More intuition on phylogenetic tree estimation

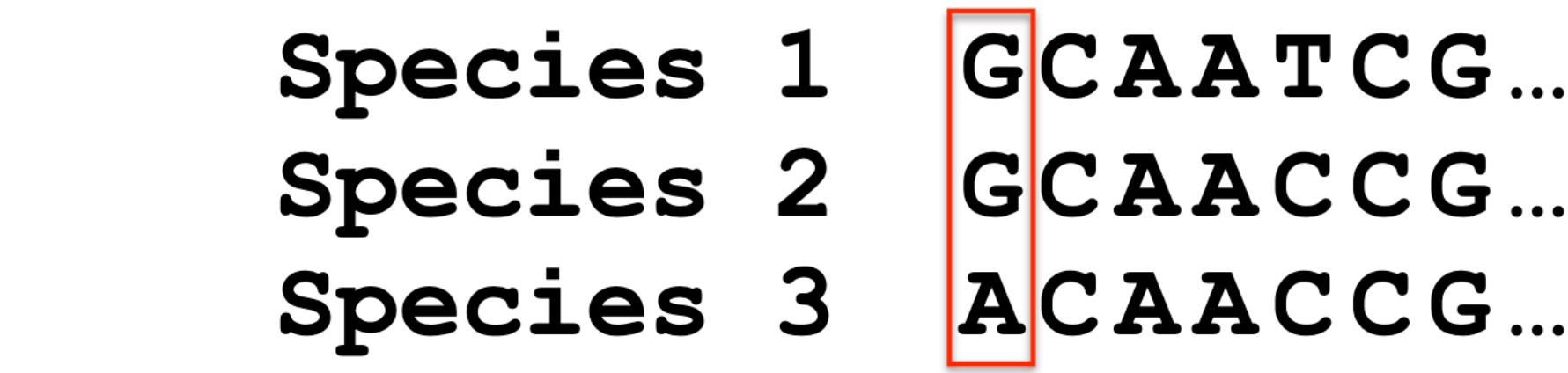
Simplification from John Huelsenbeck's lecture
Woods Hole Molecular Phylogenetics

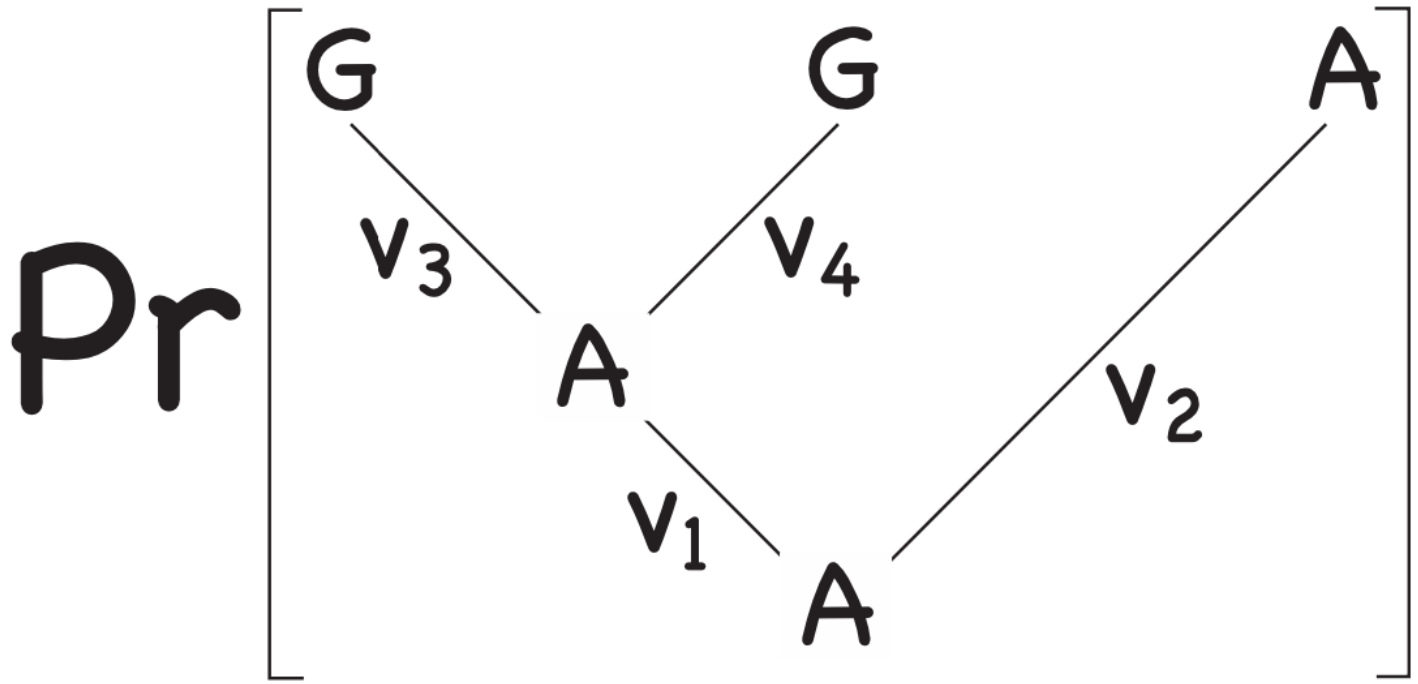
```
#NEXUS

begin data;
     dimensions ntax=5 nchar=895;
     format gap=- datatype=dna;
     matrix
     Human      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
     Chimpanzee AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
     Gorilla    AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTT.....AACCCAAACAATTCAACTCTCCCTAAGCTT
     Orangutan  AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTC.....CACCCAGACACTACAACTCTCACTAAGCTT
     Gibbon     AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTA.....AACCCAAACGCTAGAACTCTCCCTAAGCTT
     ;
end;
```

