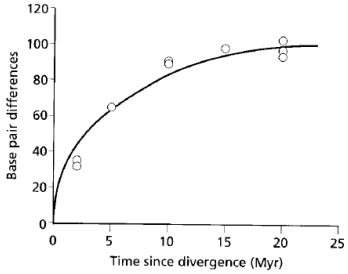
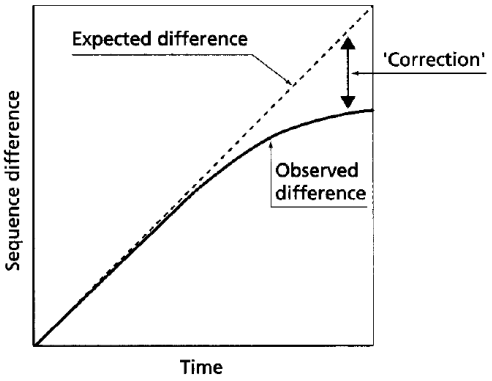
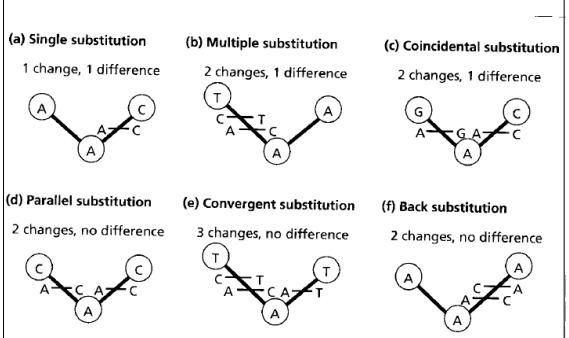
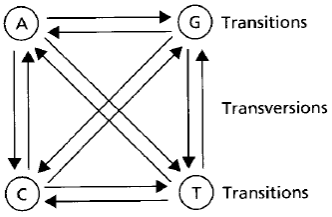
Models of DNA substitutions
Rate matrices and choosing the appropriate model

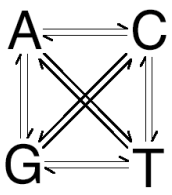








Continuous-Time Markov Chain



Simulation example from John Huelsenbeck

$$\mathbf{P}_t = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

The substitution probability matrix can be computed or determined using simulations.

Example of determining transition probabilities using simulation with the following rate matrix:

$$Q = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

From \ To	A	C	G	T
A	-0.886	0.190	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.190	-1.519	0.063
T	0.253	0.949	0.127	-1.329

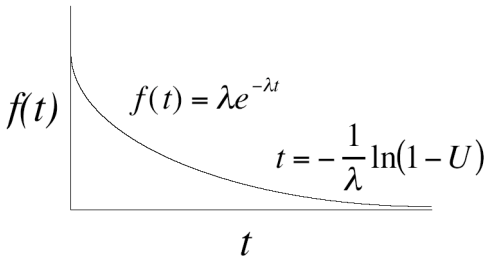
From \ To	A	C	G	T
A	-0.886	0.190	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.190	-1.519	0.063
T	0.253	0.949	0.127	-1.329

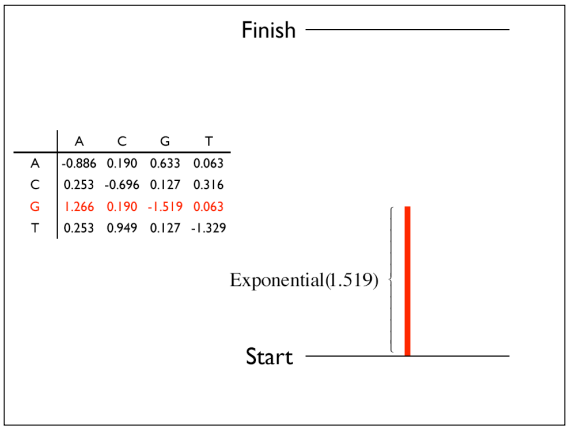
Interpretation: If the process is in state *i*, we wait an exponentially distributed amount of time with parameter $-q_{ii}$ until the next substitution occurs.

