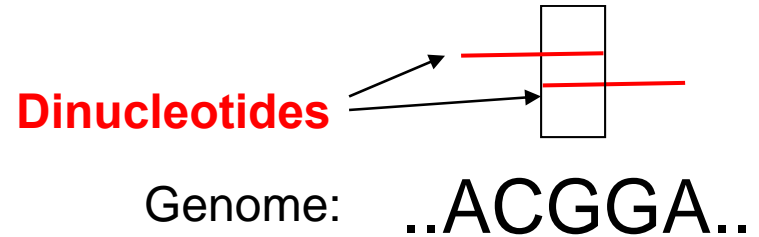


Advanced Questions in Sequence Evolution Models

- *Di-nucleotide events*

ACGGAGT
↓
ACGTCGT

- *Context-dependent models*



- *Irreversibility and rooting*



- *Probabilities of different paths*

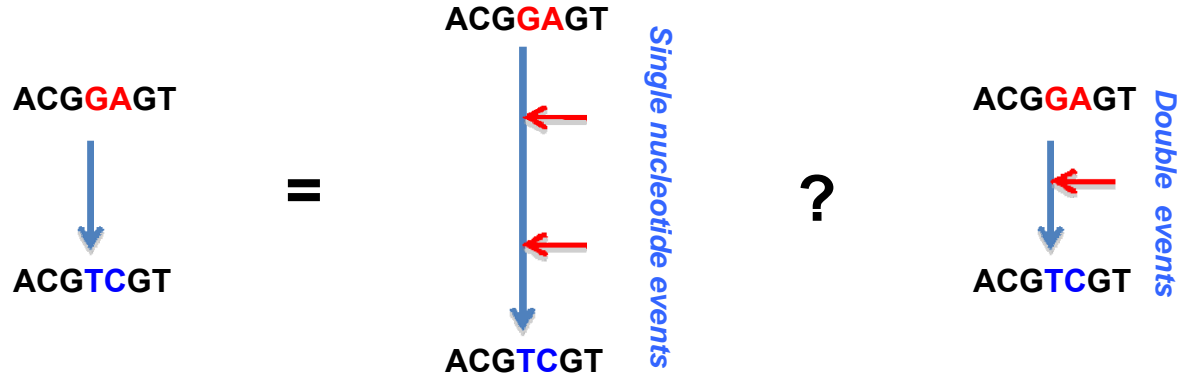


- *Rate Variation*

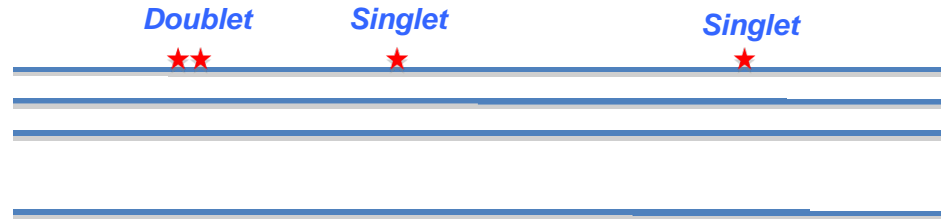
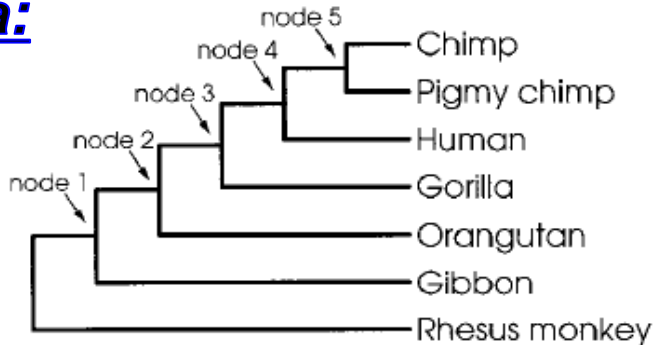
ATTGCGTCCAATATTGCGTCCAAT

