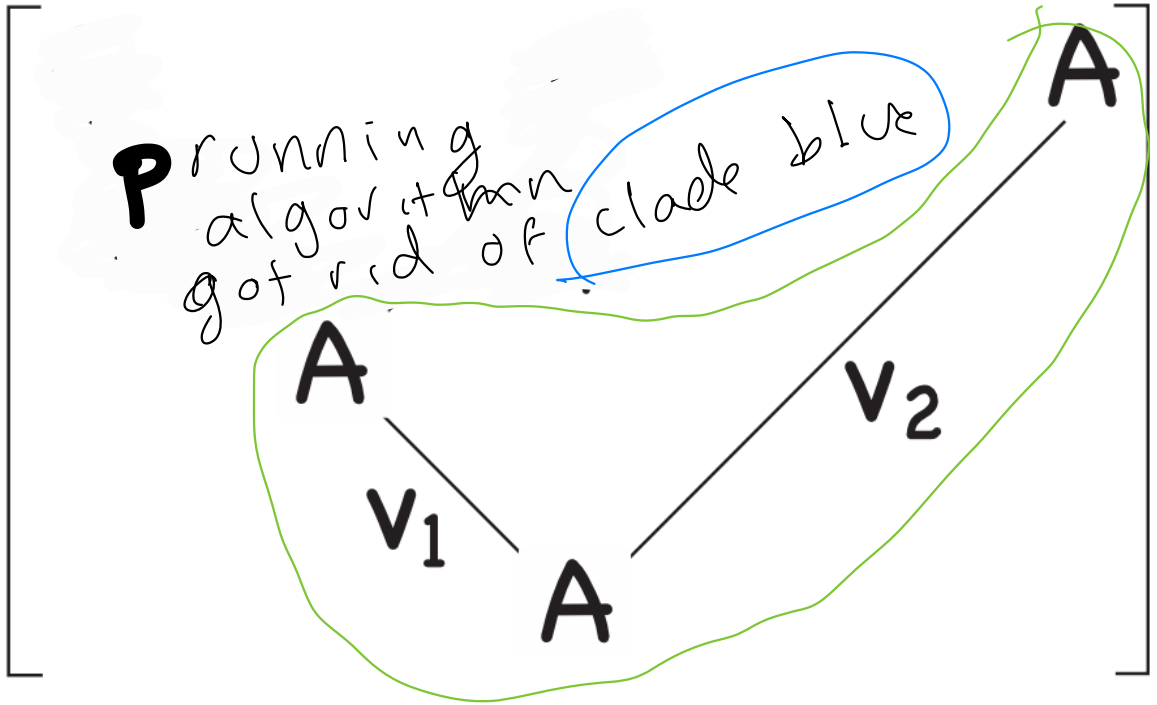
More intuition on phylogenetic tree estimation

Simplification from John Huelsenbeck's lecture
Woods Hole Molecular Phylogenetics

$$\Pr\Big[$$



$$\Big]$$

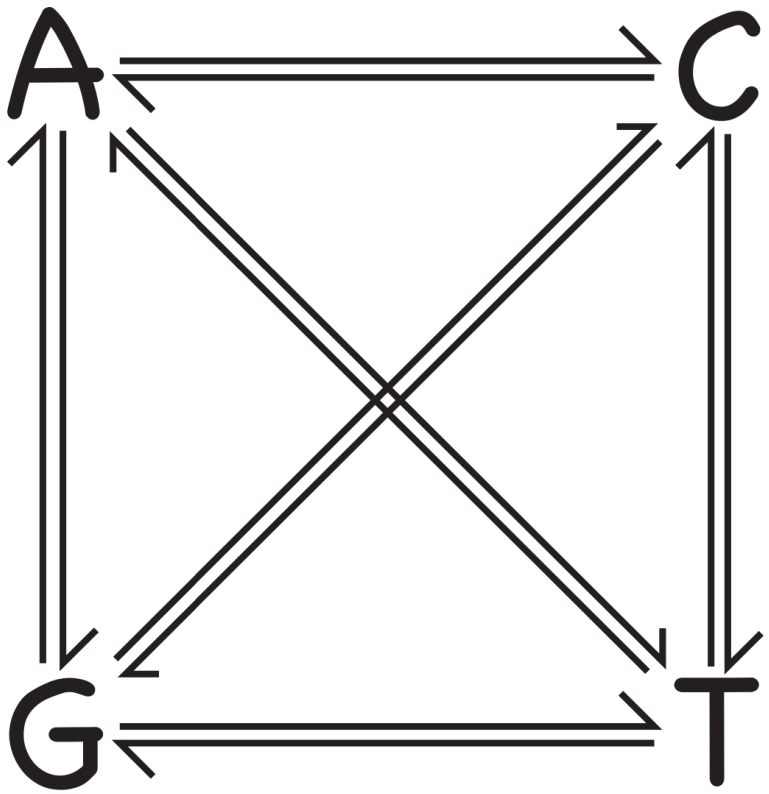
$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

Pr[]

$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

↳ stationary distribution

What is truly going on?