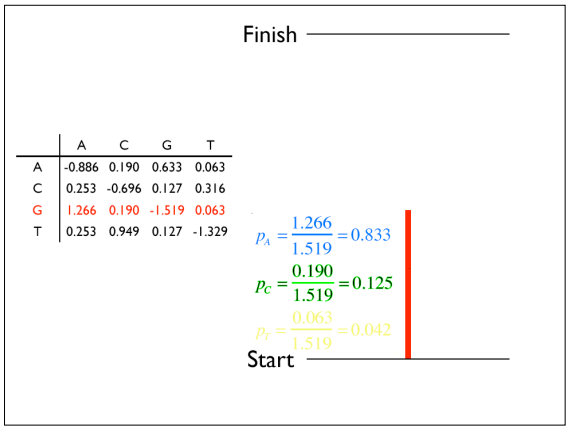
From \ To	A	C	G	T
A	-0.886	0.190	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.190	-1.519	0.063
T	0.253	0.949	0.127	-1.329

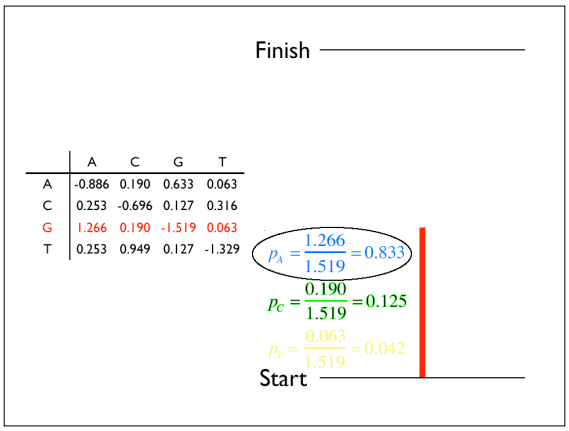
Interpretation: The change is to state *j* with probability $-q_j / q_{ii}$.

