Modeling Molecular Evolution: Markov models

Maria Anisimova

Institute of Computational Life Sciences
Zurich University of Applied Sciences - ZHAW

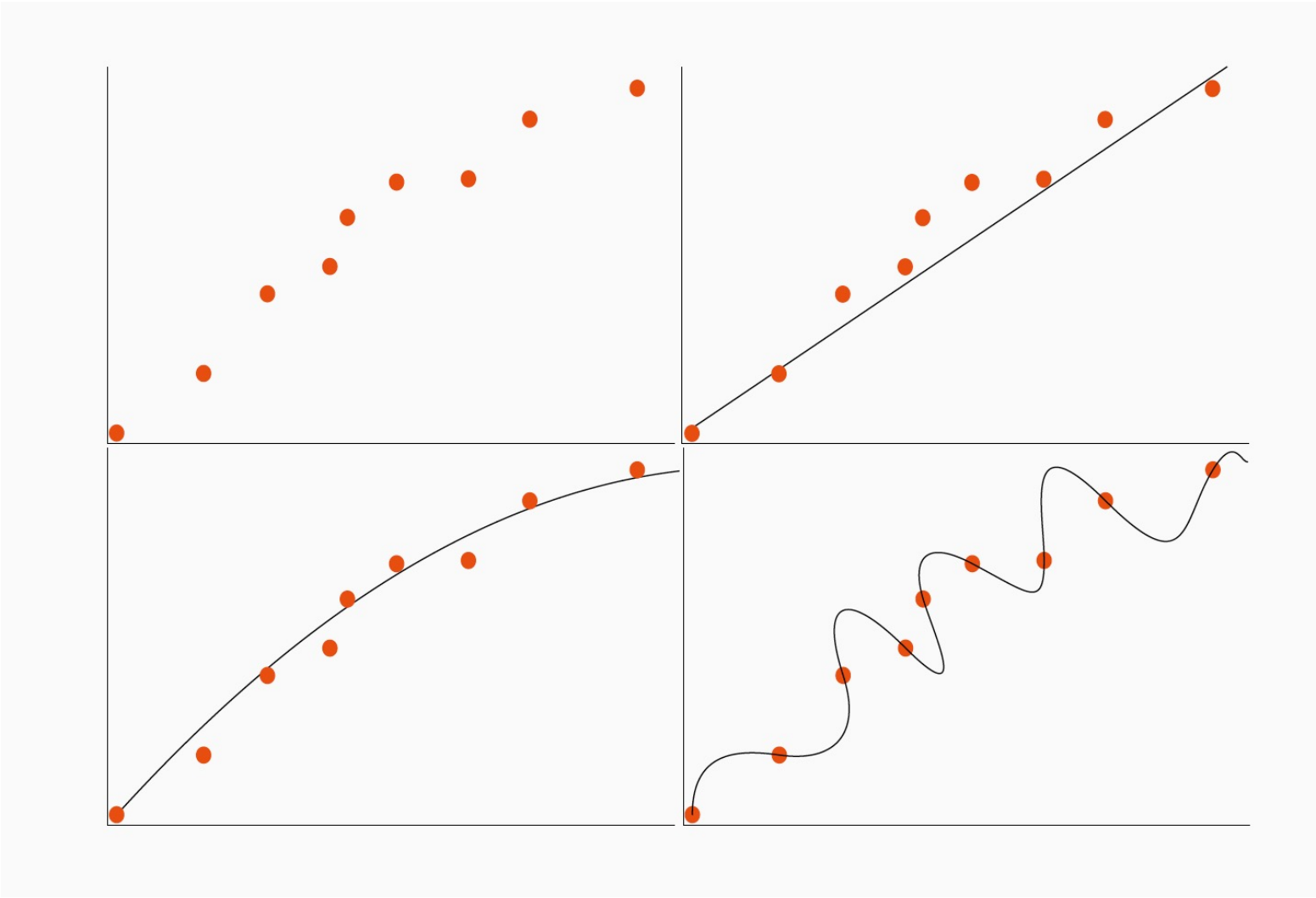
Life Sciences and Facility Management