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

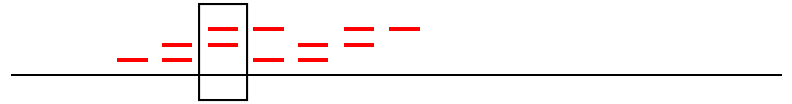
Independence with CG avoidance

Strand symmetry

Only single events

Single events with simple double events

Contagious Dependence:

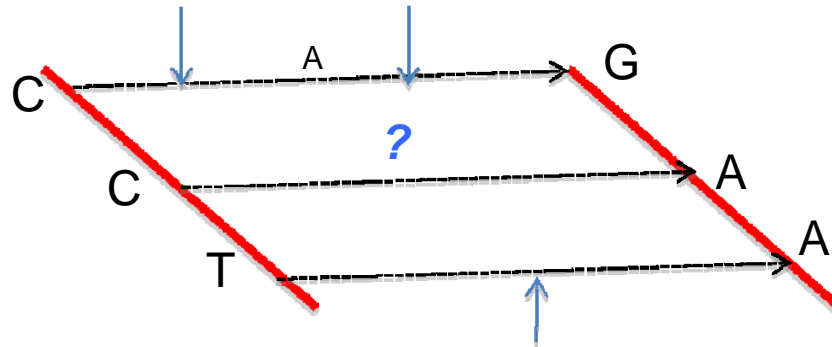


Pedersen and Jensen, 2001

Siepel and Haussler, 2003

The Problem:

What is $P[C \rightarrow A]$?



The Gibbs Sampler

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

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

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

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

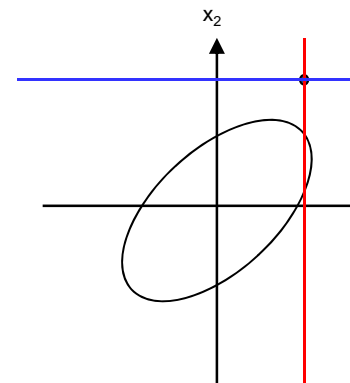
An example:

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: $x_2^{t+1} | x_1^{t+1} \sim N\{\rho x_1^{t+1}, (1 - \rho)^2\}$,
 $x_1^{t+1} | x_2^{t+1} \sim N\{\rho x_2^{t+1}, (1 - \rho)^2\}$,

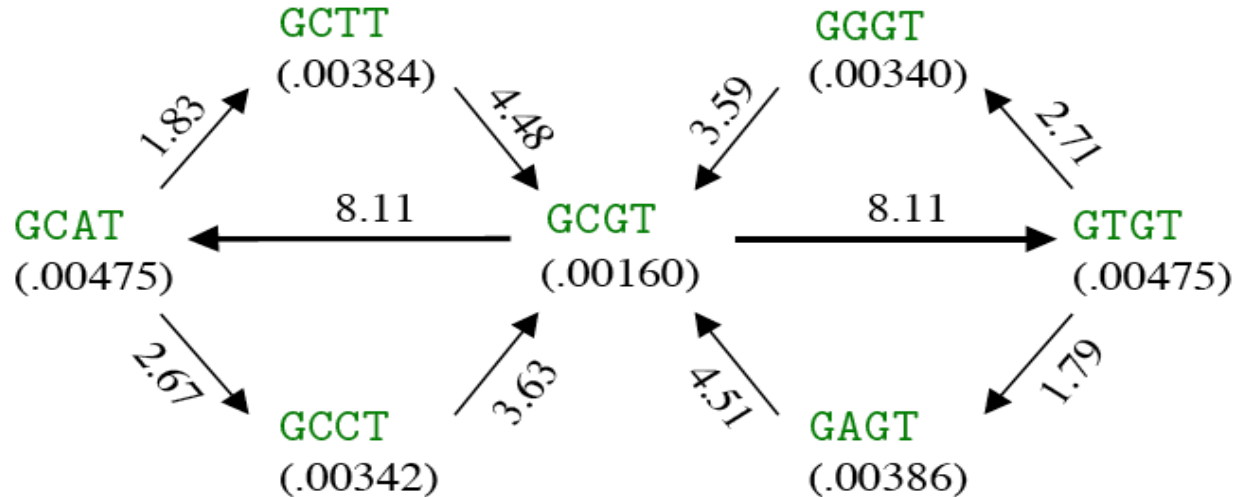
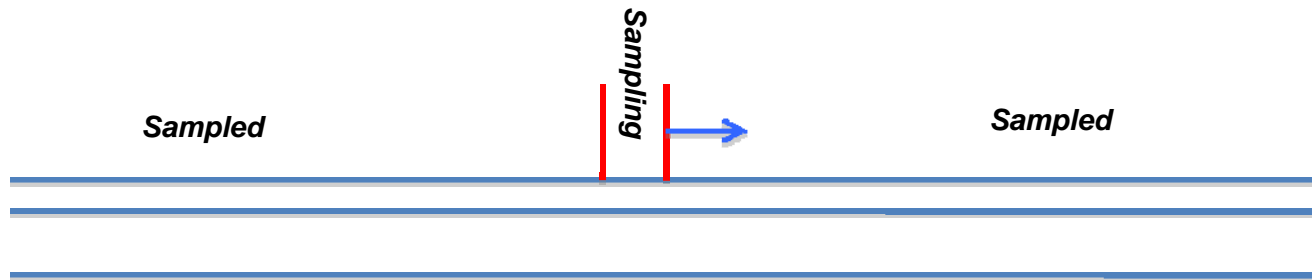
The approximating distribution after t steps of a systematic GS will be:

$$\begin{pmatrix} x_1^t \\ x_2^t \end{pmatrix} \sim N\left\{\begin{pmatrix} \rho^{2t-1} x_2^0 \\ \rho^{2t} x_2^0 \end{pmatrix}, \begin{pmatrix} 1 - \rho^{4t-2} & \rho - \rho^{4t-1} \\ \rho - \rho^{4t-1} & 1 - \rho^{4t} \end{pmatrix}\right\}$$



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 $\times 10^{-4}$

Rooting using irreversibility (Lunter)

General rate mode for nucleotides - 12 parameters:

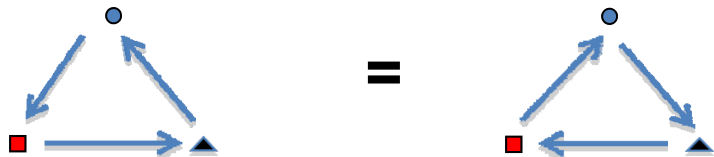
	A		C		G		T
A			$q_{A,C}$		$q_{A,G}$		$q_{A,T}$
C	$q_{C,A}$			$q_{C,G}$			$q_{C,T}$
G	$q_{G,A}$		$q_{G,C}$				$q_{G,T}$
T	$q_{T,A}$		$q_{T,C}$		$q_{T,G}$		

Reversibility

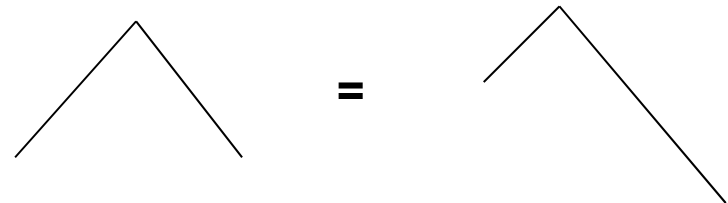
$$P(\bullet) * P(\bullet \rightarrow \blacksquare) = P(\blacksquare) * P(\blacksquare \rightarrow \bullet)$$