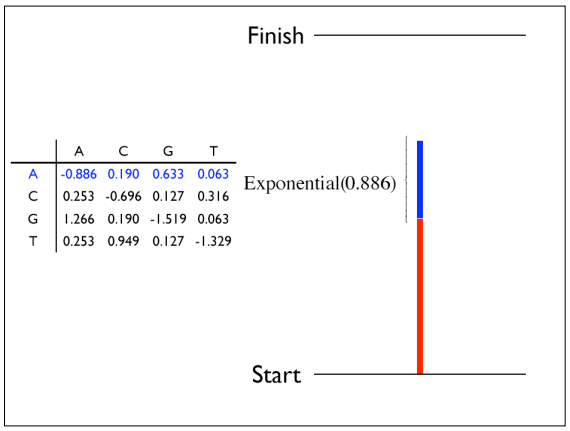
Exponential distribution

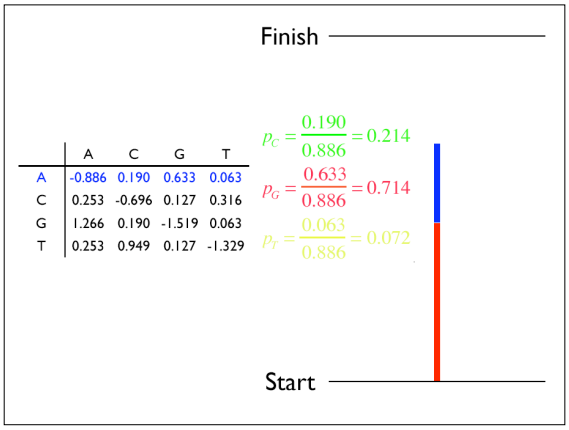


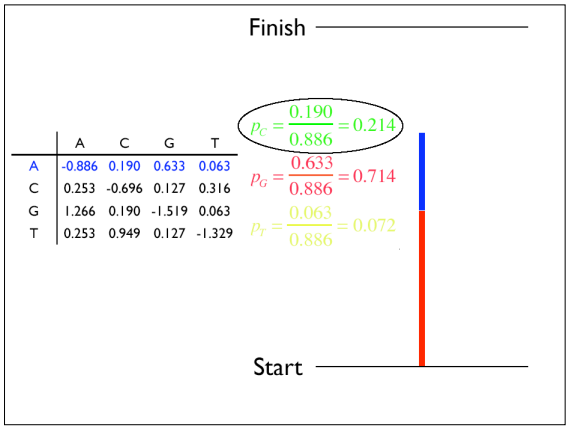

Finish
A C G T
A -0.886 0.190 0.633 0.063
C 0.253 -0.696 0.127 0.316
G 1.266 0.190 -1.519 0.063
T 0.253 0.949 0.127 -1.329
Exponential(1.519)
Start

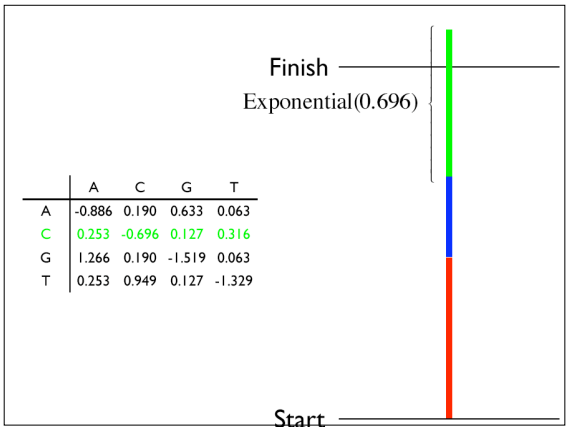

Finish
A C G T
A -0.886 0.190 0.633 0.063
C 0.253 -0.696 0.127 0.316
G 1.266 0.190 -1.519 0.063
T 0.253 0.949 0.127 -1.329
$p_A = \frac{1.266}{1.519} = 0.833$
$p_C = \frac{0.190}{1.519} = 0.125$
$p_T = \frac{0.063}{1.519} = 0.042$
Start


Finish
A C G T
A -0.886 0.190 0.633 0.063
C 0.253 -0.696 0.127 0.316
G 1.266 0.190 -1.519 0.063
T 0.253 0.949 0.127 -1.329
$p_A = \frac{1.266}{1.519} = 0.833$
$p_C = \frac{0.190}{1.519} = 0.125$
$p_T = \frac{0.063}{1.519} = 0.042$
Start

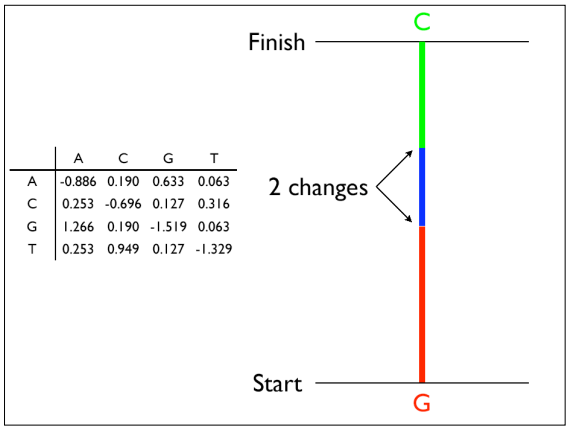

Finish
A C G T
A -0.886 0.190 0.633 0.063
C 0.253 -0.696 0.127 0.316
G 1.266 0.190 -1.519 0.063
T 0.253 0.949 0.127 -1.329
Exponential(0.886)
Start