From \ To	A	C	G	T
A	-0.886	0.19	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.19	-1.519	0.063
T	0.253	0.949	0.127	-1.329

$= Q$

How does Q relate to probability

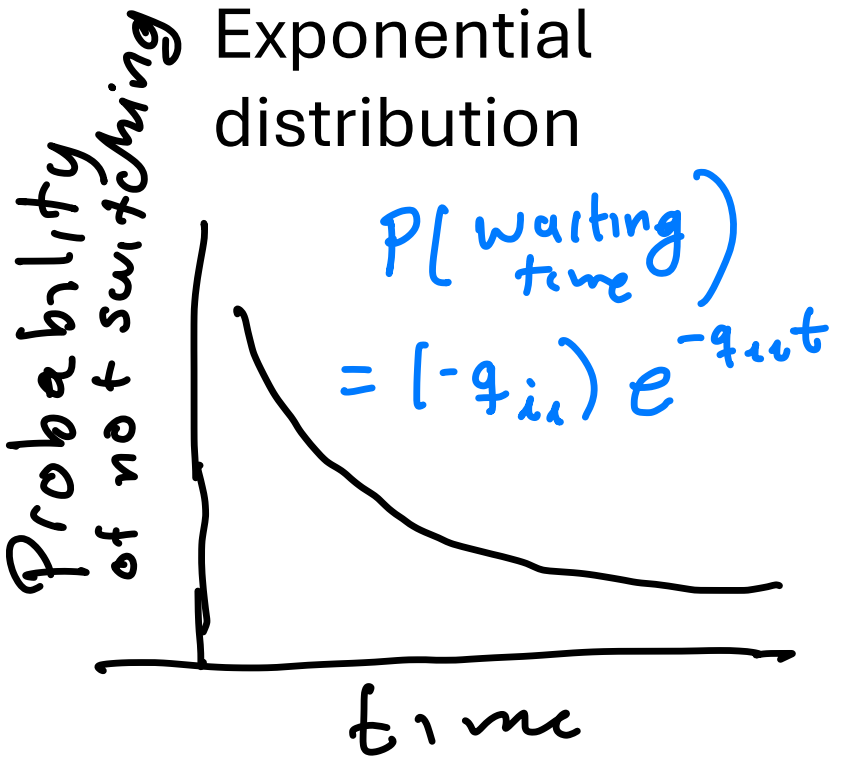
$P = e^{Qt}$ → molecular time (call those v's in previous example)

$P(X=x) = \lambda e^{-\lambda x}$ in wiki

$E[X] = \frac{1}{\lambda}$

From \ To	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−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.



Probability of not switching in a short time is large but in a long time is small

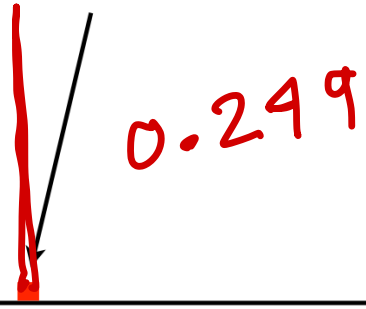
$\sim$ (distributed as) $\mathrm{Exp}(-q_{ii})$ $i = A, C, G, T$