$$\bullet \rightarrow \blacksquare = \blacksquare \rightarrow \bullet$$

Reversible rate matrix : $\pi_i q_{i,j} = \pi_j q_{j,i}$ 9 parameters



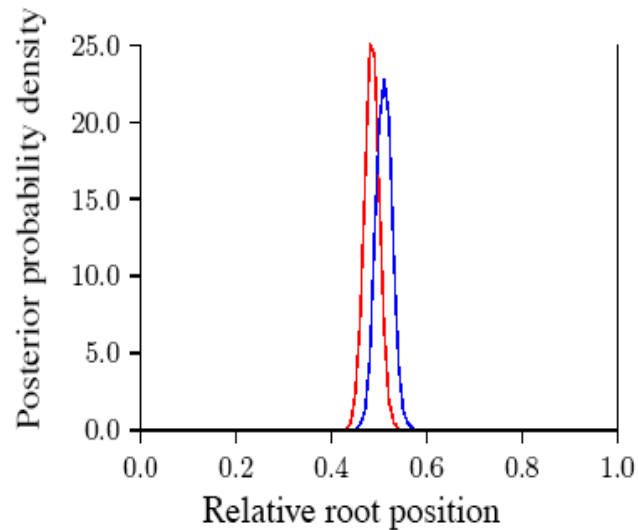
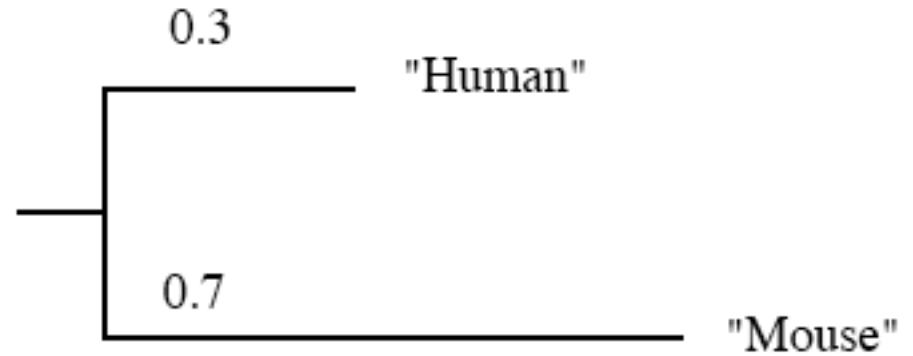
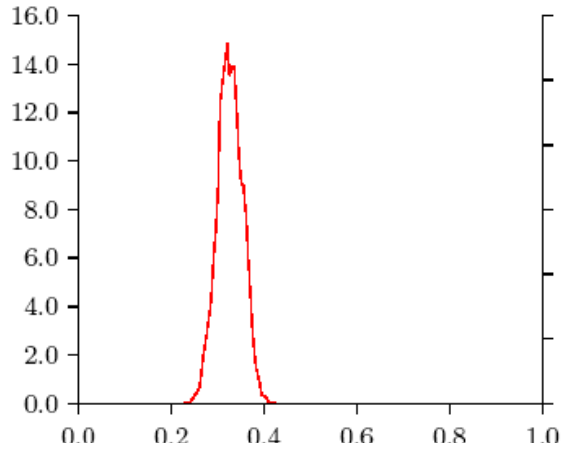
The Pulley Principle



Felsenstein 1981

Irreversibility used for rooting
Ziheng Yang 1994

Irreversibility and rooting

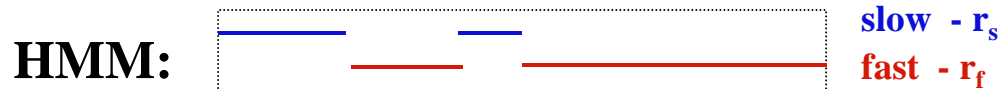
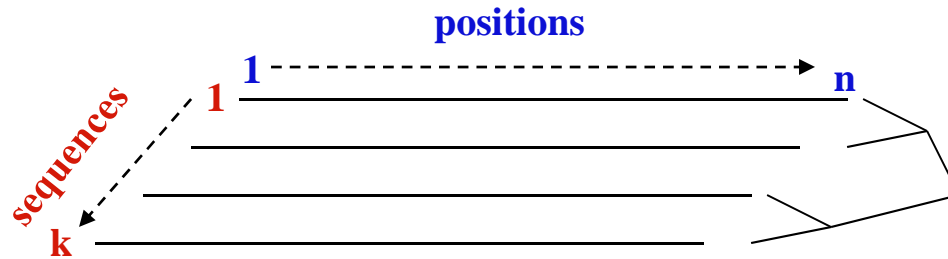