Finish
A C G T
A -0.886 0.190 0.633 0.063
C 0.253 -0.696 0.127 0.316
G 1.266 0.190 -1.519 0.063
T 0.253 0.949 0.127 -1.329
$p_C = \frac{0.190}{0.886} = 0.214$
$p_G = \frac{0.633}{0.886} = 0.714$
$p_T = \frac{0.063}{0.886} = 0.072$
Start

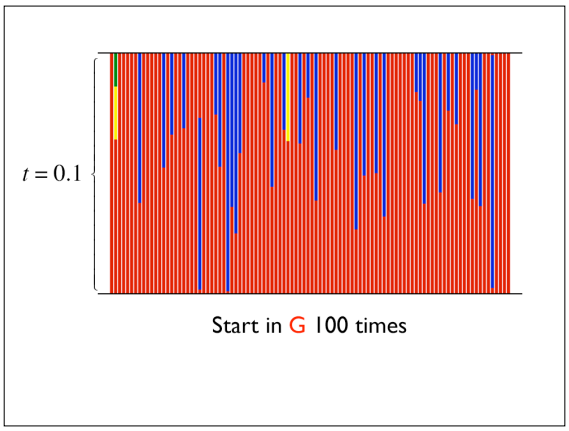

Finish
A C G T
A -0.886 0.190 0.633 0.063
C 0.253 -0.696 0.127 0.316
G 1.266 0.190 -1.519 0.063
T 0.253 0.949 0.127 -1.329
$p_C = \frac{0.190}{0.886} = 0.214$
$p_G = \frac{0.633}{0.886} = 0.714$
$p_T = \frac{0.063}{0.886} = 0.072$
Start



	A	C	G	T
A	-0.886	0.190	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.190	-1.519	0.063
T	0.253	0.949	0.127	-1.329

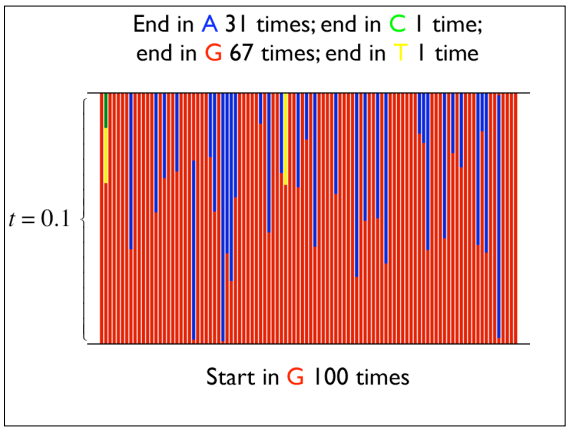


	A	C	G	T
A	-0.886	0.190	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.190	-1.519	0.063
T	0.253	0.949	0.127	-1.329



Start in G 100 times

End in A 31 times; end in C 1 time;
end in G 67 times; end in T 1 time



Start in G 100 times

Transition probabilities for t = 0.1

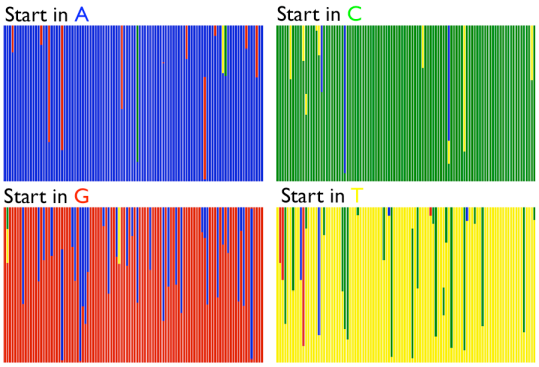
Started In	Ended In: A	C	G	T
A				
C				
G	0.31	0.01	0.67	0.01
T				

(Monte Carlo estimates of transition probabilities based on a total of 100 simulations)

Transition probabilities for t = 0.1

Started In	Ended In: A	C	G	T
A				
C				
G	0.1125	0.0182	0.8634	0.0058
T				

(Monte Carlo estimates of transition probabilies based on a total of 50,000 simulations)



