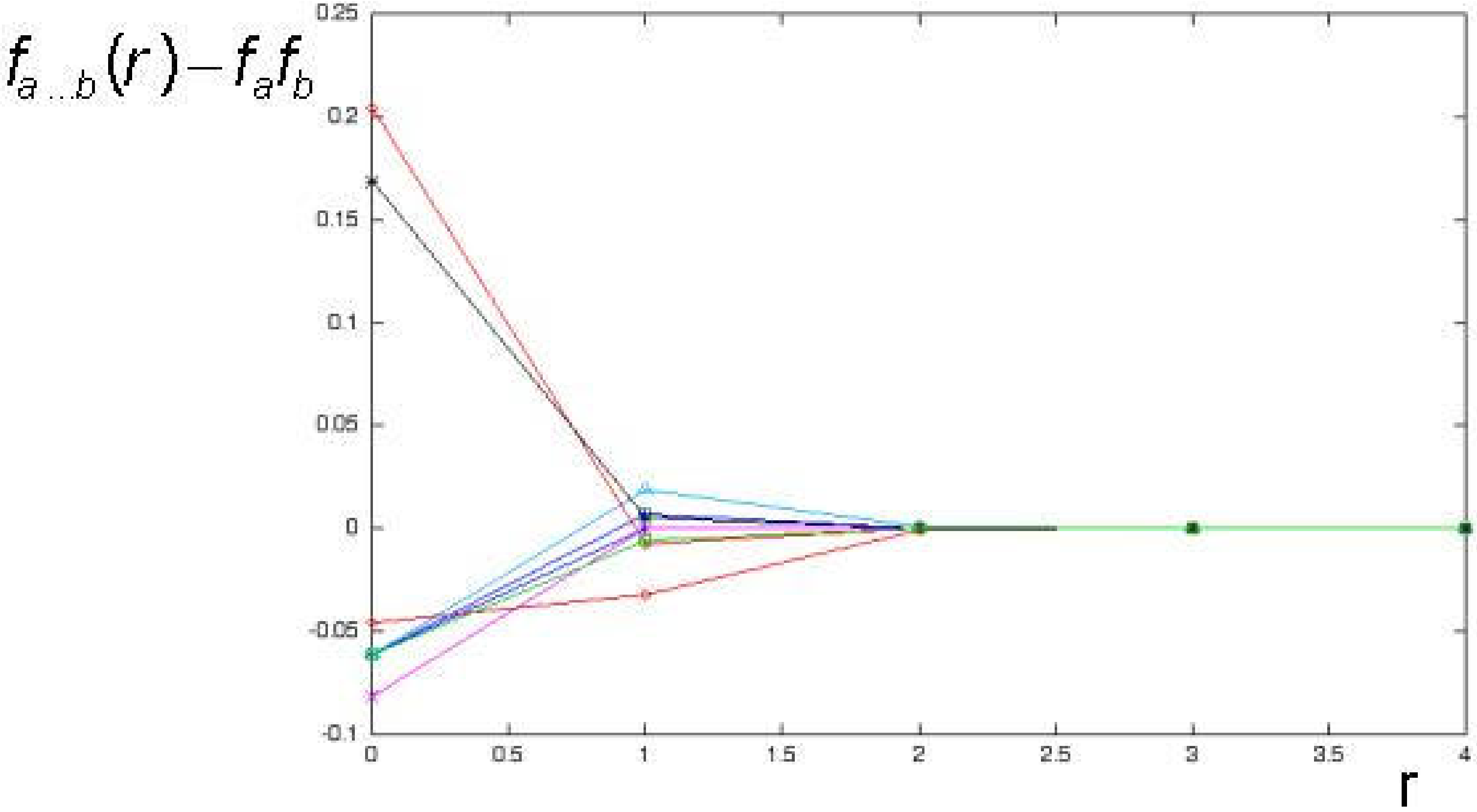
$$f_a = \sum_b f_{ab} = \sum_b f_{ba}$$

Solve the 16 non-linear differential eq. for the steady state:

$$\frac{\partial}{\partial t} f_{AA} = G(f_{AA}, f_{AC}, \ldots, f_{TT}) = 0$$

Monte Carlo Simulations

... show that the correlation length is small



➔ Cluster Approximation justified

Solution of Cluster Approx.