Inferred root positions: chr 21 .484 +/- .014 chr 10 .510 +/- .016

Inferred position 0.33 +/- 0.03, true position 0.3

Fast/Slowly Evolving States

Felsenstein & Churchill, 1996



- π_r - equilibrium distribution of hidden states (rates) at first position
- $p_{i,j}$ - transition probabilities between hidden states
- $L_{(j,r)}$ - likelihood for j'th column given rate r.
- $L^{(j,r)}$ - likelihood for first j columns given j'th column has rate r.

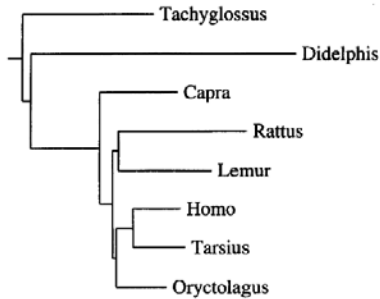
Likelihood Recursions:

$$L^{(j,f)} = (L^{(j-1,f)} p_{f,f} + L^{(j-1,s)} p_{s,f}) L_{(j,f)} \quad L^{(j,s)} = (L^{(j-1,f)} p_{f,s} + L^{(j-1,s)} p_{s,s}) L_{(j,s)}$$

Likelihood Initialisations:

$$L^{(1,f)} = \pi_f L_{(1,f)} \quad L^{(1,s)} = \pi_s L_{(1,s)}$$

Fast-Slow HMM Application



```

1                                60
Tachygloss ATGGTGCAIT TGAGTGGTTC TGAGAAGACT GCTGTGACCA ACCTGTGGGG GCATGTGAAC
Didelphis  .....C.....C.TC.GA G.....AC TGCA.....T. C.A.C.....TC TA.G.....C.G
Tarsius    .....C.....C.C.GA A.....G.C.....TG C.....CA.G.....AG..
Rattus     .....CC.A.C..A.G. G.....G.....T.ATG C.....AA.G.....
Capra      .....C.....C.C.GA G.....G.....C.....G G.T.C.....CA.G.....A
Oryctolag. ....C.....TC.AA.GA G.....T.....G.....TG C.....CA.G.....T
Homo       .....CC.....CC.GA G.....T.....C.....T.TG C.....CA.G.....
Lemur      .....ACTTGC.....C.GA G.....TG.. CA.....T CT.....CA.G.....G.T
1111111122 1111222122 2111111222 2221111222 2111111122 2111111222
111 2 221 2 1111 2 22 2 111 111 2 2
    
```

```

61                                120
Tachygloss GTCAATGAAC TCGGTGGGGA GGCCCTTGGC AGGCTGCTGG TCGTCTACCC CTGGACCCAG
Didelphis  ..T.G.CC.GA CT.....T.....A.....C.....T.....ACC
Tarsius    ..GG.A..TG ..T.....T.....G.....A.....
Rattus     GCTG.....TG ..T.....G.....T.....
Capra      ..GG.A..G T.....CT.....G.....T.....
Oryctolag. ..GG.A..G T.....G.....G.....T.....A.....
Homo       ..GG.....G T.....T.....G.....G.....T.....
Lemur      ..AG.GA..G T.....T.....G.....A.....
2222222222 2111111111 1111111111 1111111111 1211111111 2111111111
2 2 1111 11 1111 1 111 111 11111 1111111 1111111
    
```

```

121                                180
Tachygloss AGGTTCTTCG AATCCTTGGG TGACCTGTCC AGCGCGGATG CTGTGATGGG AAACGCCAAG
Didelphis  .....T..T..GGAG...T.....T.....TCTC.T.GC.....C.....TC.....TT.T...
Tarsius    .....T.....C.....T.....G.....TCTC.T.CC.....T.....A.C.T.T...
Rattus     .....A..T..TAG...T.....G.....TCT.....TC.....A.C.....T.....C.T...
Capra      .....T.....GCA...T.....G.....T.....TCT.....T.....AA.C.T.T...
Oryctolag. ....G.....T.....G.....T.....TCT.AC.....T.....A.C.T.C.T...
Homo       .....T.....G.....T.....G.....T.....TCTC.T.....T.....C.C.T...
Lemur      .....G.....T.....G.....T.....TCTC.TC.....T.....G.....C.T...
1111111111 2222111111 1111111111 2222222211 1111111122 2112211111
1111 11 22211111 1 11111 2 22 1 11 11122 1 22111111
    
```

```

181                                240
Tachygloss GTCAAGGCCG ATGGTGGCAA GGTGCTGACC TCCTTCGGGG ATGCCCTGAA GAACCTGAC
Didelphis  ..TC.A.....T.....T.....T.....A.AG.C.....C.FT.G...