Institute of Computational Life Sciences

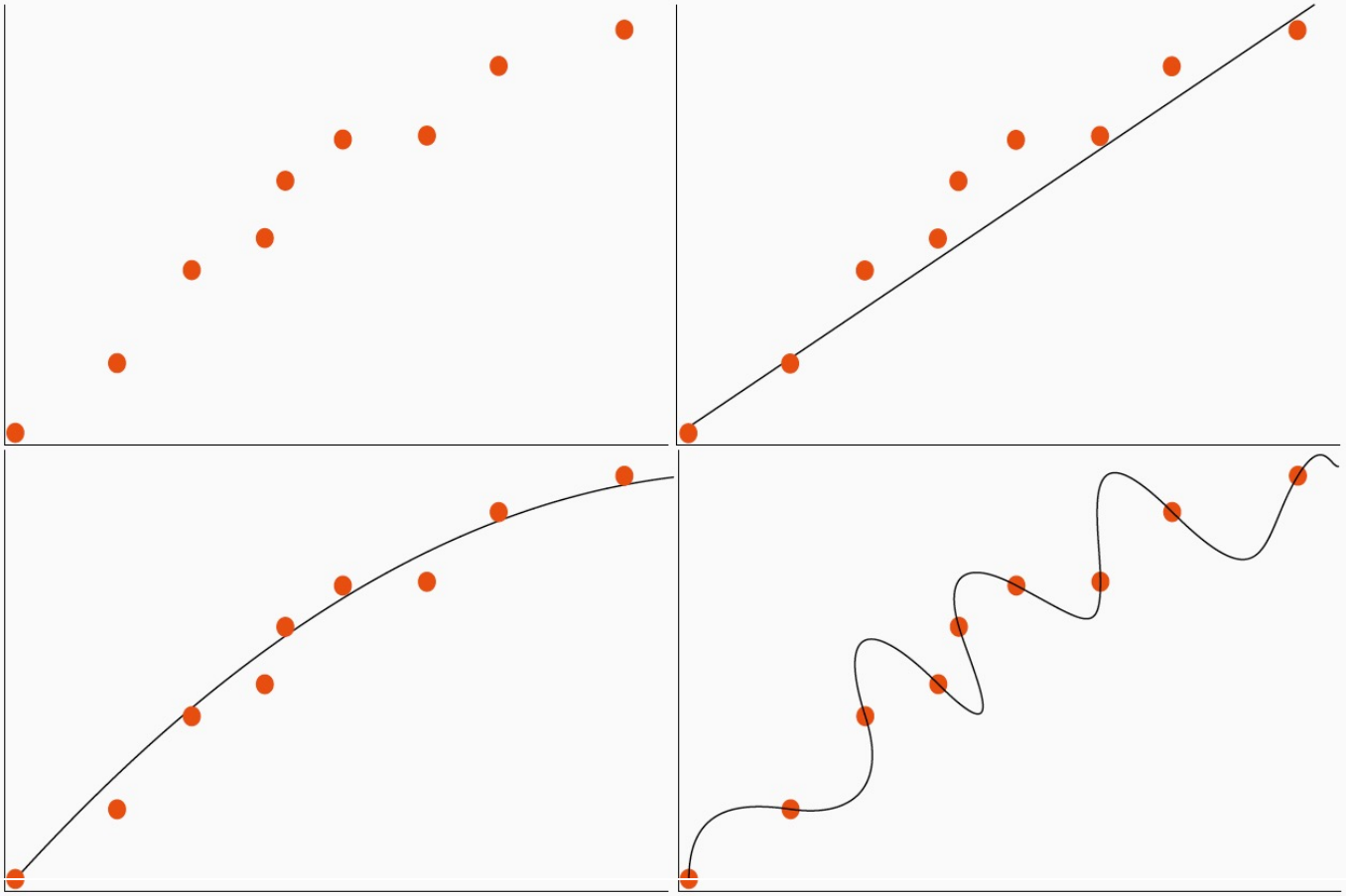
10.05.2022, Lund

Swiss Institute of Bioinformatics

All models are wrong…



All models are wrong…



… but some are useful. (*Box 1976*)