The frequencies are given by:

$$f_A = f_T = \frac{1}{4} + \frac{\Delta}{2} \qquad f_C = f_G = \frac{1}{4} - \frac{\Delta}{2}$$

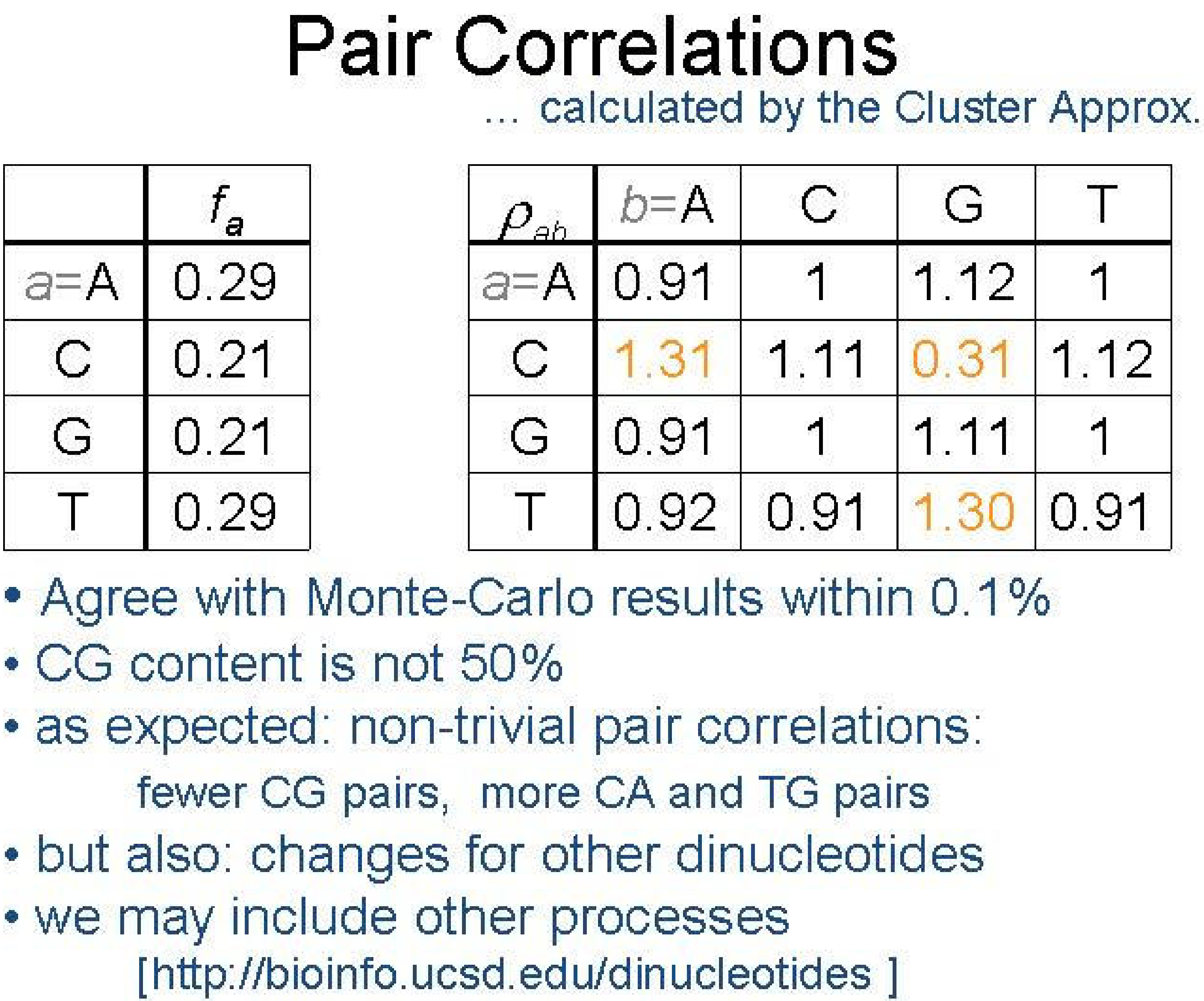
$$\Delta = \frac{(3p+q)r}{16(p+q)(3p+q) + 4(7p+3q)r}$$