Tarsius    .....C.....CAA.....A.....A.....TA.T.....C.G.A..GC TC.F.G...
Rattus     ..G.....CAAG.....A.A.A.G.....AAT.....G.....AC.T.G...
Capra      ..G.....CAAG.....AGA.....TA.TA.C.G.A.....C.T.T...
Oryctolag. ..G.....T.....CAAG.....G.T.G.....A.T.....G.GT.....TC.....G...
Homo       ..G.....T.....CAAG.....A.....CGGT G.....TA.T.....G.....GC TC.....G...
Lemur      ..G.....T.....CAAG.....GT G.....TA.T.....A.GT.....C.TC.....G...
1111111111 1111122211 1111222222 2111122211 1222221122 2222211111
111 1 111 11 1111 222 111 1 2 2 111
    
```

```

241                                300
Tachygloss AACCTCAAGG GAACCTTGGC CAAGCTGAGC GAACCTGACT GCGACAAGCT GCACCTGGAC
Didelphis  .....G.....T..T.AT.....T.....T.....G.C.....T.....T.....
Tarsius    .....C.....T.....T.....T.....T.....G.....AT.....T.....
Rattus     .....C.....T.....TC.T.....T.....C.....T.....T.....
Capra      G.....T.....TC.....T.....G.....T.....T.....
Oryctolag. ....A.....C.....T.....T.....T.....T.....
Homo       .....C.....T.....CA.....T.....G.....T.....T.....
Lemur      .....C.....T.....TC.A.....T.....G.....T.....T.....
1111111111 1111111111 2222111111 1111111111 1111111111 1111111111
1111111111 1111 11 1111 1 1 1111 111111 1 111 111111
    
```

```

301                                360
Tachygloss CCGCAAGATT TCAATGCCCT GGGTAACGTG CTGGTCTGGG TCGTGGCCGG TCACTGGAC
Didelphis  ..T.....C.....GATG.....G.TA.C.A.T..GA.CT G.....TGA G.....TG...
Tarsius    ..T.....C.....GG.T.T.....C.....T.....GT.....G.....A.C.....TG...
Rattus     ..T.....C.....GG.T.....C.....TA..A.T..GA.T..GT.....G.A.C.C.GG...
Capra      ..T.....C.....G.T.....C.....TA.....G.T.....G.....T.....C.....CATG...
Oryctolag. ..T.....C.....GG.T.....C.....T.....T.....G.....T.T.A.....TG...
Homo       ..T.....C.....GG.T.....C.....T.....TGT.....G.....A.....TG...
Lemur      ..TC.....C.....C.T.....C.....G.....T.....G.....TA.A.....TG...
1111111111 1112222111 1111112211 1111122222 2111122222 2111111111
111 111 1 11 1 1 11 1 1 22 11 222 21 11
    
```

```

361                                420
Tachygloss AAGGAATTCA CCCCAGGCGC CAGGCTGCC TGGCAGAAGC TGGTGTCTGG TGTTTCCAC
Didelphis  .....T..T..T..T..T..ATG T.....T.....T.....C.....G.....A.....G...T
Tarsius    ..A.....T.....GC.....T T.....AT.....G.....G.....GG.TACT
Rattus     .....TGT..A.....TC.....G.....G.....A.GG..AGT
Capra      ..GT.....GCT.CT G.....AG ..T.....G.....G.....G.....A.T
Oryctolag. ..A.....T.....TC.T G.....AT.....G.....G.....G.....GG.A.T
Homo       ..A.....T.....ACCA.T G.....AT.....G.....G.....GG.TA.T
Lemur      ..T.C.....G.....G.C.T G.....TT.....AG.....G.....GG.A.T
1221111111 1111222211 2111111111 1221111111 1111111111 1111111111
1 11111 222222 111 1 1111 11111 111 1 1
    
```

```

421                                484
Tachygloss GCCCTGGCCC ACAAGTACCA CTGA
Didelphis  .....A.....
Tarsius    .....T.....
Rattus     .....T.....A.....
Capra      .....GA.T.....A.....
Oryctolag. ....T.....A.....
Homo       .....T.....A.....
Lemur      .....T.....T.....
1111111111 1111111111 1111111111 1121
111 111 1 111 1 11 1 21
    
```