Uses of Molecular Evolution Models

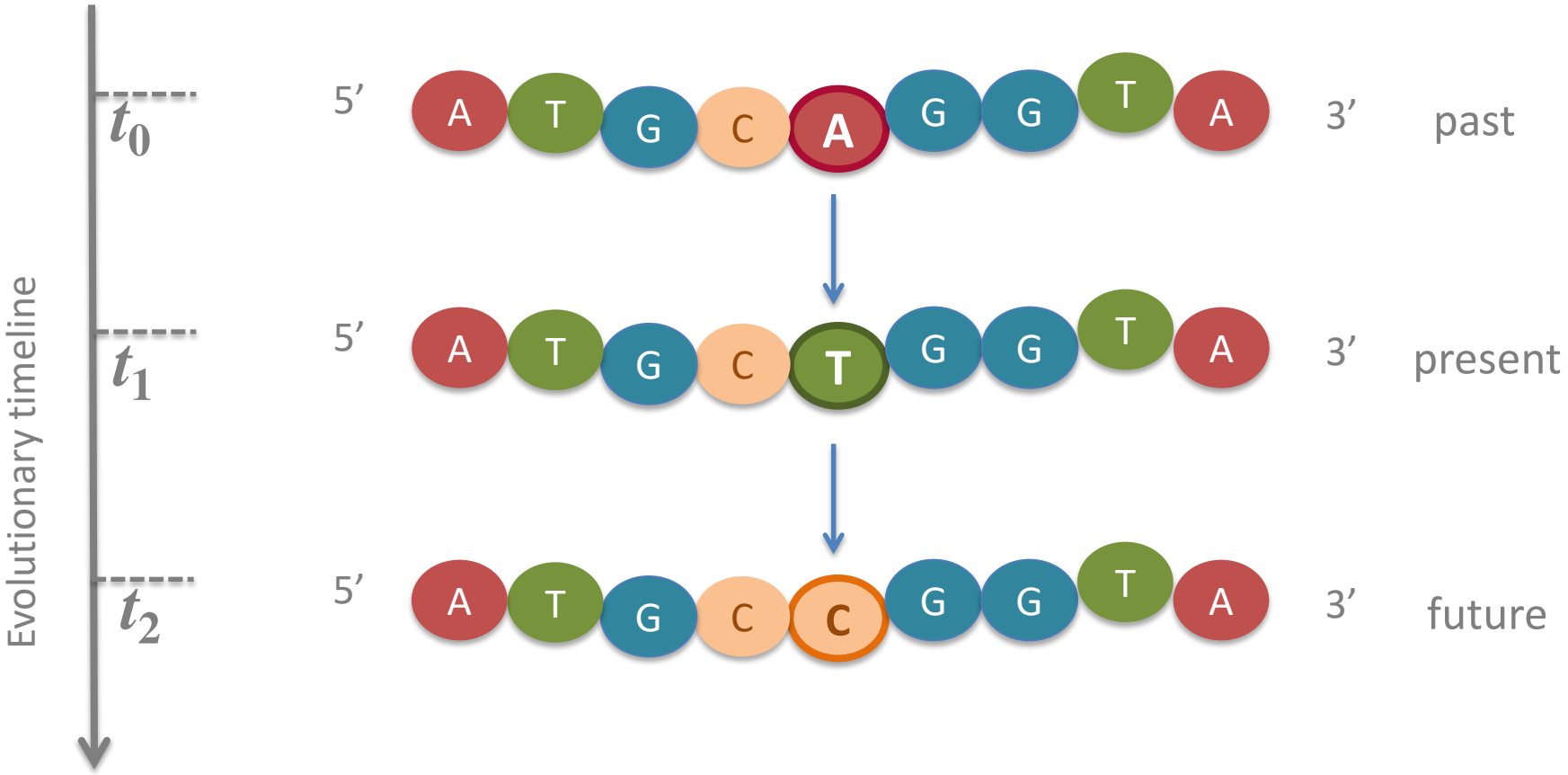
- Test hypotheses
- Study molecular evolution patterns
- Find homologs, conservation due to function
- Which sites are conserved /under positive selection?
- Infer sites involved in evasion from immune response and used in vaccine design
- Infer mutation rates, biases across clades/species
- Molecular dating
- Study evolution of gene families using phylogenetics
- How does environment/ecology affect genomes?
- Connection between genotype & phenotype?