Transition probabilities for t = 0.1

Started In \ Ended In	A	C	G	T
A	0.88	0.02	0.09	0.01
C	0.03	0.90	0.00	0.07
G	0.31	0.01	0.67	0.01
T	0.04	0.20	0.04	0.72

(Monte Carlo estimates of transition probabilies based on a total of 100 simulations)

Transition probabilities for t = 0.1

Started In \ Ended In	A	C	G	T
A	0.9180	0.0182	0.0577	0.0060
C	0.0249	0.9346	0.0125	0.0279
G	0.1125	0.0182	0.8634	0.0058
T	0.0241	0.0877	0.0113	0.8767

(Monte Carlo estimates of transition probabilies based on a total of 50,000 simulations)

Transition probabilities for any rate matrix, Q, can be calculated as

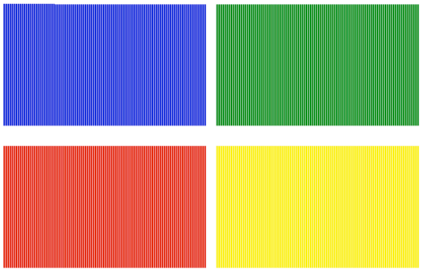
$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Monte Carlo (50,000 reps)

Started In \ Ended In	A	C	G	T
A	0.9180	0.0182	0.0577	0.0060
C	0.0249	0.9346	0.0125	0.0279
G	0.1125	0.0182	0.8634	0.0058
T	0.0241	0.0877	0.0113	0.8767

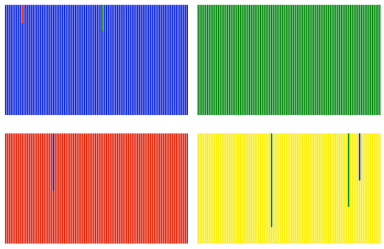
Exact: $\mathbf{P}(t) = e^{\mathbf{Q}t}$

Started In \ Ended In	A	C	G	T
A	0.9191	0.0184	0.0563	0.0061
C	0.0245	0.9344	0.0123	0.0287
G	0.1127	0.0183	0.8627	0.0061
T	0.0245	0.0862	0.0123	0.8770



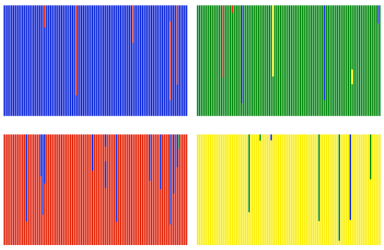
$\mathbf{P}(0.00) =$

	A	C	G	T
A	1.0000	0.0000	0.0000	0.0000
C	0.0000	1.0000	0.0000	0.0000
G	0.0000	0.0000	1.0000	0.0000
T	0.0000	0.0000	0.0000	1.0000



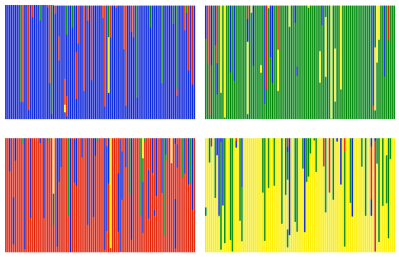
P(0.01) =

	A	C	G	T
A	0.9912	0.0019	0.0062	0.0006
C	0.0025	0.9931	0.0013	0.0031
G	0.0125	0.0019	0.9849	0.0006
T	0.0025	0.0094	0.0013	0.9868



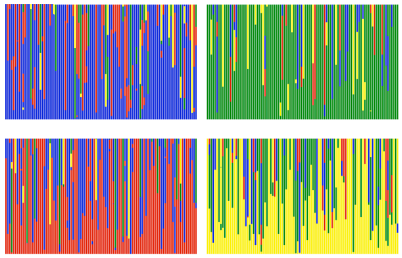
P(0.10) =

	A	C	G	T
A	0.9191	0.0183	0.0563	0.0061
C	0.0243	0.9344	0.0122	0.0287
G	0.1127	0.0184	0.8627	0.0061
T	0.0245	0.0861	0.0122	0.8770