Probabilities of different paths

A → → *T*

- *What are the number of events?*
- *What are the kinds of events?*

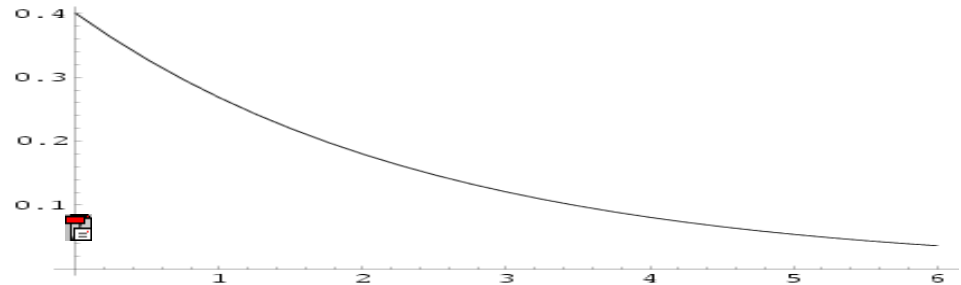
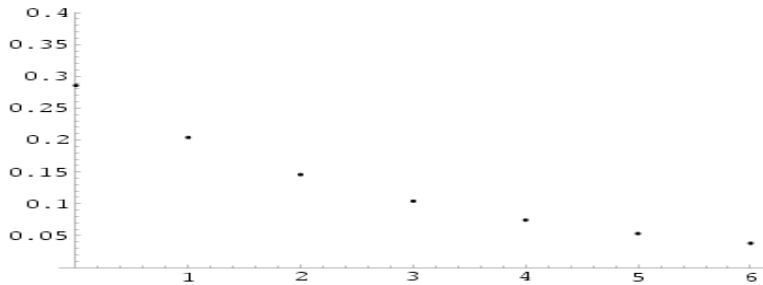
Geometric/Exponential Distributions

The Geometric Distribution: $\{1, \dots\}$ $\text{Geo}(p)$:

$$P\{Z=j\}=p^j(1-p) \quad P\{Z>j\}=p^j \quad E(Z)=1/p.$$

The Exponential Distribution: \mathbb{R}^+ $\text{Exp}(a)$

$$\text{Density: } f(t) = ae^{-at}, \quad P(X>t) = e^{-at}$$



Properties: $X \sim \text{Exp}(a)$ $Y \sim \text{Exp}(b)$ independent

- i. $P(X>t_2|X>t_1) = P(X>t_2-t_1) \quad (t_2 > t_1)$
- ii. $E(X) = 1/a.$
- iii. $P(Z>t)=(\approx)P(X>t)$ small a ($p=e^{-a}$).
- iv. $P(X < Y) = a/(a + b).$
- v. $\min(X, Y) \sim \text{Exp}(a + b).$