Andrey Markov

1856 - 1922

Russian mathematician
Described the rules of a process:
inspired by Eugene Onegin of Pushkin

Markov model of substitution



Memoriless property:

$$\Pr(C_{future} \mid T_{present} \,\&\, A_{past}) = \Pr(C_{future} \mid T_{present})$$

Markov model of substitution

The future depends only on the current state

States $X(t)$: discrete or continuous

Time t: discrete (e.g., # generations) or continuous

Simple & convenient mathematically

Typical assumptions:

- Independence of evolution at sites
- Stationarity
- Homogeneity
- Time reversibility

More formally…

A **discrete** Markov process $X(t)$ in time t is a family of R.V. such that

for any (continuous or discrete) states $x_0, x_1, \ldots, x_t, x_{t+1}$ and any discrete t :

$$\Pr\{X(t+1)=x_{t+1} \mid X(t)=x_t,\ X(t-1)=x_{t-1}, \ldots,\ X(1)=x_1,\ X(0)=x_0\}$$

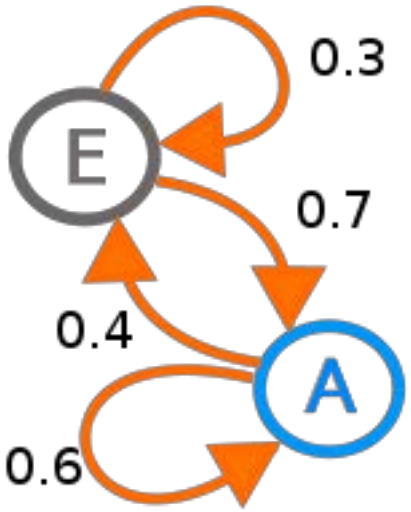
$$= \Pr\{X(t+1)=x_{t+1} \mid X(t)=x_t\}$$

A **continuous** Markov process has continuous index, defined for a family of R.V. $\{X(t),\ 0 \le t < \infty\}$

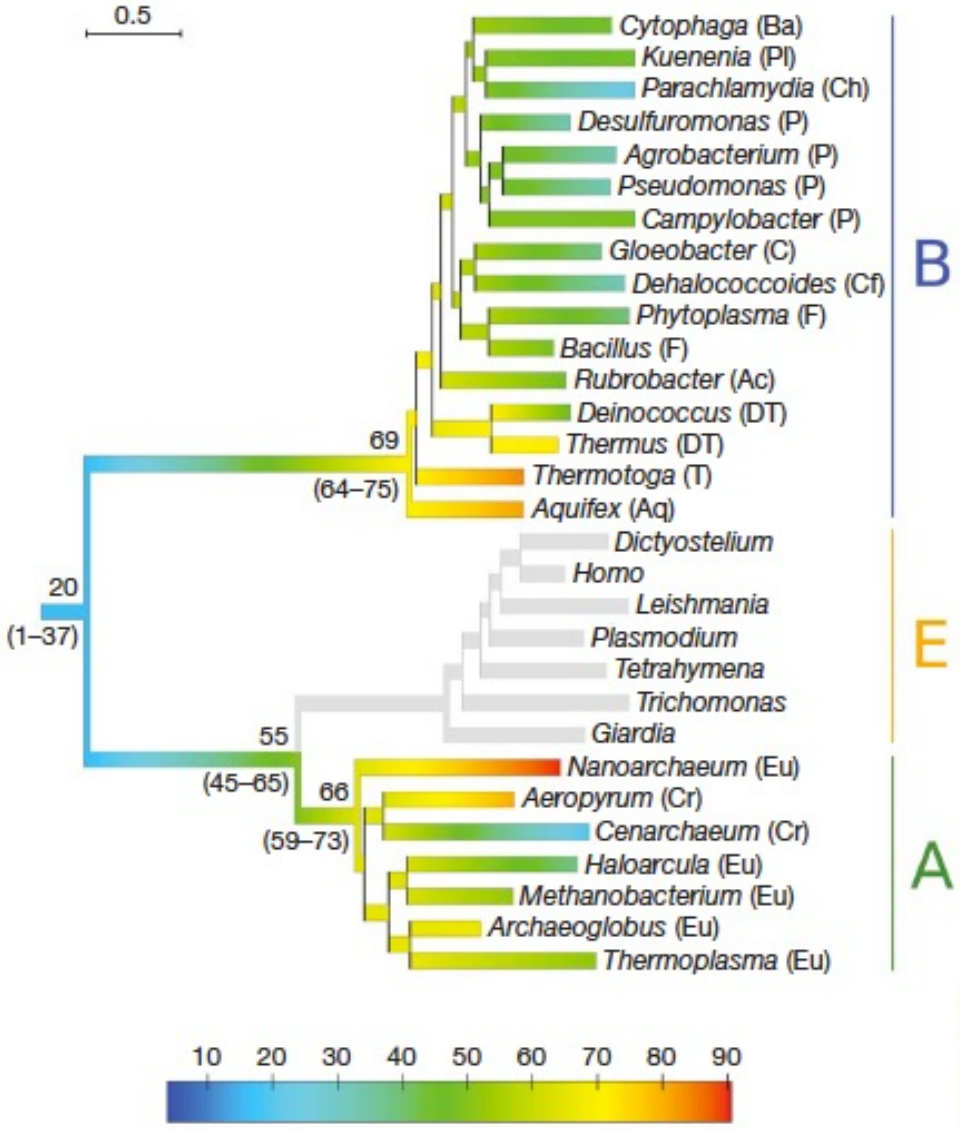
Generating matrix is needed!

