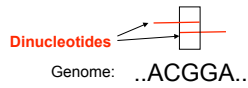


## Advanced Questions in Sequence Evolution Models

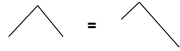
• **Di-nucleotide events**



• **Context-dependent models**



• **Irreversibility and rooting**



• **Probabilities of different paths**

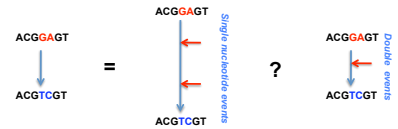


• **Rate Variation**

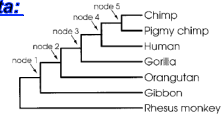
ATTGCGTCCAATATTGCGTCCAAT

## Di-nucleotide events

**The Problem:**



**Data:**



**Analysis and Conclusion:**

Assuming JC69 + doublet mutations.

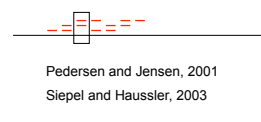
00:  $10^{-8}$  doublet mutation rate, ~10% of singlet rate  
03: much less for a large more reliable data set

## Context-dependent models

**From singlet models to doublet models:**

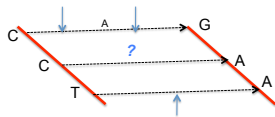
- Independence
- Independence with CG avoidance
- Strand symmetry
- Only single events
- Single events with simple double events

**Contagious Dependence:**



**The Problem:**

What is PIC → AJ?



## The Gibbs Sampler

$x^{(t)} = (x_1^{(t)}, \dots, x_d^{(t)})$  for iteration t. At iteration t + 1

For  $i=1, \dots, d$ : Draw  $x_i^{(t+1)}$  from conditional distribution  $\pi(\cdot | x_{[-i]}^{(t)})$  and leave remaining components unchanged, i.e.  $x_{[-i]}^{(t+1)} = x_{[-i]}^{(t)}$

Both random & systematic scan algorithms leaves the true distribution invariant.

$$\pi(x_i^{(t+1)} | x_{[-i]}^{(t)}) \times \pi(x_{[-i]}^{(t)}) = \pi(x_{[-i]}^{(t+1)}, x_i^{(t+1)})$$

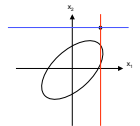
**An example:**

Target Distribution is  $x = (x_1, x_2)$  is  $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$  distributed.

The conditional distributions are then:  $x_1^{(t+1)} | x_2^{(t)} \sim N\{\rho x_2^{(t)}, (1 - \rho)^2\}$ ,  
 $x_2^{(t+1)} | x_1^{(t)} \sim N\{\rho x_1^{(t)}, (1 - \rho)^2\}$ .

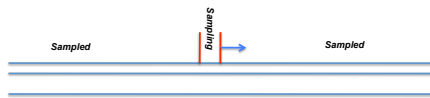
The approximating distribution after t steps of a systematic GS will be:

$$\begin{pmatrix} x_1^{(t)} \\ x_2^{(t)} \end{pmatrix} \sim N\left(\begin{pmatrix} \rho^{2t-1} x_1^0 \\ \rho^{2t} x_2^0 \end{pmatrix}, \begin{pmatrix} 1 - \rho^{2t-2} & \rho - \rho^{2t-1} \\ \rho - \rho^{2t-1} & 1 - \rho^{2t} \end{pmatrix}\right)$$



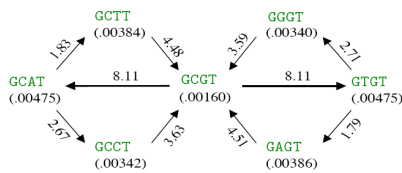
## Basic Dinucleotide model

• **Jensen-Pedersen sampler (2000) sampler**



**The Data:**

100 kb non-coding from chromosomes 22 and 10 from mouse and human.



- (.00160) etc.: Equilibrium probability of sequence
- 8.11 etc.: Net equilibrium flow  $\times 10^{-4}$

From Lunter & Hein, 2004

## Rooting using irreversibility (Lunter)

General rate mode for nucleotides - 12 parameters:

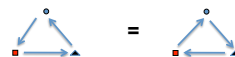
|   |           |           |           |           |
|---|-----------|-----------|-----------|-----------|
|   | A         | C         | G         | T         |
| A |           | $q_{A,C}$ | $q_{A,G}$ | $q_{A,T}$ |
| C | $q_{C,A}$ |           | $q_{C,G}$ | $q_{C,T}$ |
| G | $q_{G,A}$ | $q_{G,C}$ |           | $q_{G,T}$ |
| T | $q_{T,A}$ | $q_{T,C}$ | $q_{T,G}$ |           |

**Reversibility**

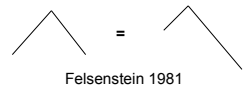
$$P(o \rightarrow \bullet) P(\bullet \rightarrow o) = P(\bullet \rightarrow o) P(o \rightarrow \bullet)$$

$$o \rightarrow \bullet = \bullet \rightarrow o$$

Reversible rate matrix:  $\pi_1 q_{1,2} = \pi_2 q_{2,1}$ , 9 parameters



**The Pulley Principle**



Felsenstein 1981

Irreversibility used for rooting  
Ziheng Yang 1994