Finish

	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

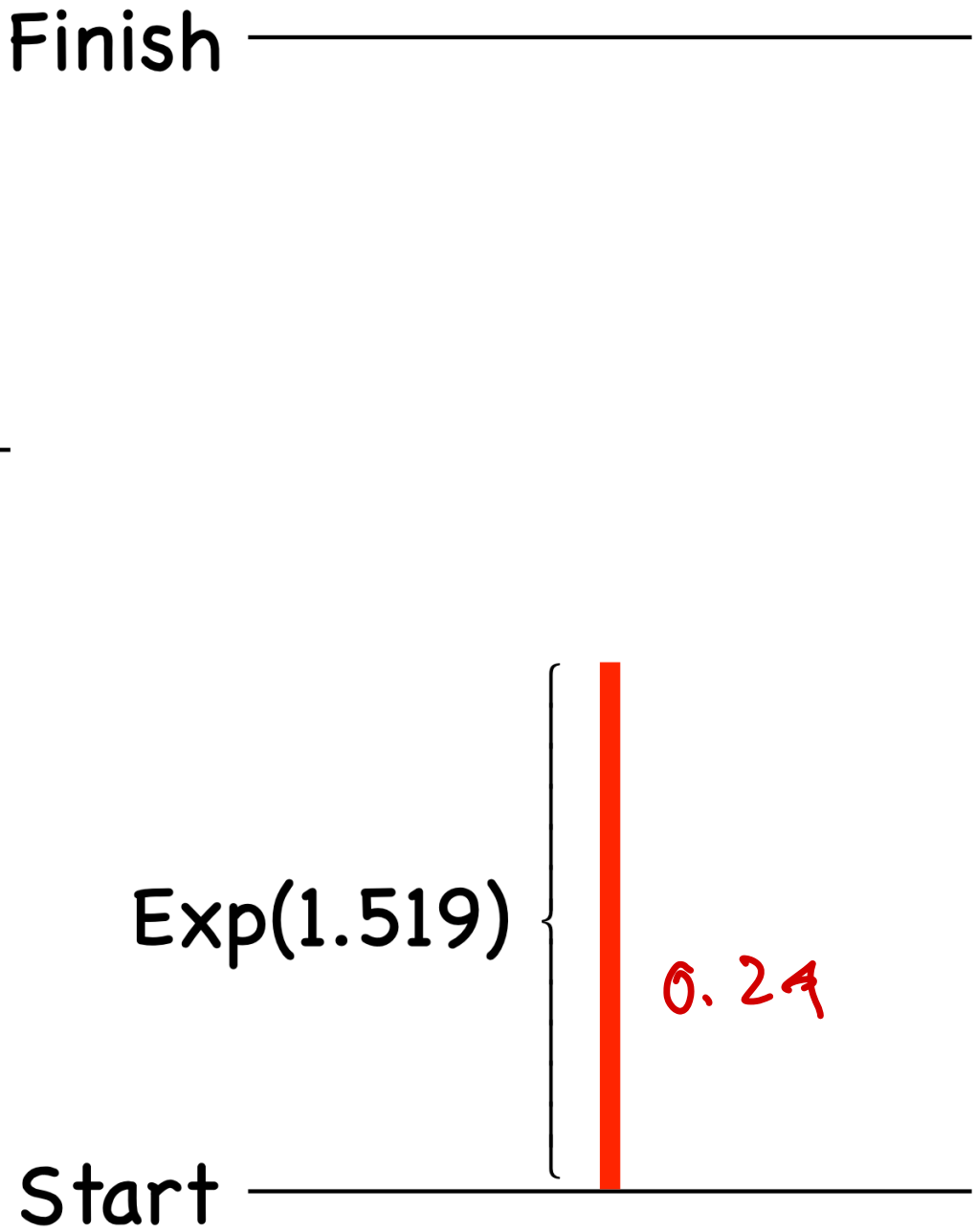
1.011
0.249
0.22
0.04

Start in state G



Start

	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

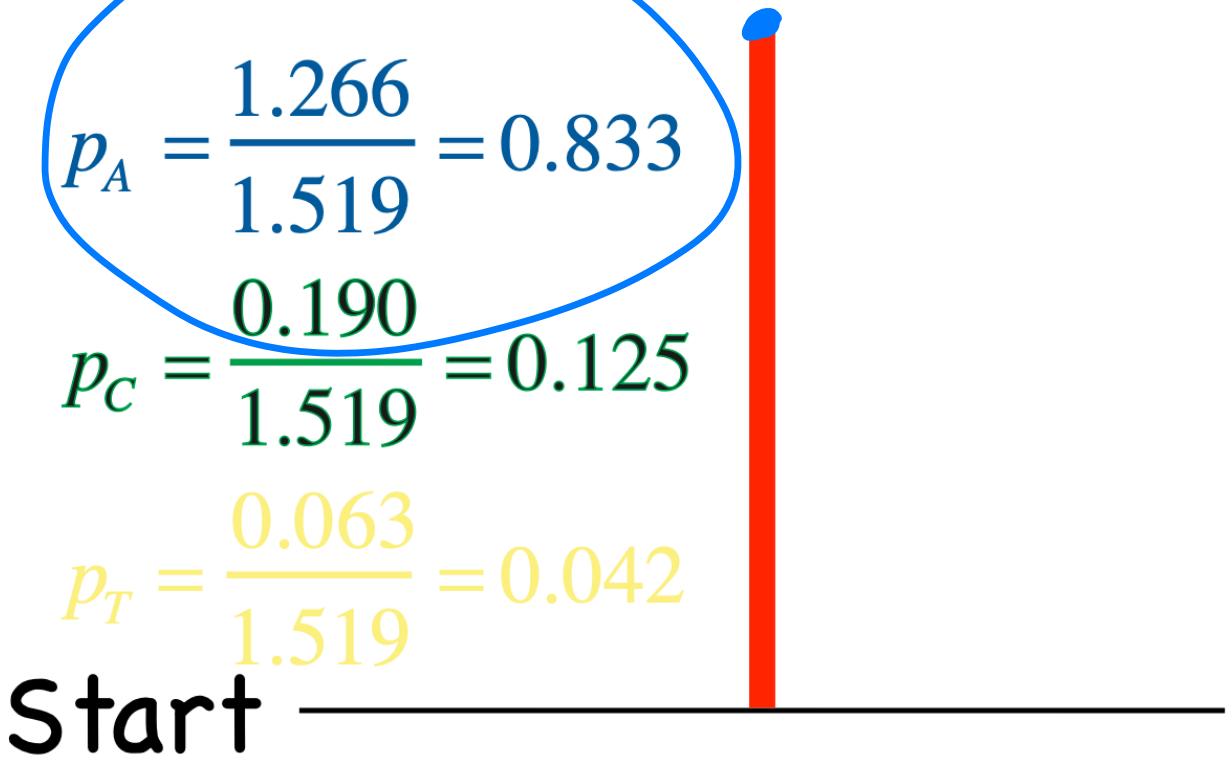


	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

$$p_A = \frac{1.266}{1.519} = 0.833$$

$$p_C = \frac{0.190}{1.519} = 0.125$$