For a **homogeneous** Markov process:

$$\Pr\{X(t+1)=x \mid X(t)=y\} = \Pr\{X(t)=x \mid X(t-1)=y\} \text{ for any } t$$



Example of a simple 2-state process:
http://en.wikipedia.org/wiki/Markov_chain



From Boussau et al. 2008, Nature

Nonthermophilic LUCA?

Figure 2 | Evolution of thermophily over the tree of life. Protein-derived nhPhyloBayes OGT estimates (and their 95% confidence intervals for key ancestors) for prokaryotic organisms are colour-coded from blue to red for low to high temperatures. Colours were interpolated between temperatures estimated at nodes. The eukaryotic domain, in which OGT cannot be estimated, has been shaded. The colour scale is in °C; the branch length scale is in substitutions per site. A, archaeal; B, bacterial; E, eukaryotic domains. Ac, Actinobacteria; Aq, Aquificae; Ba, Bacteroidetes; C, Cyanobacteria; Cf, Chloroflexi; Ch, Chlamydiae; Cr, Crenarchaeota; DT, Deinococcus/Thermus; Eu, Euryarchaeota; F, Firmicutes; P, Proteobacteria; Pl, Planctomycetes; T, Thermotogae.

Markov model of DNA substitution

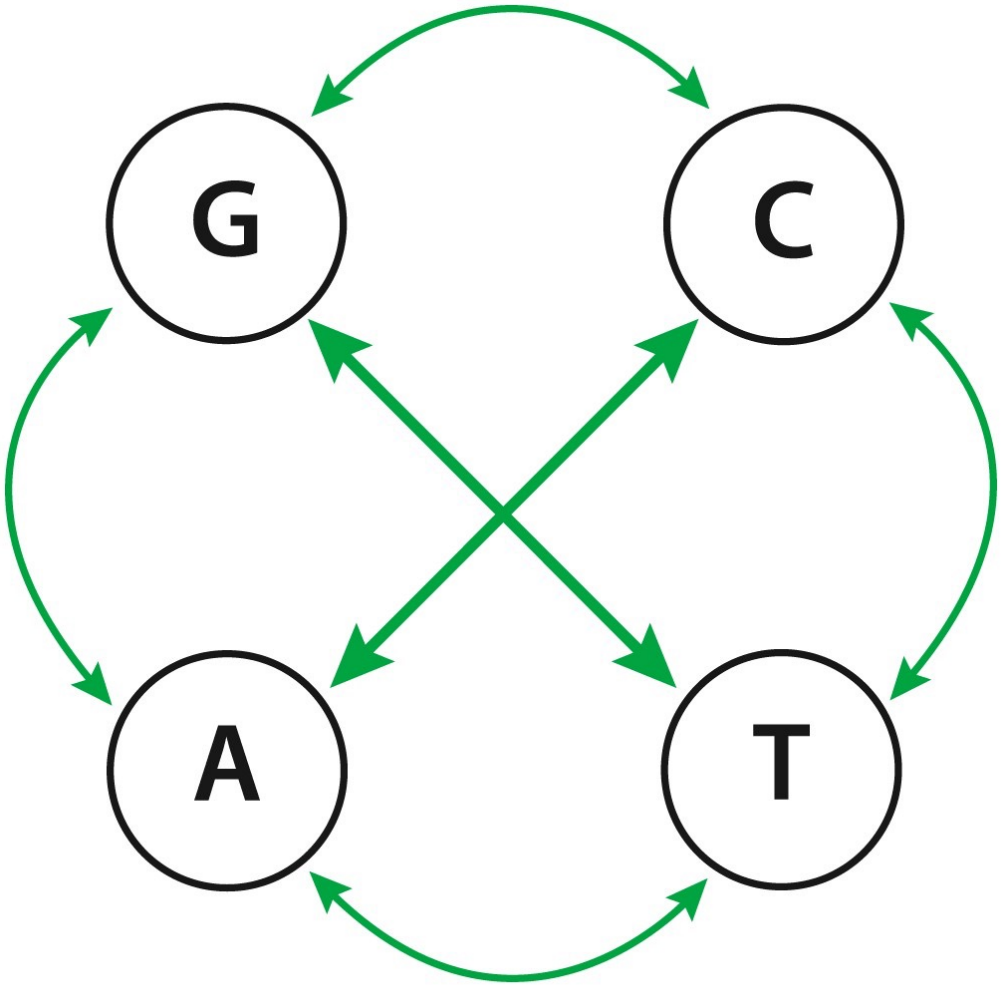


Figure 3.13 Phylogenomics: A Primer (© Garland Science 2013)

Markov model of DNA substitution

Continuous-time Markov process describes substitutions at any site

Character at time *t* is $X(t) \in \{A,C,G,T\}$

Process generating matrix *Q*

$$Q = \begin{pmatrix} q_{TT} & q_{TC} & q_{TA} & q_{TG} \\ q_{CT} & q_{CC} & q_{CA} & q_{CG} \\ q_{AT} & q_{AC} & q_{AA} & q_{AG} \\ q_{GT} & q_{GC} & q_{GA} & q_{GG} \end{pmatrix}$$

q_{ij} are instantaneous rates from *i* to *j*

Process leaves state *i* at rate: $-q_{ii} = \sum_{j \neq i} q_{ij}$

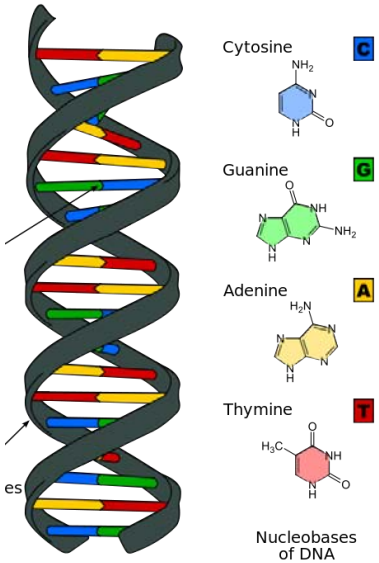
$\Pr\{ X(t+\Delta t)=j \mid X(t)=i\}_{i \neq j} = q_{ij} \Delta t$

If q_{ij} constant over time the process is homogeneous

Q determines transition matrix $P(t) = \{p_{ij}(t)\} = \{\Pr\{ X(t)=j \mid X(0)=i \}\}$, $t > 0$

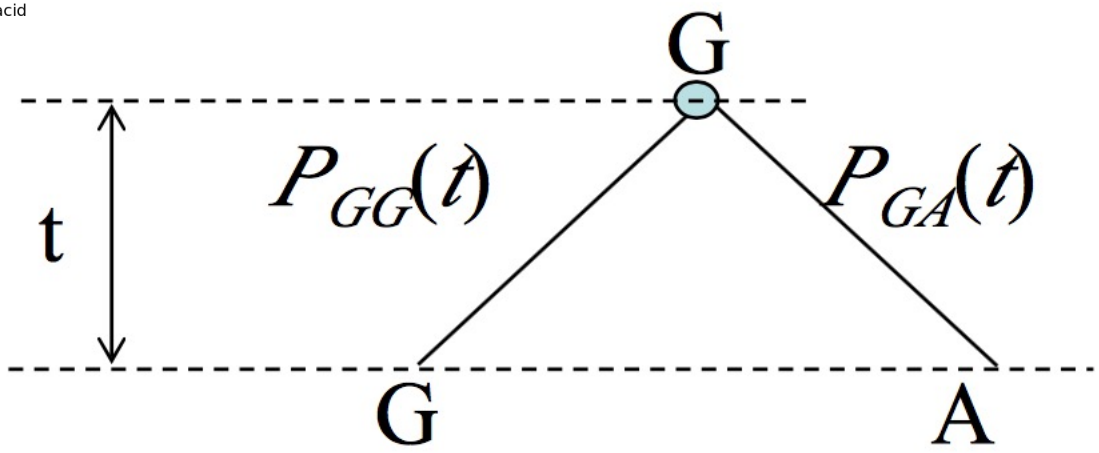
$$\frac{dP(t)}{dt} = P(t)Q \text{ and } P(0) = I \quad \Rightarrow \boxed{P(t) = \exp(Qt)}$$