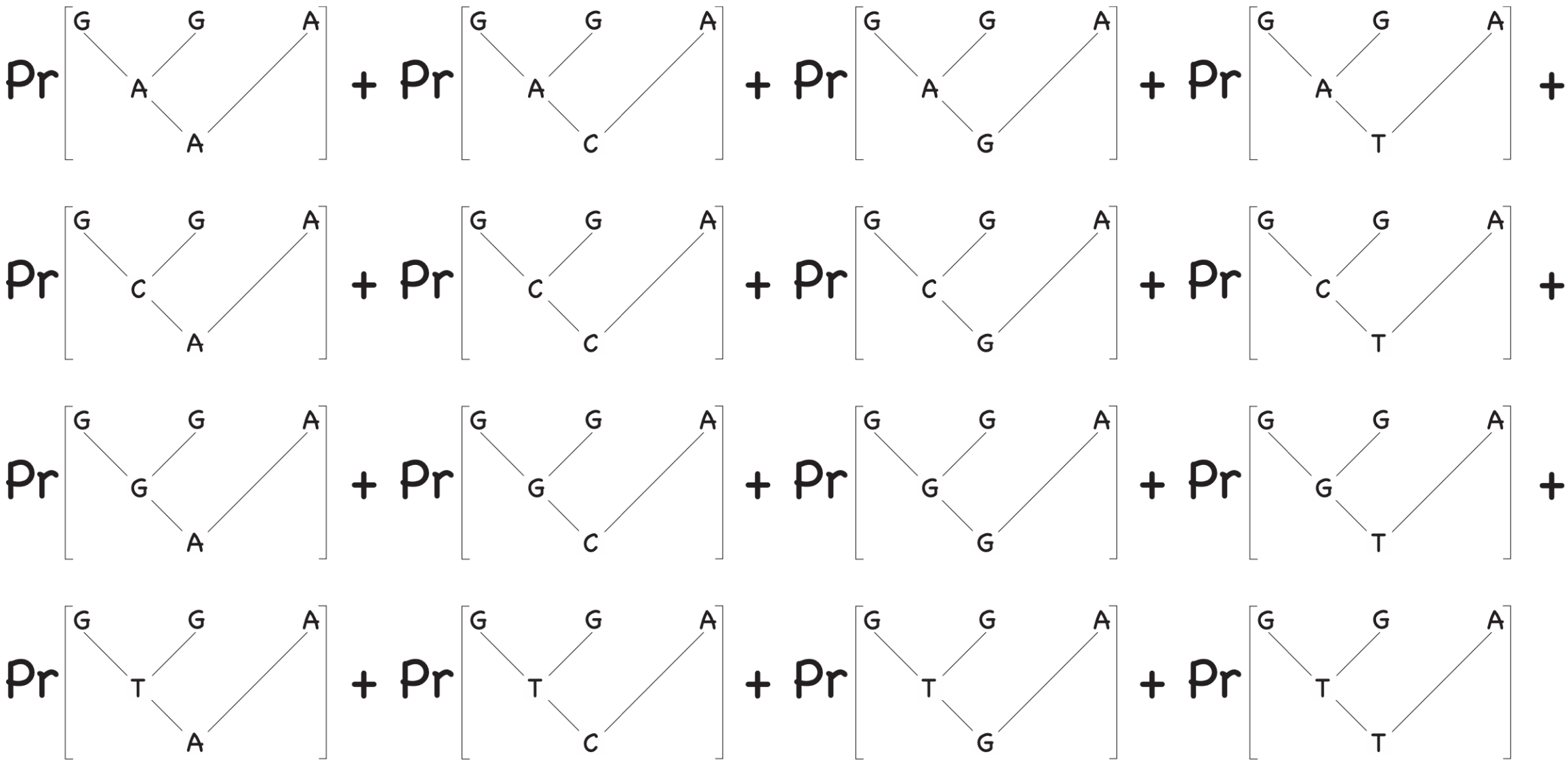
```
Human      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
Chimpanzee AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
Gorilla    AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTT.....AACCCAAACAATTCAACTCTCCCTAAGCTT
Orangutan  AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTC.....CACCCAGACACTACAACTCTCACTAAGCTT
Gibbon     AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTA.....AACCCAAACGCTAGAACTCTCCCTAAGCTT
```

$$\Pr\begin{pmatrix} A \\ A \\ A \\ A \\ A \end{pmatrix} \times \Pr\begin{pmatrix} A \\ A \\ A \\ A \\ A \end{pmatrix} \times \Pr\begin{pmatrix} G \\ G \\ G \\ G \\ G \end{pmatrix} \times \Pr\begin{pmatrix} C \\ C \\ C \\ C \\ C \end{pmatrix} \times \Pr\begin{pmatrix} T \\ T \\ T \\ T \\ T \end{pmatrix} \times \Pr\begin{pmatrix} T \\ T \\ T \\ T \\ T \end{pmatrix} \times \Pr\begin{pmatrix} C \\ C \\ C \\ C \\ T \end{pmatrix} \cdots$$