Connected correlation functions:

$$\hat{f}_{CA} = \frac{(1+\Delta)\Delta}{4} \qquad \hat{f}_{CG} = -\frac{r(1-2\Delta)^2 - 16(p+q)\Delta}{16r}$$

$$\hat{f}_{CC} = -\frac{(2\Delta-1)(4r\Delta^2 + 8(2p+2q+r)\Delta - r)}{32(\Delta-1)}$$

$$\ddot{f}_{AC} = \ddot{f}_{AT} = \ddot{f}_{GC} = \ddot{f}_{GT} = 0 \qquad \hat{f}_{ab} = f_{ab} - f_a f_b$$



Pair Correlations

... calculated by the Cluster Approx.

	f_a
a=A	0.29
C	0.21
G	0.21
T	0.29

ρ_{ab}	b=A	C	G	T
a=A	0.91	1	1.12	1
C	1.31	1.11	0.31	1.12
G	0.91	1	1.11	1
T	0.92	0.91	1.30	0.91

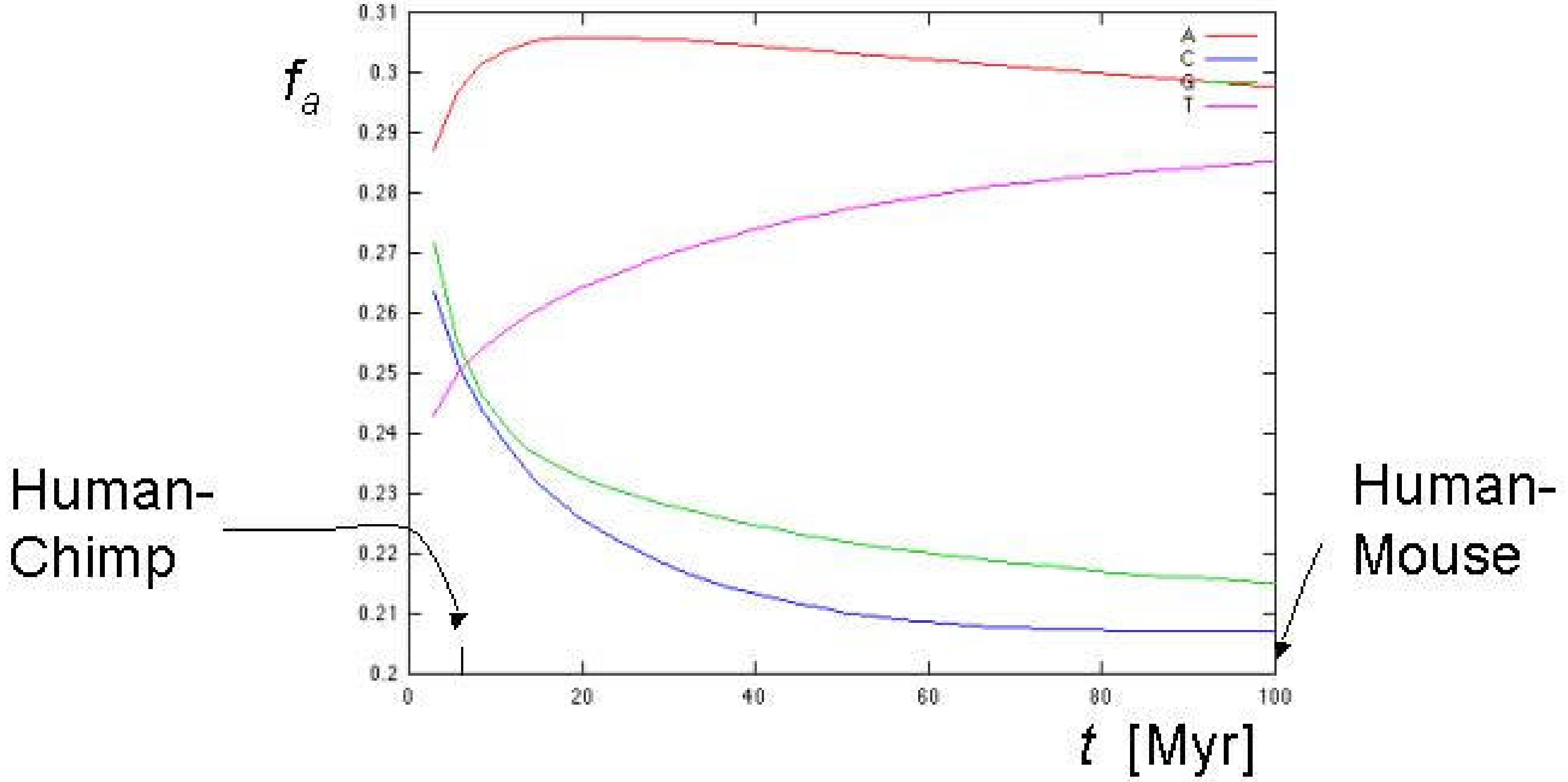
- Agree with Monte-Carlo results within 0.1%
- CG content is not 50%
- as expected: non-trivial pair correlations:

  fewer CG pairs, more CA and TG pairs
- but also: changes for other dinucleotides
- we may include other processes

  [http://bioinfo.ucsd.edu/dinucleotides]

Dynamic Behavior

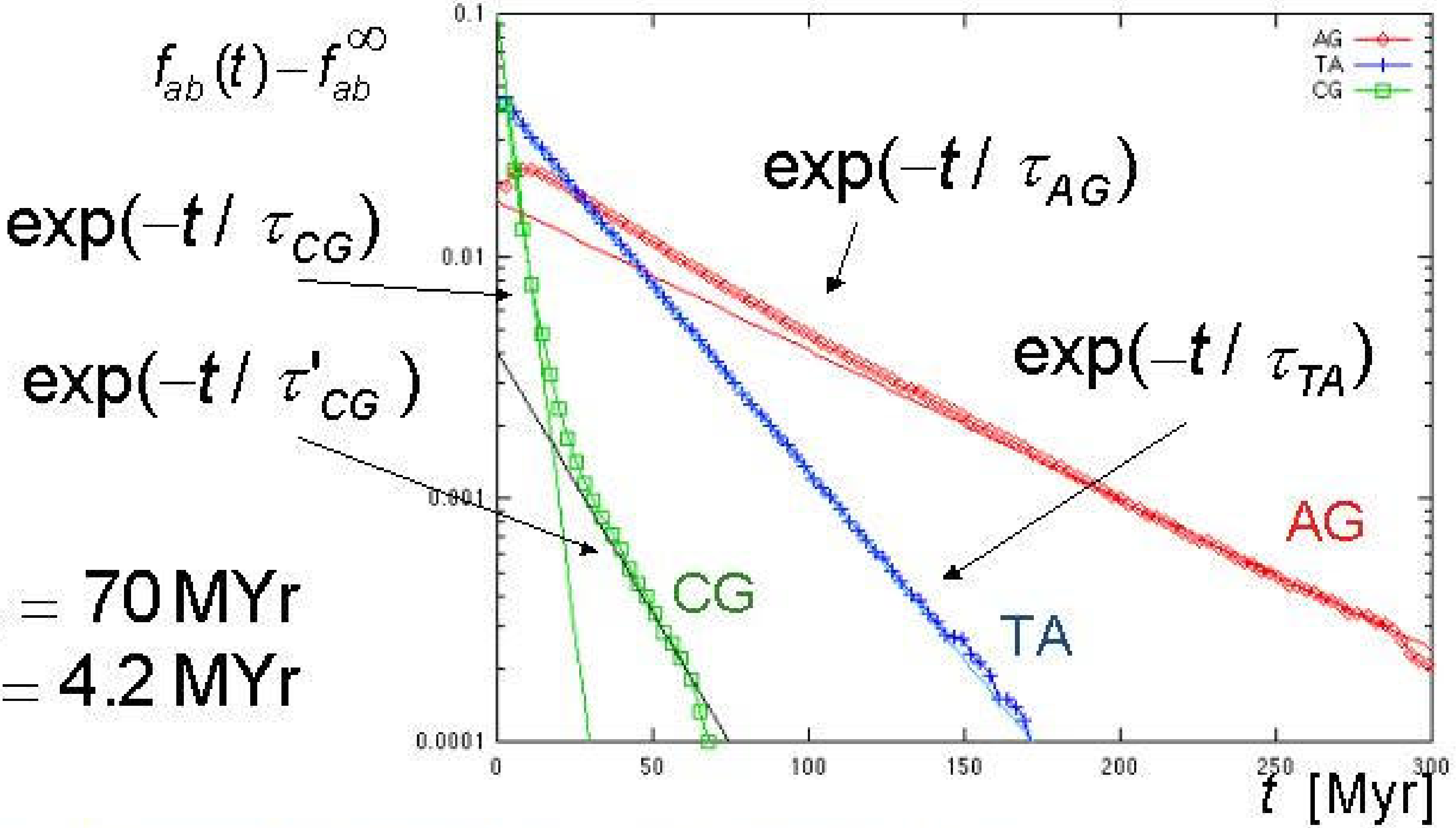
… by Monte-Carlo Simulations



Stationary state is reached over some tens of Myr.