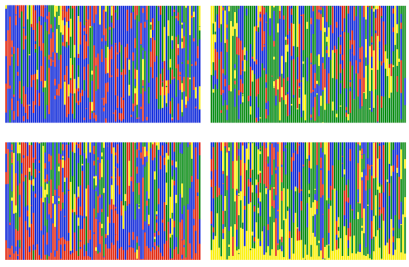
P(0.50) =

	A	C	G	T
A	0.7079	0.0813	0.1835	0.0271
C	0.1085	0.7377	0.0542	0.0995
G	0.3670	0.0813	0.5244	0.0271
T	0.1085	0.2985	0.0542	0.5387



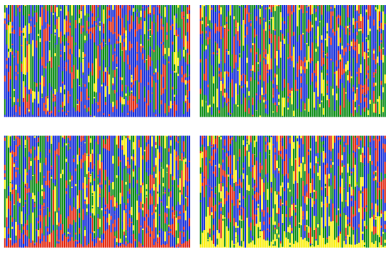
P(1.00) =

	A	C	G	T
A	0.5803	0.1406	0.2320	0.0468
C	0.1875	0.5871	0.0937	0.1314
G	0.4641	0.1406	0.3483	0.0468
T	0.1875	0.3942	0.0937	0.3243



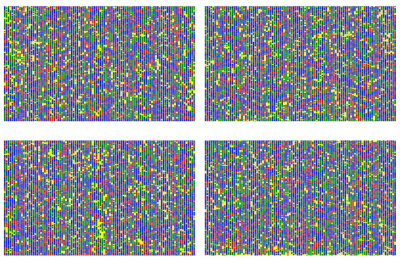
P(5.00) =

	A	C	G	T
A	0.4113	0.2873	0.2056	0.0957
C	0.3831	0.3190	0.1915	0.1062
G	0.4112	0.2873	0.2056	0.0957
T	0.3831	0.3188	0.1915	0.1065



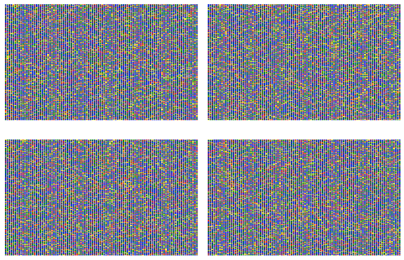
P(10.00) =

	A	C	G	T
A	0.4005	0.2994	0.2002	0.0998
C	0.3992	0.3008	0.1996	0.1002
G	0.4005	0.2994	0.2002	0.0998
T	0.3992	0.3008	0.1996	0.1002



$\mathbf{P}(100.00) =$

	A	C	G	T
A	0.4000	0.3000	0.2000	0.1000
C	0.4000	0.3000	0.2000	0.1000
G	0.4000	0.3000	0.2000	0.1000
T	0.4000	0.3000	0.2000	0.1000



$\mathbf{P}(1000.00) =$

	A	C	G	T
A	0.4000	0.3000	0.2000	0.1000
C	0.4000	0.3000	0.2000	0.1000
G	0.4000	0.3000	0.2000	0.1000
T	0.4000	0.3000	0.2000	0.1000

Stationary probabilities (also called equilibrium frequencies, prior probabilities) of the states are the probability of finding the Markov chain in the state after an infinite amount of time.

$$\pi_A = 0.4$$
$$\pi_C = 0.3$$
$$\pi_G = 0.2$$
$$\pi_T = 0.1$$

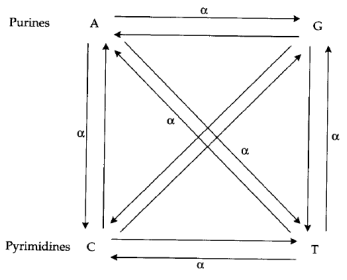
$$Q = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix}\mu$$