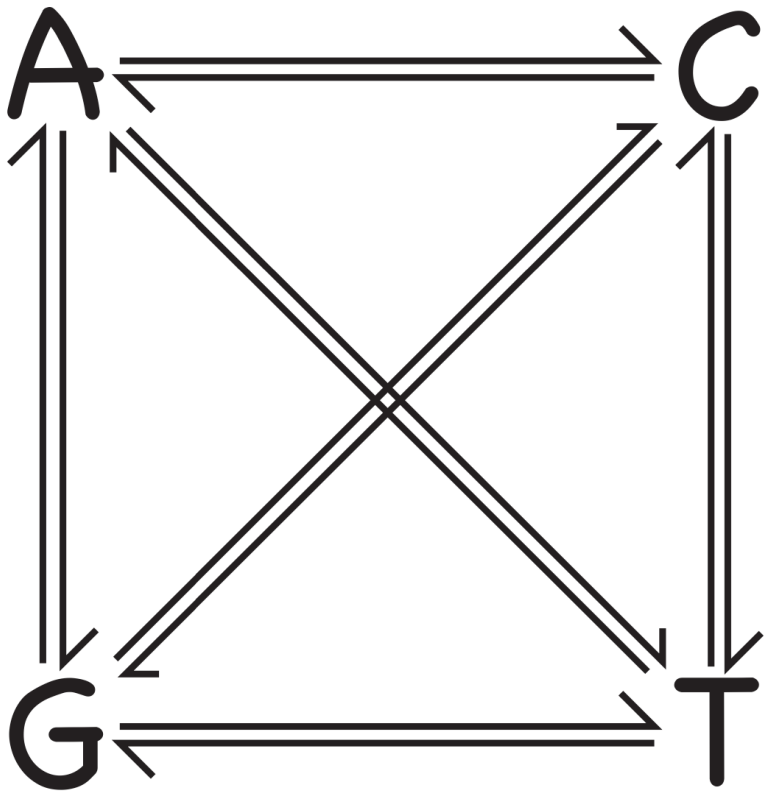

Species 1 GCAATCG…
Species 2 GCAACCG…
Species 3 ACAACCG…



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


Pr G G A A + Pr G G A C + Pr G G A G + Pr G G A T +
Pr G G A C + Pr G G A C + Pr G G A G + Pr G G A T +
Pr G G A G + Pr G G A C + Pr G G A G + Pr G G A T +
Pr G G A T + Pr G G A C + Pr G G A G + Pr G G A T

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

Exponential distribution

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.

		To			
		A	C	G	T
From	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: The change is to state j with probability $-q_{ij}/q_{ii}$.

Finish

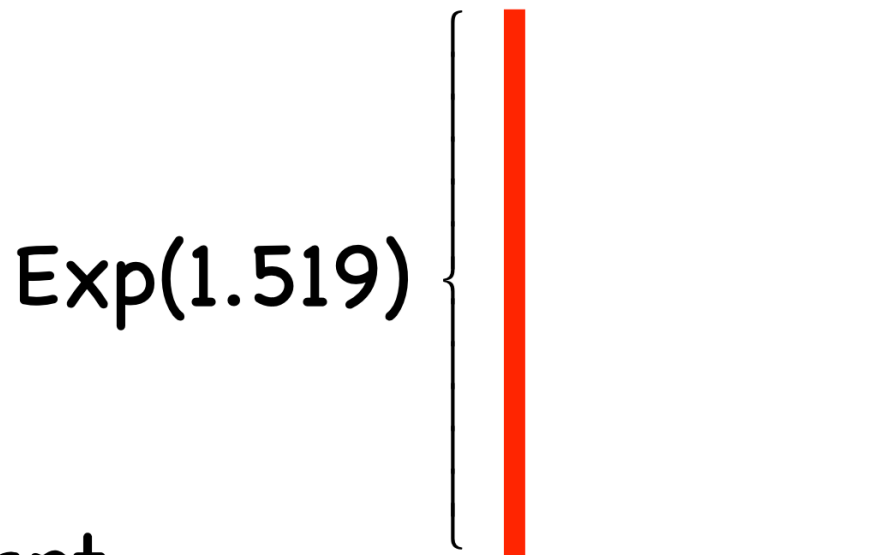
	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