Initial conditions & dynamics matter on smaller time-scales

Relaxation of Dinucleotide Corr.



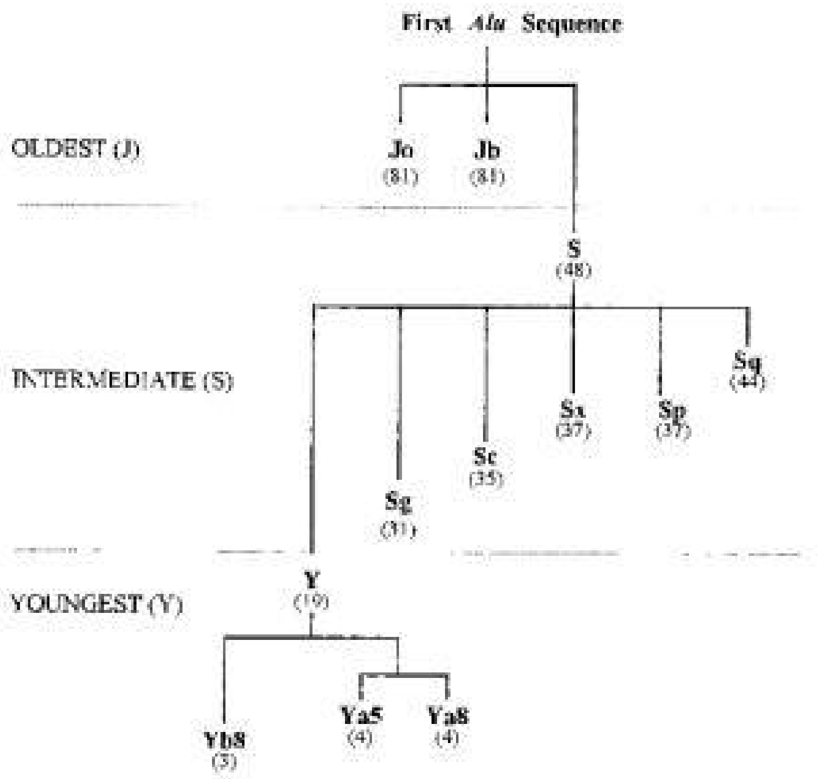
$\tau_{AG} = 70\,\mathrm{MYr}$

$\tau_{CG} = 4.2\,\mathrm{MYr}$

- Many different timescales ➜ different clocks
- Time-scales (solid lines) calculated by cluster approx.