$\kappa = 5$
$\pi_A = 0.4$
$\pi_A = 0.3$
$\pi_A = 0.2$
$\pi_A = 0.1$

$$Q = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

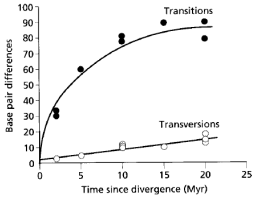
$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \quad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}]$$

Jukes Cantor model:
D = -3/4ln(1-4/3p)



But transitions and transversions have different rates
Ratio is often called k.

Type of sequences	Transition/transversion ratio (κ)
mtDNA	9.0
12S rRNA	1.75
α- and β-globins	0.66
Pseudo η-globin	2.70



$$\mathbf{P}_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}^t \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

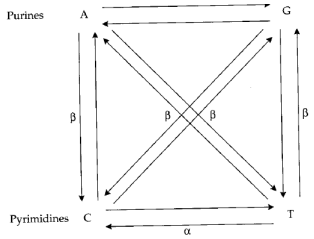
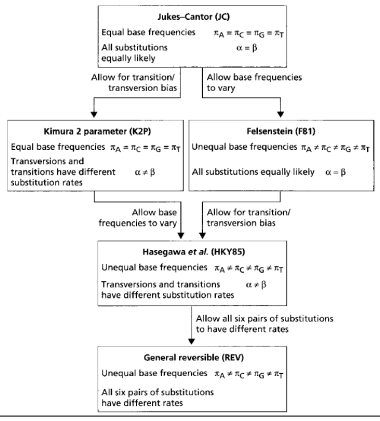


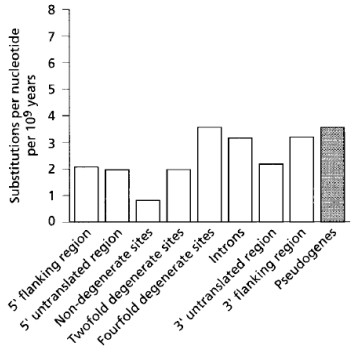
FIGURE 3.4 Two-parameter model of nucleotide substitution. The rate of transition (α) may not be equal to the rate of transversion (β).



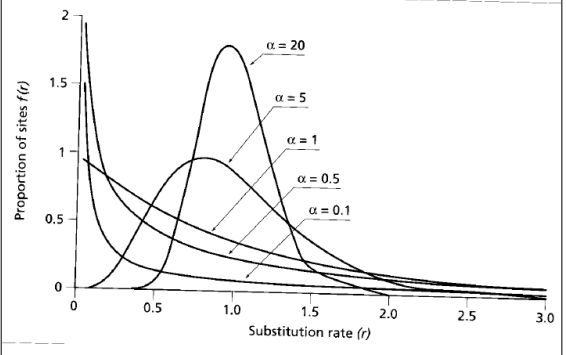
What assumptions did we make?

- All nucleotides change independently
- The subsituion rate is constant over time and in different lineages
- The base composition is at equilibrium
- The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time..

Variation in rates between sites is common



Model variation in sites with gamma parameter



Type of sequences	α
Nuclear genes	
Albumin genes	1.05
Insulin genes	0.40
c-myc genes	0.47
Prolactin genes	1.37
16S-like rRNAs, stem region	0.29
16S-like rRNAs, loop region	0.58
ψη-globin pseudogenes	0.66
Viral genes	
Hepatitis B virus genomes	0.26
Mitochondrial genes	
12S rRNAs	0.16
Position 1 of four genes	0.18
Position 2 of four genes	0.08
Position 3 of four genes	1.58
D-loop region	0.17
Cytochrome *b*	0.44

How do we choose an appropriate model?