$$p_T = \frac{0.063}{1.519} = 0.042$$

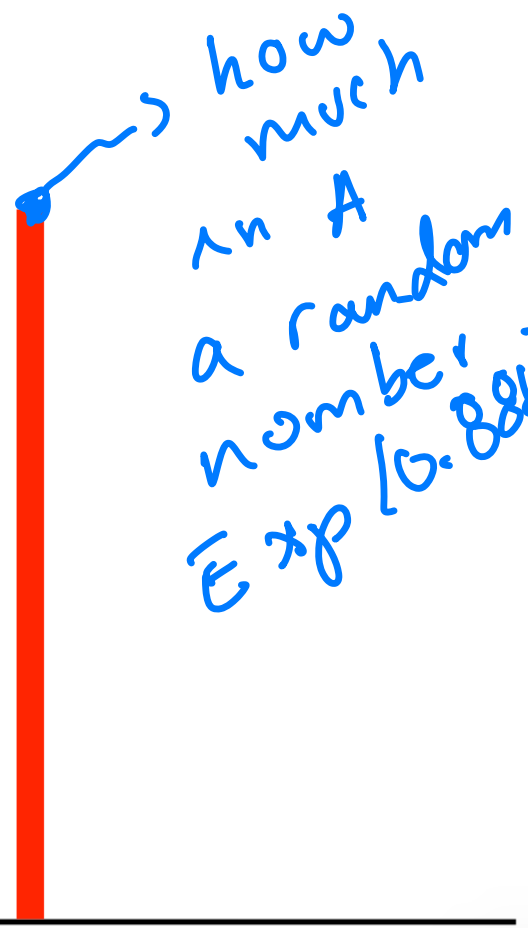


	A	C	G	T
A	-0.886	0.19	0.633	0.063
C	0.253	-0.696	0.127	0.316
G	1.266	0.19	-1.519	0.063
T	0.253	0.949	0.127	-1.329

$$p_A = \frac{1.266}{1.519} = 0.833$$

$$p_C = \frac{0.190}{1.519} = 0.125$$

$$p_T = \frac{0.063}{1.519} = 0.042$$



Finish

	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

Exp(0.886)