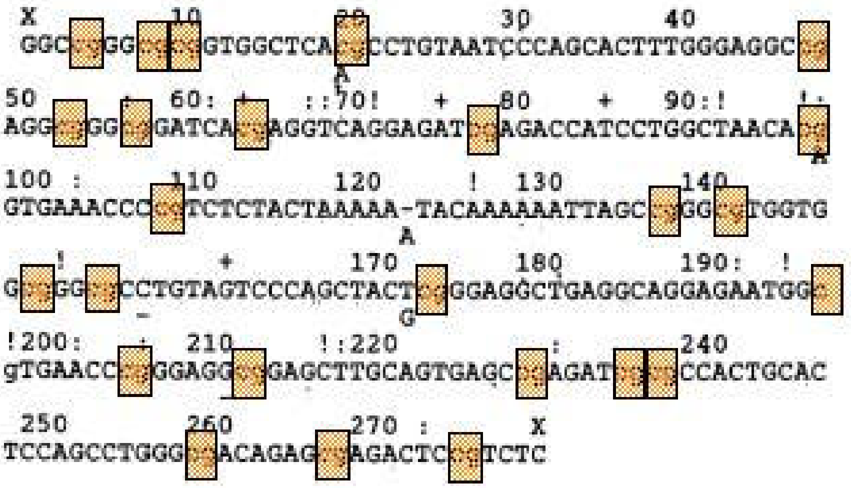
Alu Repeats

- Retrotransposon
- 280 bp long
- the master sequence is well conserved and CpG rich
- the oldest Alu's are about 60 Myr old, we may use the relaxation of the CG dinucleotide
- about 10^6 Alu sequences in the Human Genome (≈10%)



Changes on CpG and non-CpG sites

Master sequence:

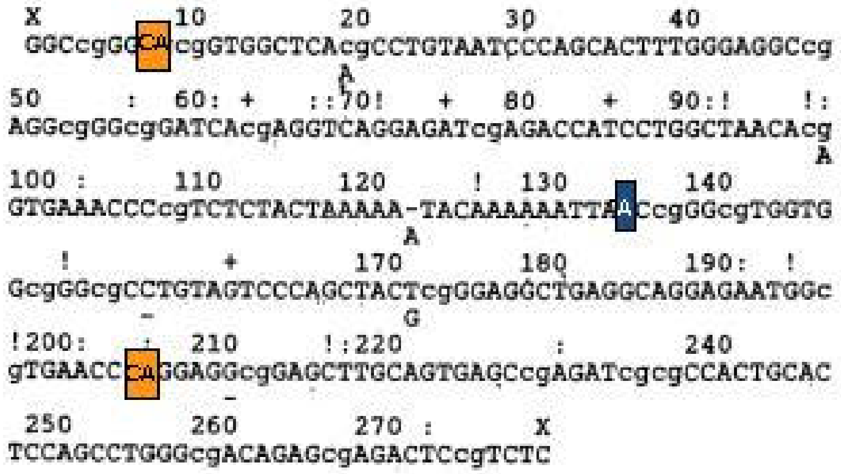


```
 X        10        20        30        40
 GGCcgGGcgcgGTGGCTCAcgCCTGTAATCCCAGCACTTTGGGAGGCcg
                    A
                    C
50     :  60: +   ::70!   +   80     +   90:!    !:
AGGcgGGcgGATCAcgAGGTCAGGAGATcgAGACCATCCTGGCTAACAcg
                                                 A
100 :     110       120     !  130       140
GTGAAACCCcgTCTCTACTAAAAA-TACAAAAAATTAGCcgGGcgTGGTG
                        A
  !         +        170       180       190:  !
GcgGGcgCCTGTAGTCCCAGCTACTcgGGAGGCTGAGGCAGGAGAATGGc
         -               G
!200:   :  210    !:220          :        240
gTGAACCcgGGAGGcgGAGCTTGCAGTGAGCcgAGATcgcgCCACTGCAC
               -
 250       260       270 :      X
TCCAGCCTGGGcgACAGAGcgAGACTCcgTCTC
```