- Calculate likelihood of model given data
- Compare likelihoods of nested models

What is maximum likelihood?

Comparison to probability theory:

Probability of # heads in 5 coin tosses

Heads	Prob.
0	.03
1	.16
2	.31
3	.31
4	.16
5	.03

$P(x) = (n!/(n-x)!)p^x q^{n-x}$

Same calculation for coins with different bias.

	Bias of coin towards Heads				
Heads	.1	.3	.5	.7	.9
0	.59	.17	.03	.00	.00
1	.33	.36	.16	.03	.00
2	.07	.31	.31	.13	.01
3	.01	.13	.31	.31	.07
4	.00	.03	.16	.36	.33
5	.00	.00	.03	.17	.59

One way to get the likelihood is to estimate them using Makov Chain Monte Carlo methods.

-analogy to walking up hill.

Markov Chain Monte Carlo

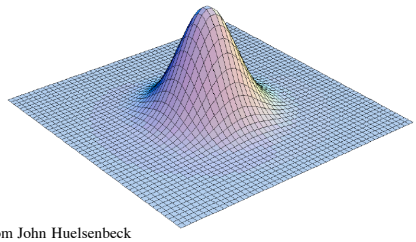
- Start with proposed state
- Perturb old state and calculate probability of new state
- Test if new state is better than old state, accept if ratio of new to old is greater than a randomly drawn number between 0 and 1.
- Move to new state if accepted, if not stay at old state
- Start over

Caveats: The proposal mechanism is at the discretion of the programmer, but must satisfy a few basic requirements: all states must be reachable, the chain must be aperiodic, and the mechanism must be stochastic.

Markov chain Monte Carlo

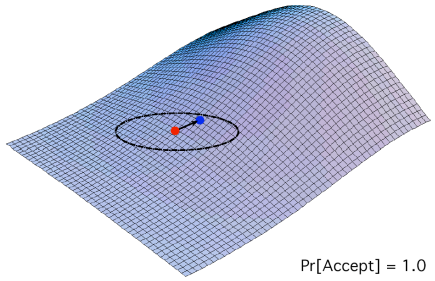
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing Machines. *J. Chem. Phys.* 21:1087–1091.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

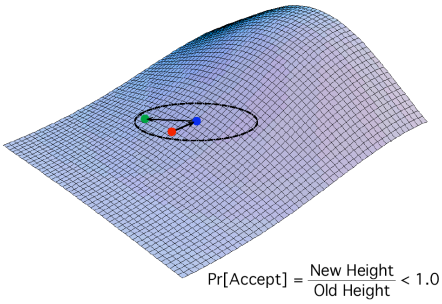


-Graphs from John Huelsenbeck

Circle represents amount of potential proposed change.

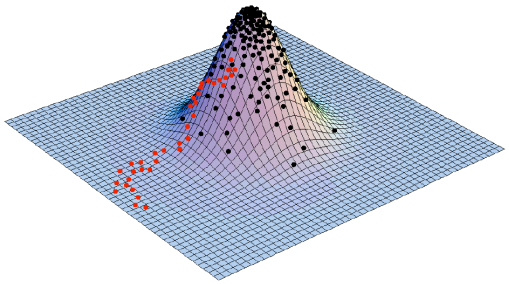


Pr[Accept] = 1.0



$$\Pr[\text{Accept}] = \frac{\text{New Height}}{\text{Old Height}} < 1.0$$

Red dots = "burn in" period.



Likelihood ratio test statistic:

$$\Delta l = \log\left(\frac{\max\{L(\text{neutral model})\}}{\max\{L(\text{selection model})\}}\right)$$

$$= \log(\max\{L(\text{neutral model})\}) - \log(\max\{L(\text{selection model})\})$$

$-2\Delta l$ approximates χ^2 with n degrees freedom, where n is the difference in number of parameters between the nested models.

Test models using ModelTest:

Posada and Crandall 2001, *Syst. Biol.* **50**: 580-601