Start

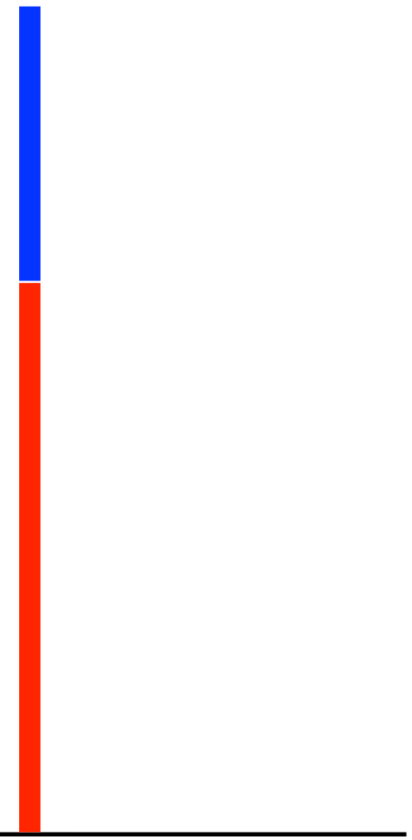
Finish

	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

$$p_C = \frac{0.190}{0.886} = 0.214$$

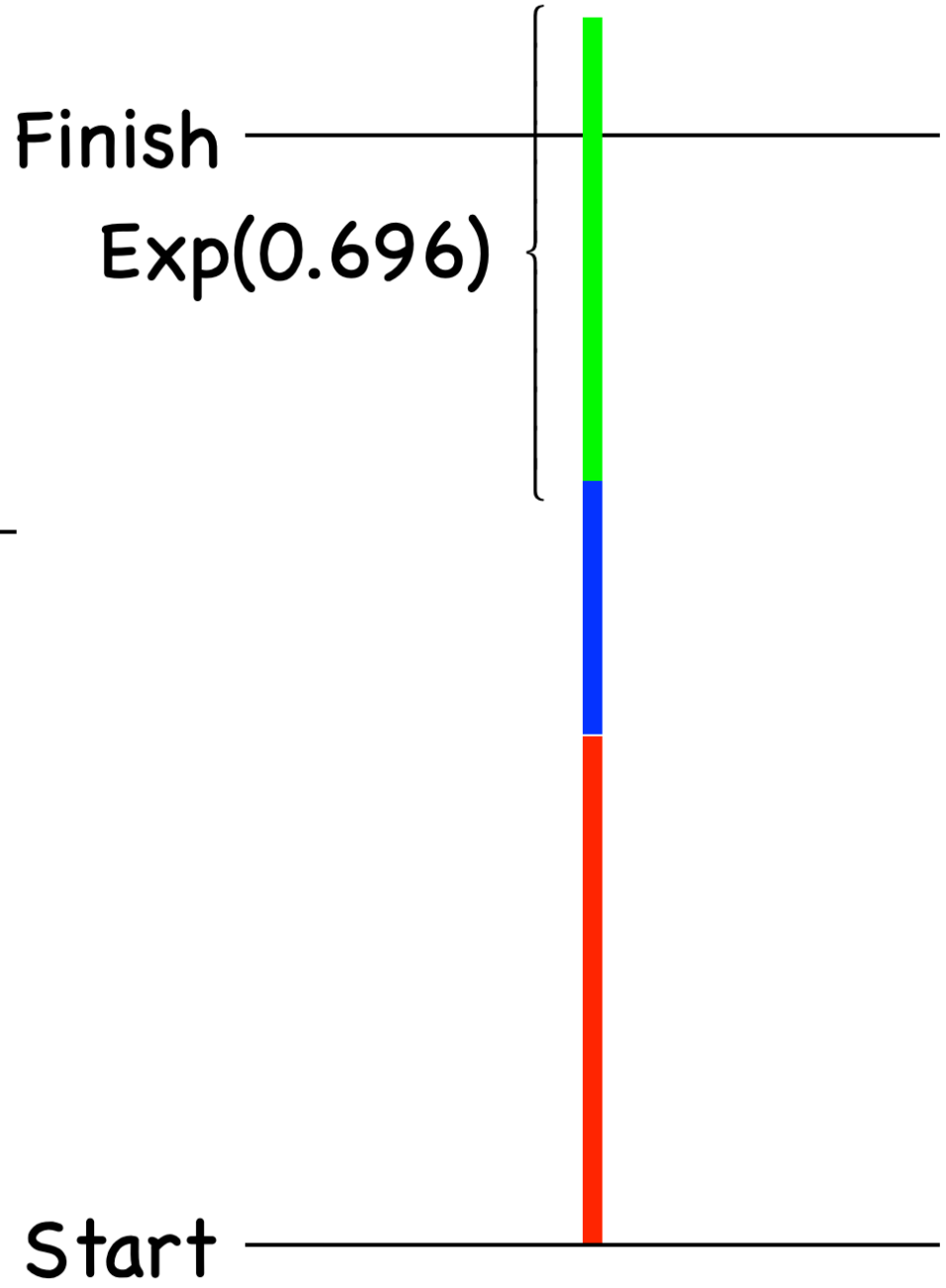
$$p_G = \frac{0.633}{0.886} = 0.714$$