[Britten et al, 1988]

▨ CpG site

evolved sequence:



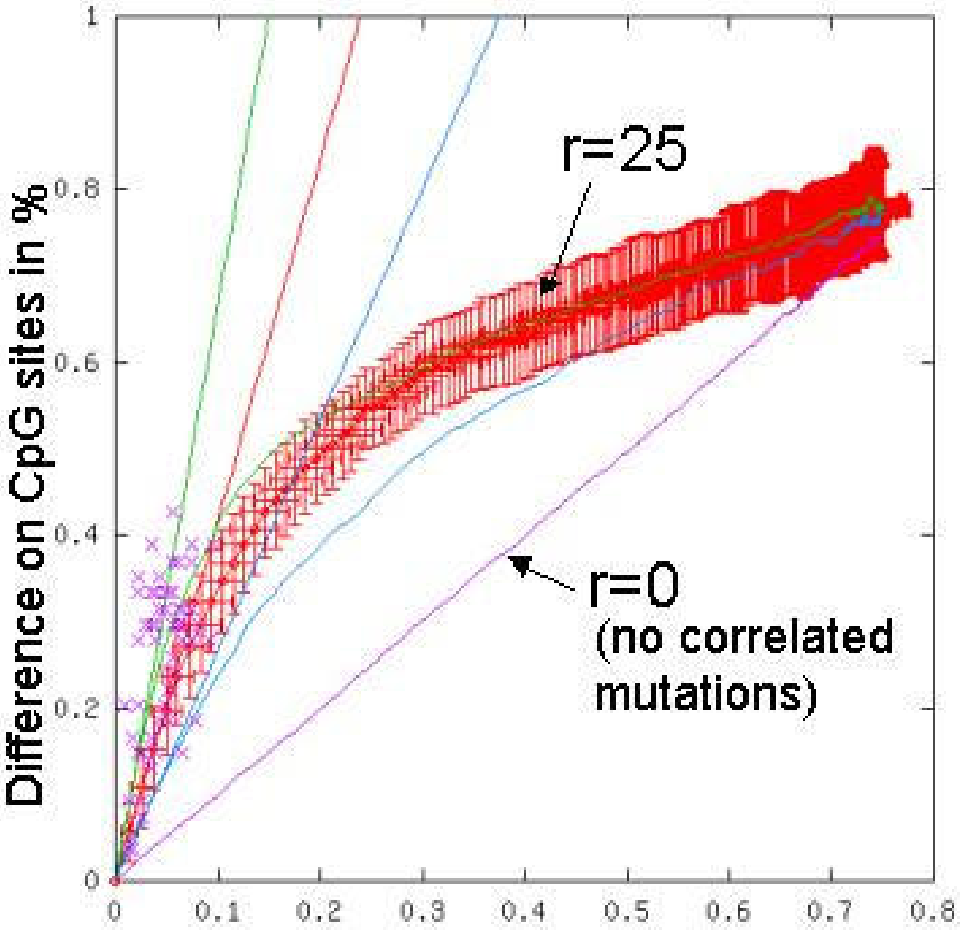
```
 X        10        20        30        40
 GGCcgGGCAcgGTGGCTCAcgCCTGTAATCCCAGCACTTTGGGAGGCcg
                    A
                    C
50     :  60: +   ::70!   +   80     +   90:!    !:
AGGcgGGcgGATCAcgAGGTCAGGAGATcgAGACCATCCTGGCTAACAcg
                                                 A
100 :     110       120     !  130       140
GTGAAACCCcgTCTCTACTAAAAA-TACAAAAAATTAACcgGGcgTGGTG
                        A
  !         +        170       180       190:  !
GcgGGcgCCTGTAGTCCCAGCTACTcgGGAGGCTGAGGCAGGAGAATGGc
         -               G
!200:   :  210    !:220          :        240
gTGAACCCAGGAGGcgGAGCTTGCAGTGAGCcgAGATcgcgCCACTGCAC
               -
 250       260       270 :      X
TCCAGCCTGGGcgACAGAGcgAGACTCcgTCTC
```