Start in state G

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



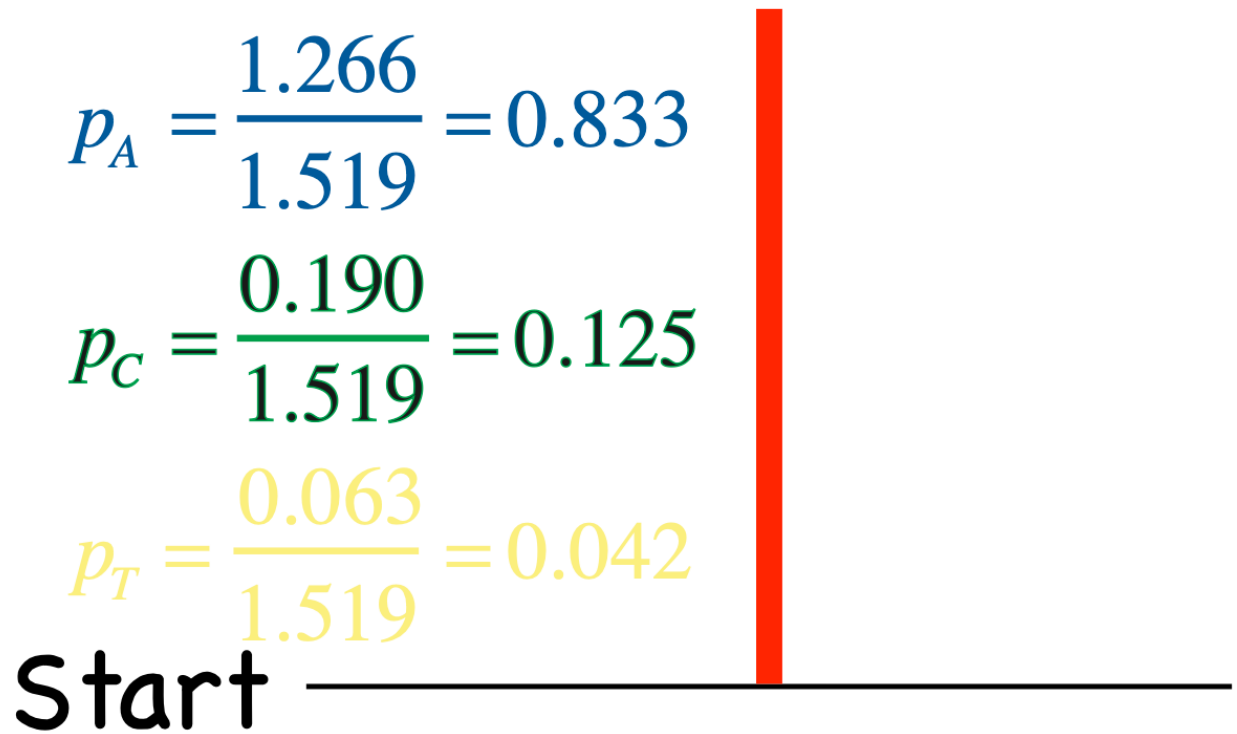
Start

	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

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

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

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

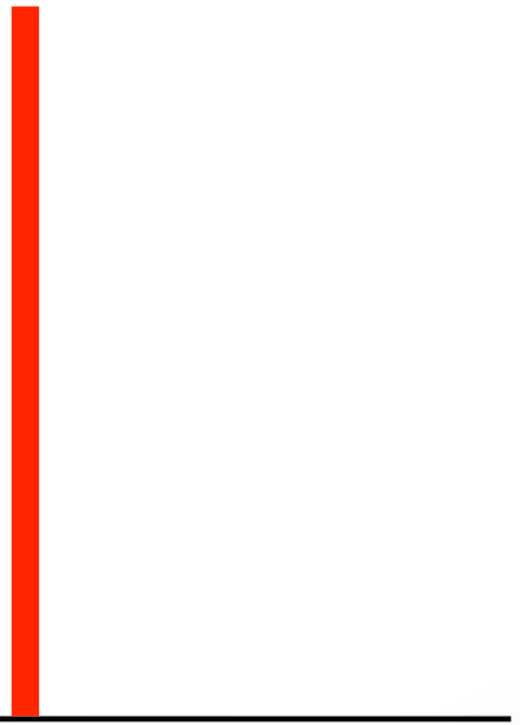