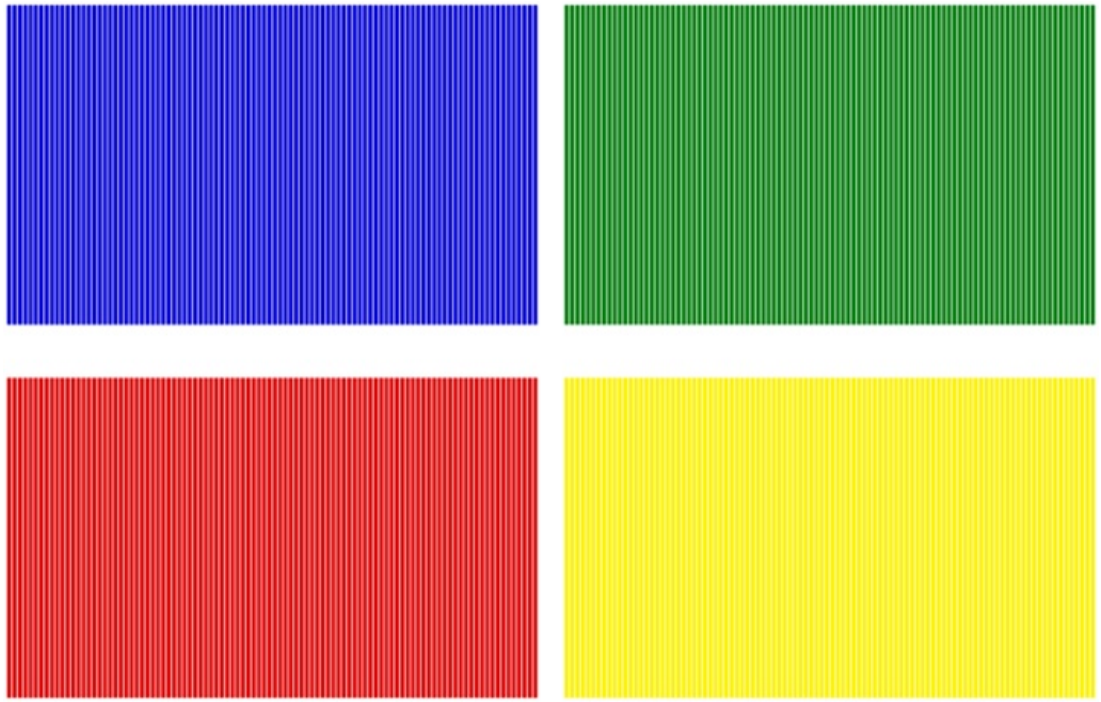
$$p_T = \frac{0.063}{0.886} = 0.072$$



Start

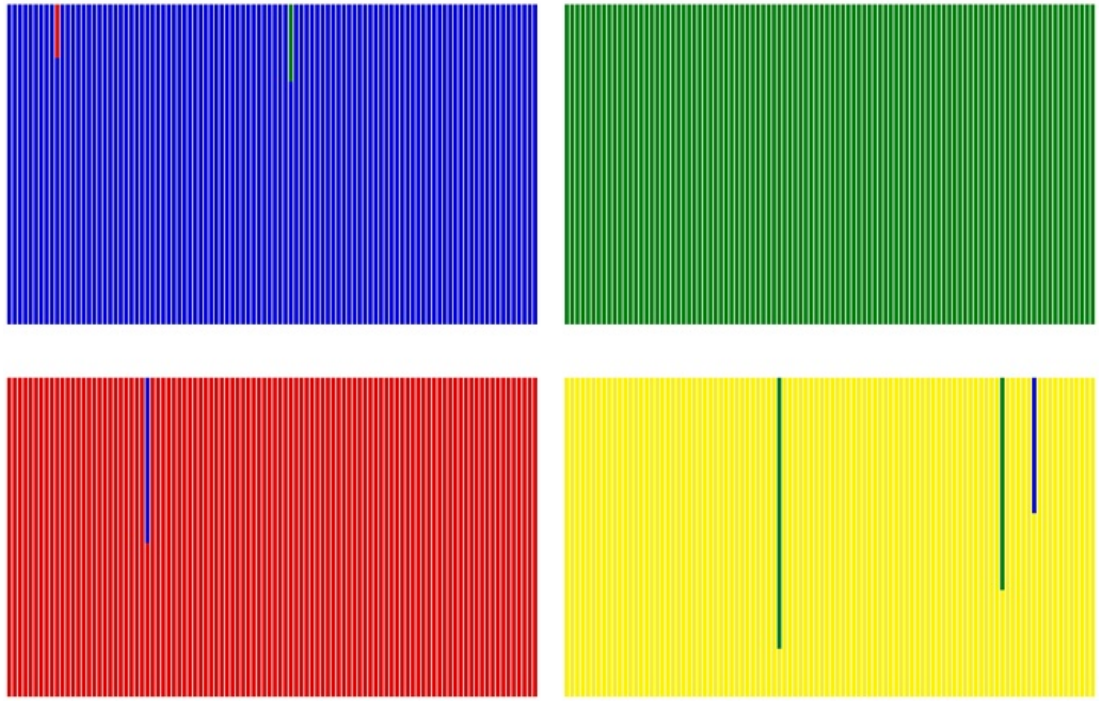
	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329