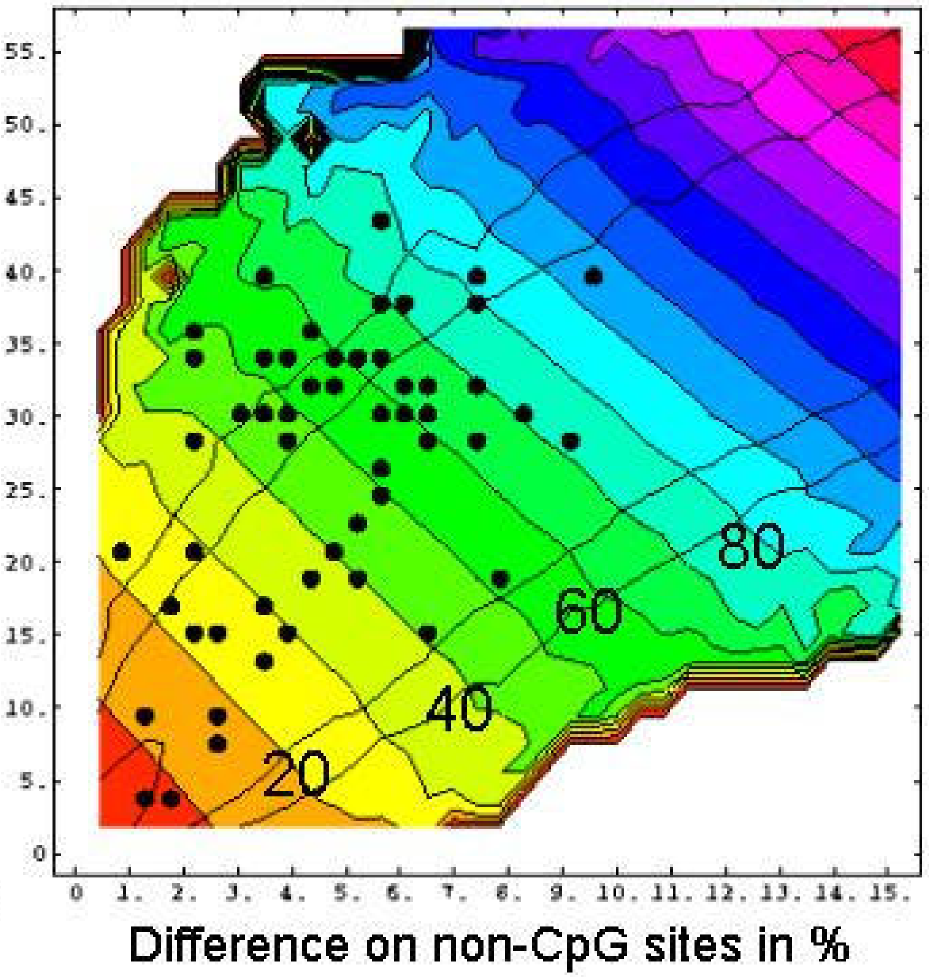
🟧 🟦 changed CpG, non-CpG site

We count the number of changes on CpG and non-CpG positions and compare them with expectations from the model

Changes on CpG and non-CpG sites

... Monte-Carlo Simulations of the model





[Alu data taken from Britten et al, 1988]

Summary

- The pattern of dinucleotide correlations let us deduce the underlying mutation processes
- Different correlations relax with different rates → different clocks

Outlook

- Incorporation of the model into DNA Sequence Evolution
  - Useful for comparative Genomics approach to gene finding, motif finding, …