	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329

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

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

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



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

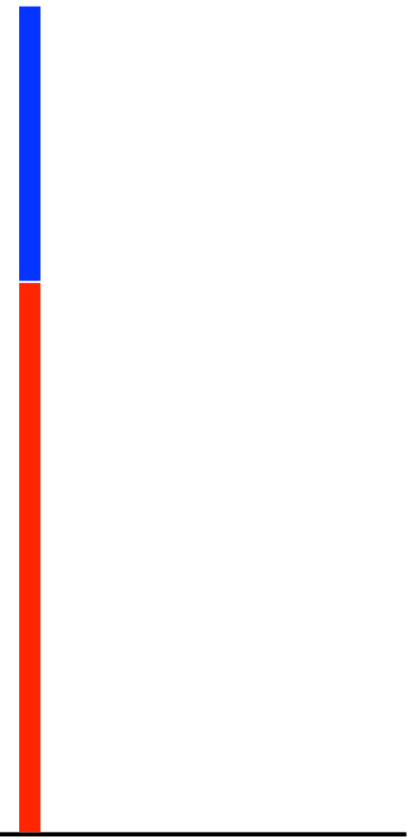
Finish

	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

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

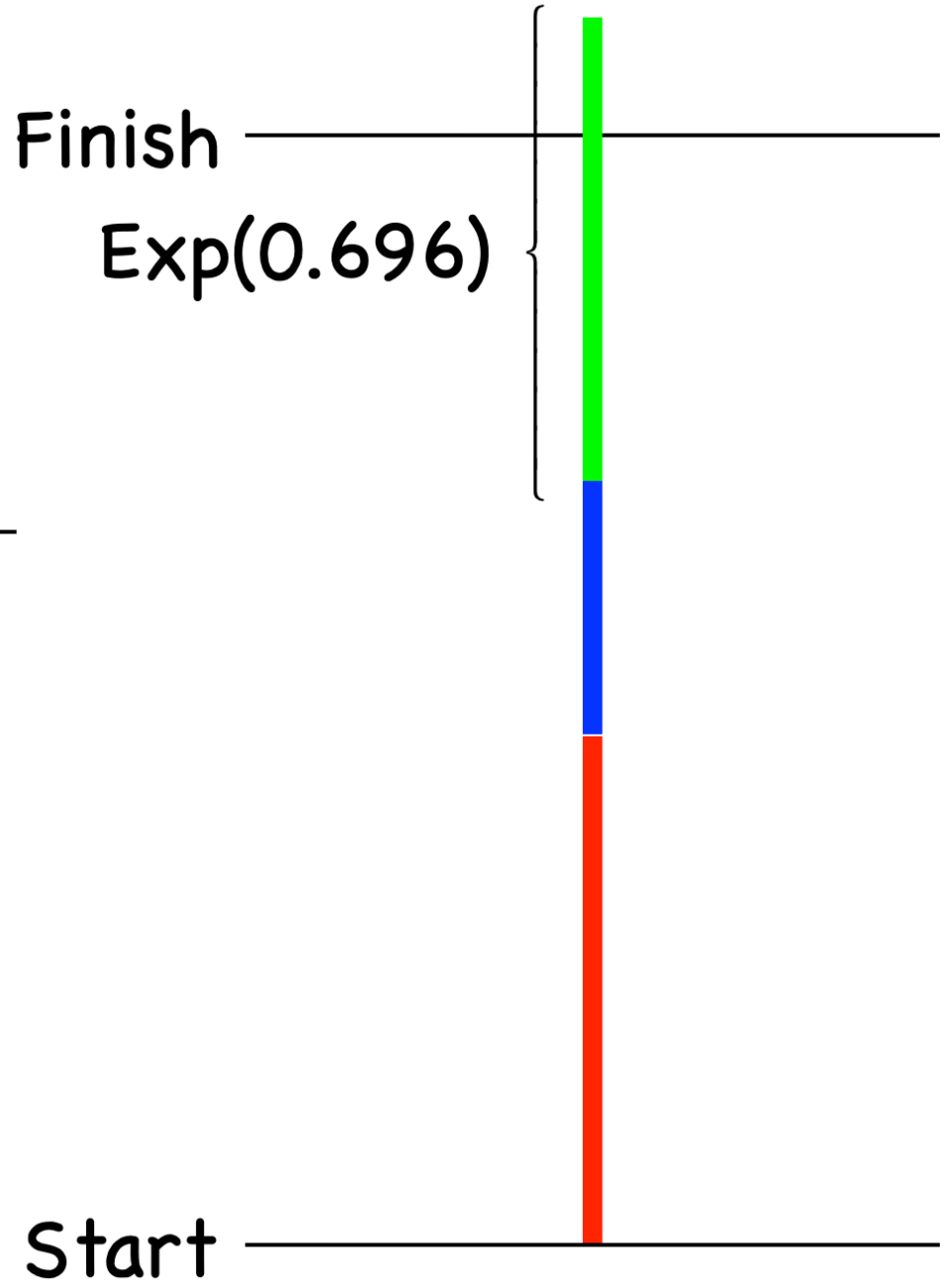
$$p_G = \frac{0.633}{0.886} = 0.714$$

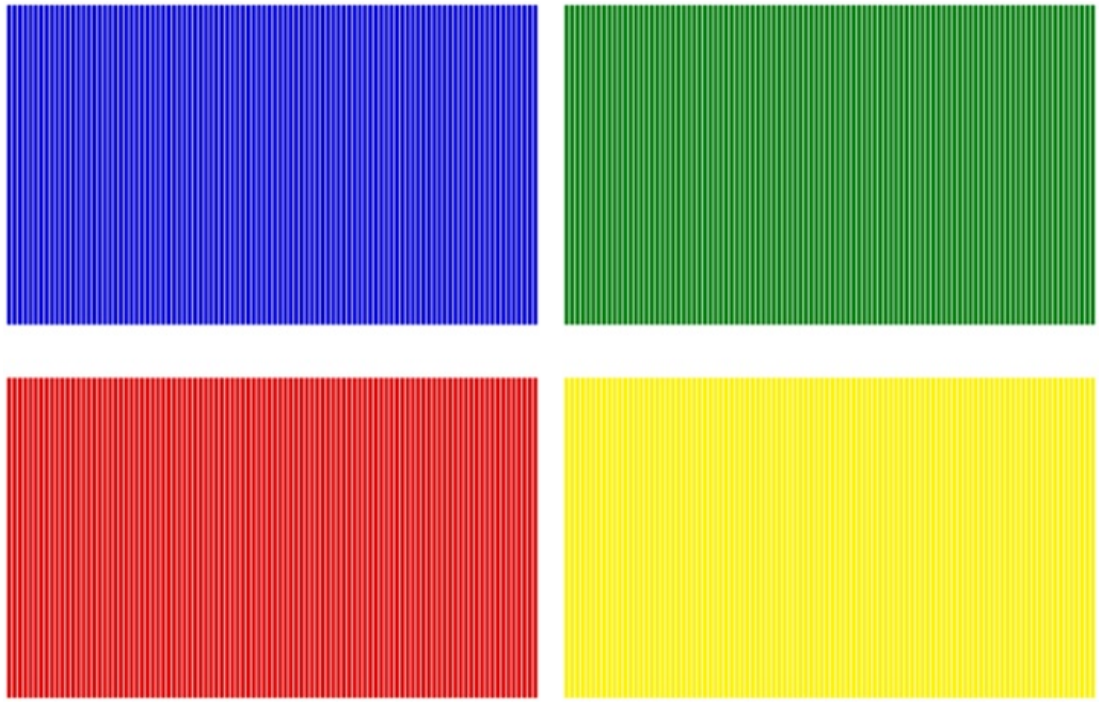
$$p_T = \frac{0.063}{0.886} = 0.072$$



Start

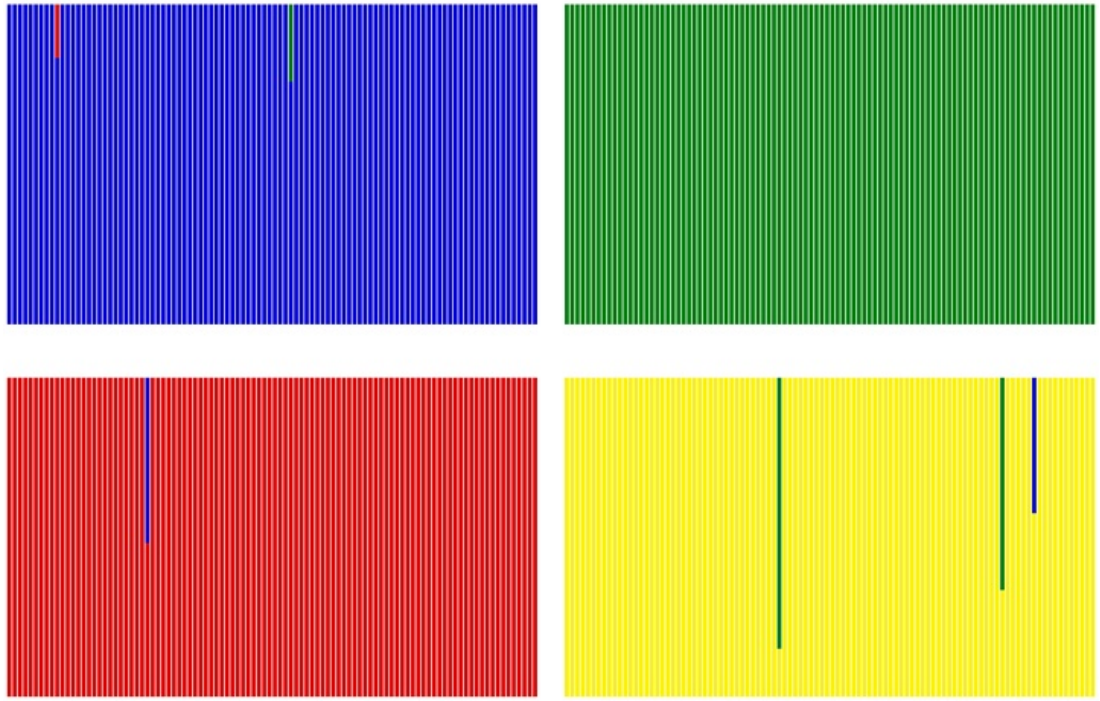
	A	C	G	T
A	−0.886	0.19	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.19	−1.519	0.063
T	0.253	0.949	0.127	−1.329