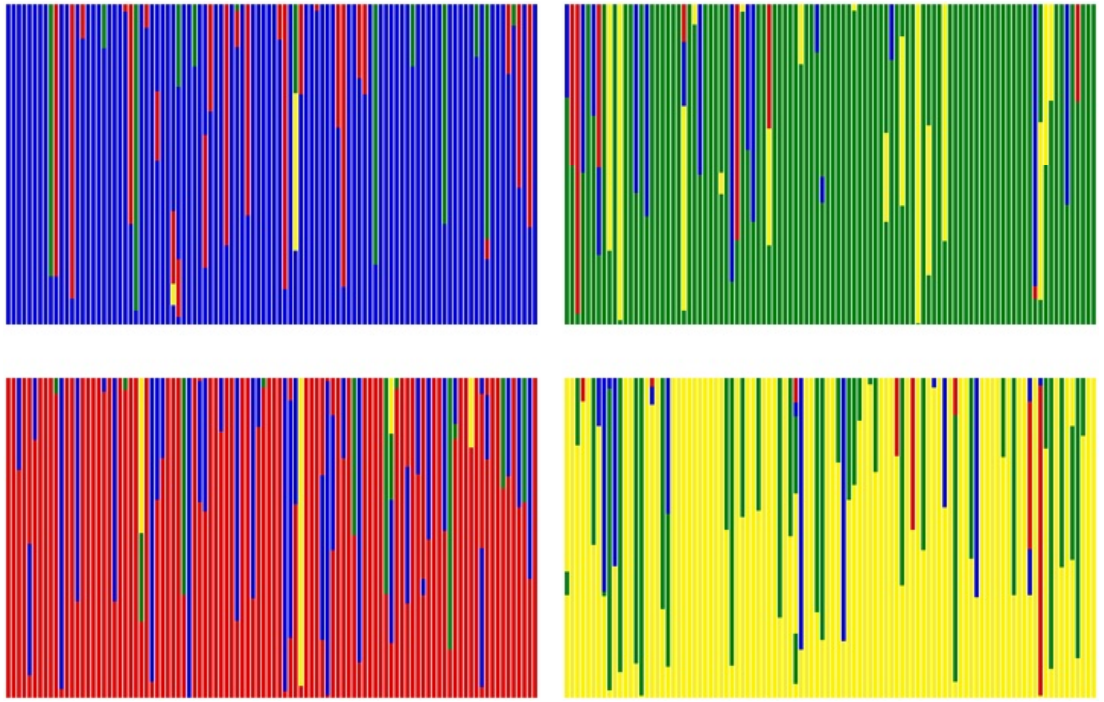
P(0.00) =

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1



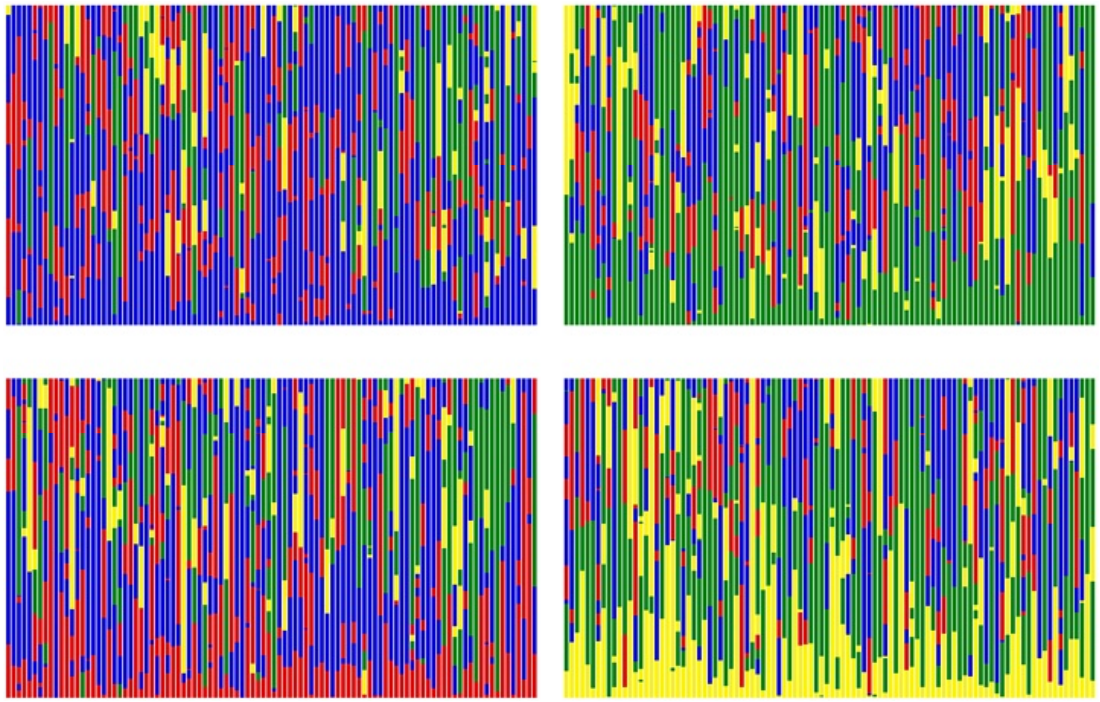
$\mathbf{P}(0.01) =$

	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



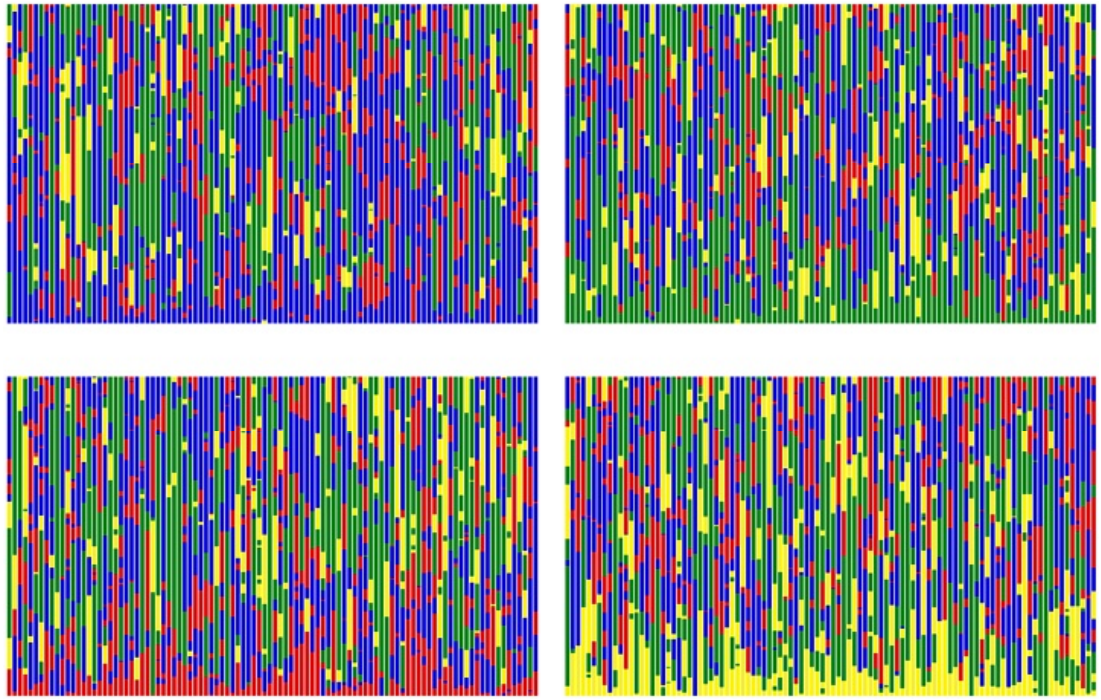
$P(0.50) =$

	A	C	G	T
A	0.7079	0.0813	0.1835	0.0271
C	0.1085	0.7377	0.0542	0.0995
G	0.367	0.0813	0.5244	0.0271
T	0.1085	0.2985	0.0542	0.5387