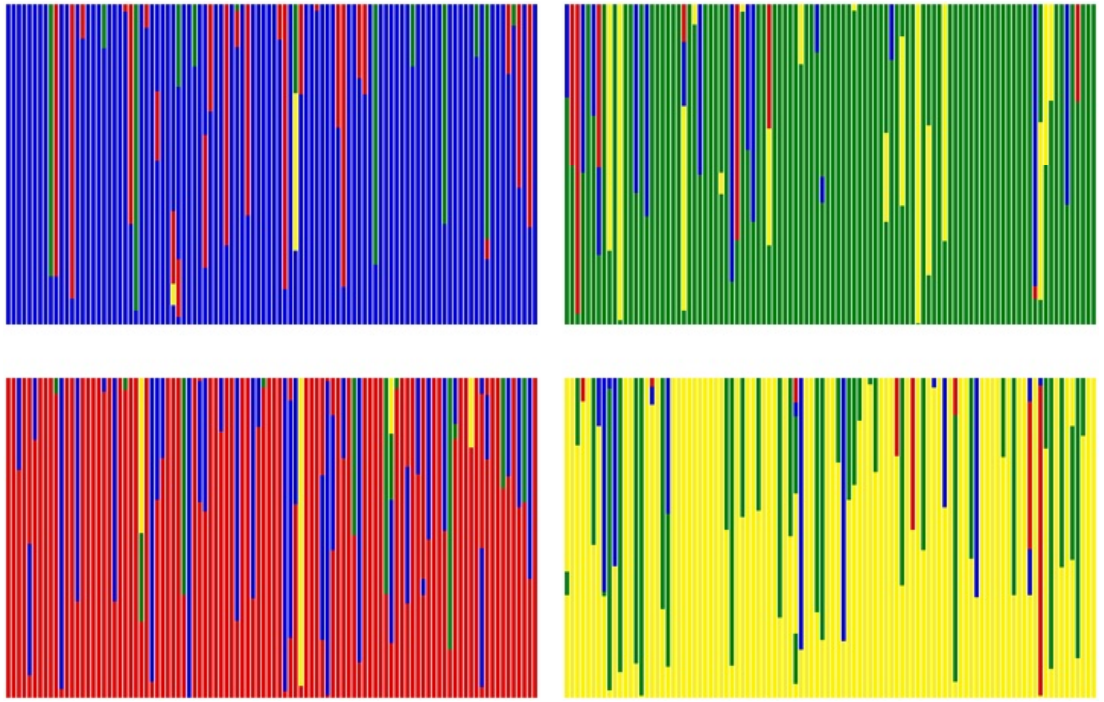
P(0.00) =

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



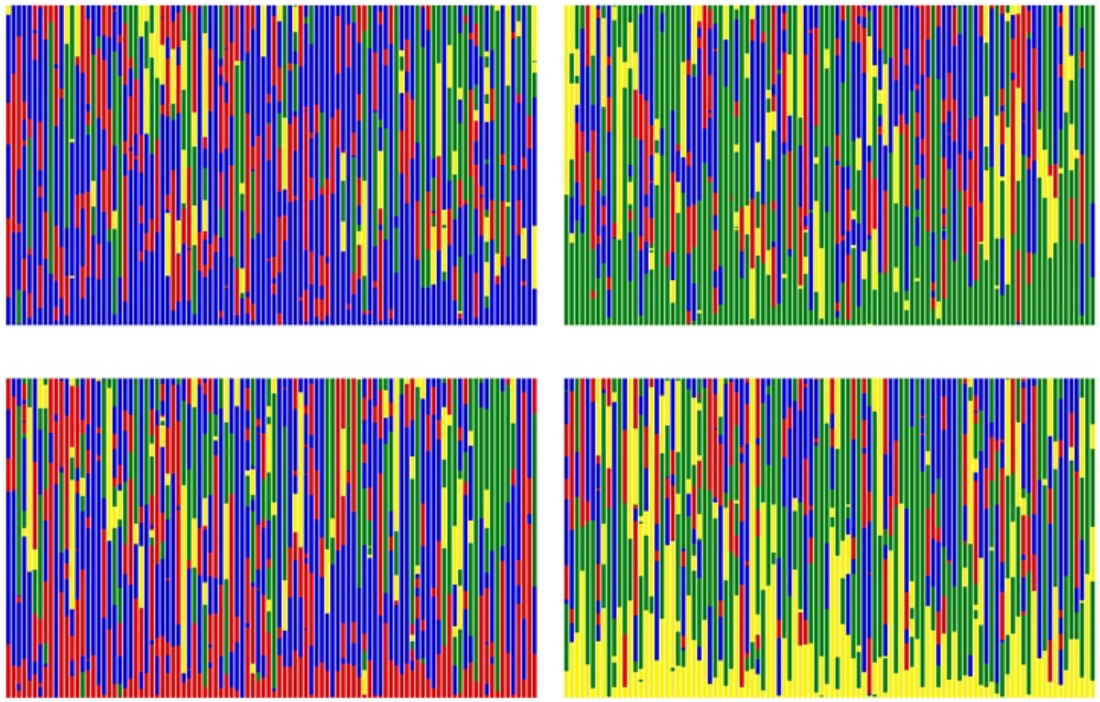
$\mathbf{P}(0.01) =$

	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



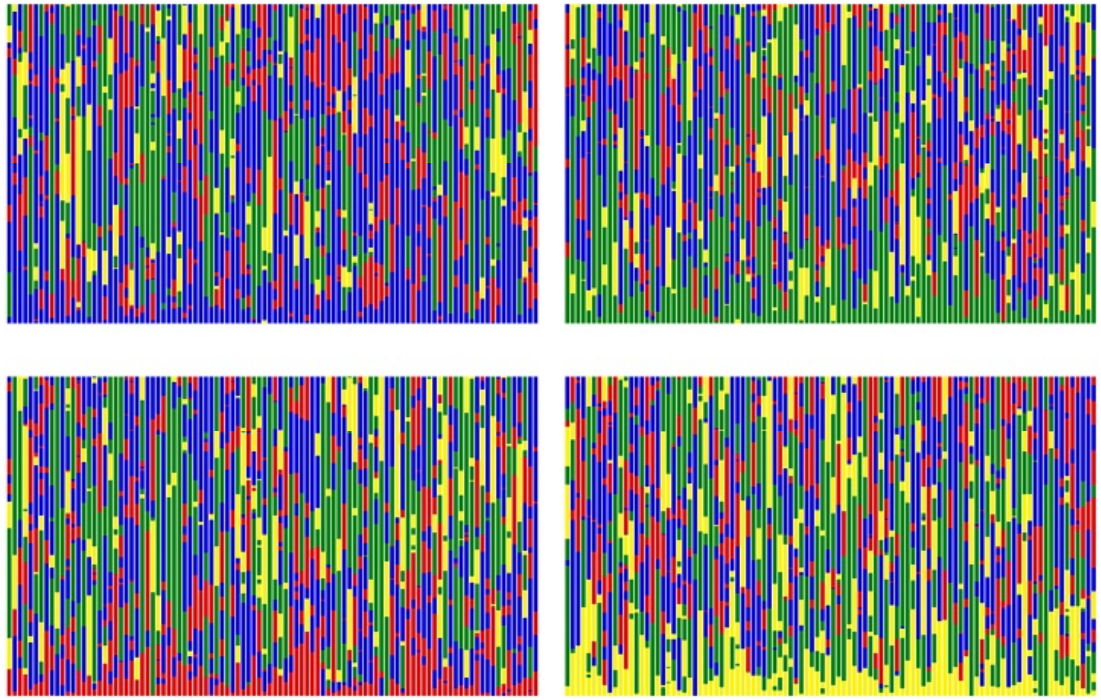
$P(0.50) =$

	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



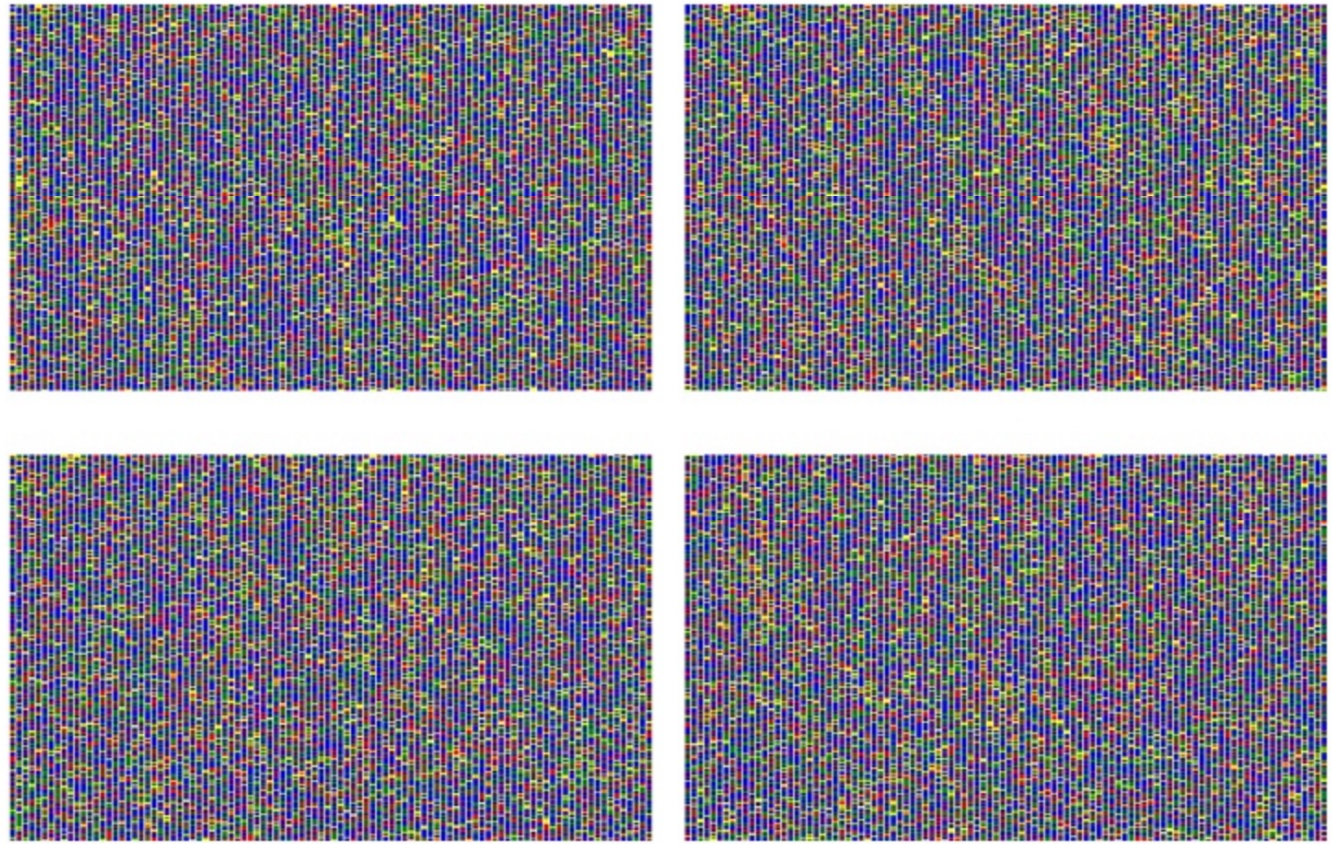
$P(5.00) =$

	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



$\mathbf{P}(1000.00) =$

	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

Working with real data

- Tips in the phylogeny: genes, species, individuals, or some other evolving entity, are referred to as taxa.

Add taxa

- Adding taxa breaks up branches in the phylogeny. This often improves the performance of maximum likelihood
- Improve model evaluation and model parameter estimation
- Adding more outgroup taxa often improves rooting
- Adding taxa can considerably increase the computational demand of analyses

Types of real data

- Full genome: Largely restricted to organisms with small genomes. Expensive and difficult to build phylogenies with large genome sequences, though this is rapidly changing.

- Transcriptome: RNA gives a snapshot of an enriched subset of the genome mRNA is then copied to complimentary DNA (cDNA) and sequenced. Less stable than DNA.

- Targeted enrichment: We design short bait sequences that are similar to conserved genome regions. Data are difficult to combine across different studies that used different baits.

- RAD-seq: Sequence data from specific regions scattered across the genome. It uses intrinsic properties of the genome for enrichment, rath,er than user-designed baits