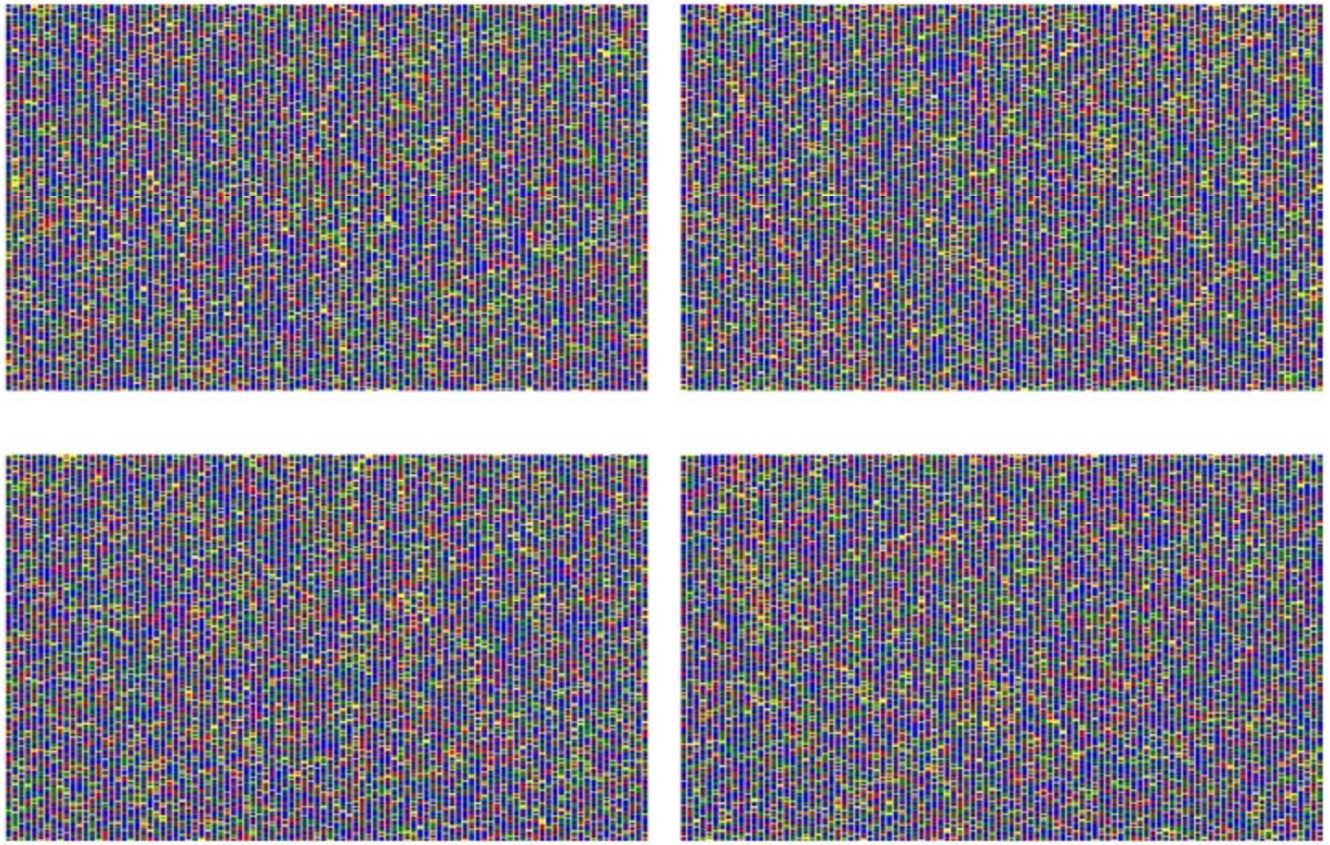
$P(5.00) =$

	A	C	G	T
A	0.4113	0.2873	0.2056	0.0957
C	0.3831	0.319	0.1915	0.1062
G	0.4112	0.2873	0.2056	0.0957
T	0.3831	0.3188	0.1915	0.1065



$\mathbf{P}(10.00) =$

	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



Stationary distribution!

$\pi_A = 0.4$

$\pi_C = 0.3$

$\pi_G = 0.2$

$\pi_T = 0.1$

we start our likelihood with this probability

P(1000.00) =

	A	C	G	T
A	0.4	0.3	0.2	0.1
C	0.4	0.3	0.2	0.1
G	0.4	0.3	0.2	0.1
T	0.4	0.3	0.2	0.1