Markov model of DNA substitution



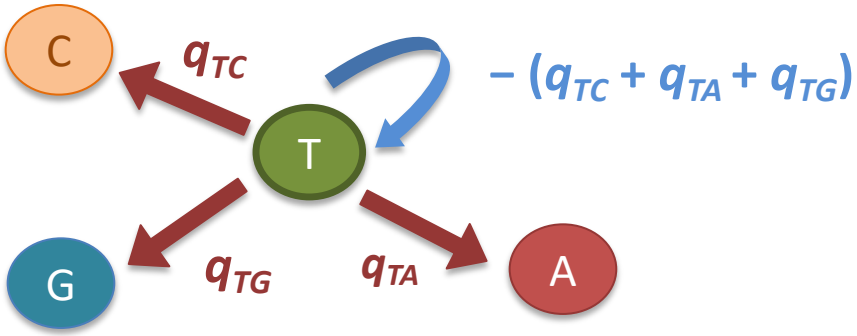
$i = \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}$

$P_{GG}(t)$ and $P_{GA}(t)$ Independent from i



The instantaneous rate matrix

$$Q = \{q_{ij}\} = \begin{pmatrix} -\sum_{j \neq A} q_{Tj} & q_{TC} & q_{TA} & q_{TG} \\ q_{CT} & -\sum_{j \neq C} q_{Cj} & q_{CA} & q_{CG} \\ q_{AT} & q_{AC} & -\sum_{j \neq G} q_{Aj} & q_{AG} \\ q_{GT} & q_{GC} & q_{GA} & -\sum_{j \neq T} q_{Gj} \end{pmatrix}$$



Total rate of change = Rate of staying in the same state

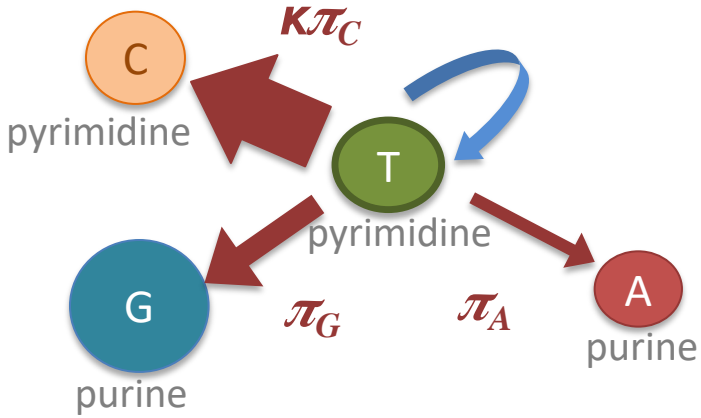
$$q_{TC} + q_{TG} + q_{TA} = -(q_{TC} + q_{TA} + q_{TG})$$

HKY model, Hasegawa-Kishino-Yano (1985)

$$Q_{\text{HKY}} = \begin{pmatrix} \bullet & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & \bullet & \pi_A & \pi_G \\ \pi_T & \pi_C & \bullet & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & \bullet \end{pmatrix}$$

Unequal frequencies

$\pi_T, \pi_C, \pi_A, \pi_G$



Transition (ts) vs. transversion (tv) rate ratio:

$\kappa = ts/tv$

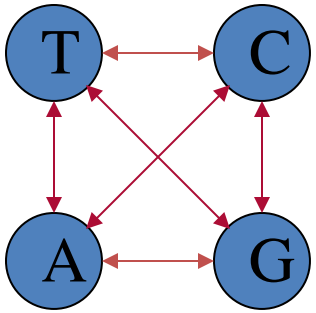
Common models of nucleotide evolution

$$Q_{\mathrm{JC69}} = \begin{pmatrix} \bullet & \lambda & \lambda & \lambda \\ \lambda & \bullet & \lambda & \lambda \\ \lambda & \lambda & \bullet & \lambda \\ \lambda & \lambda & \lambda & \bullet \end{pmatrix}$$

Jukes and Cantor (1969)

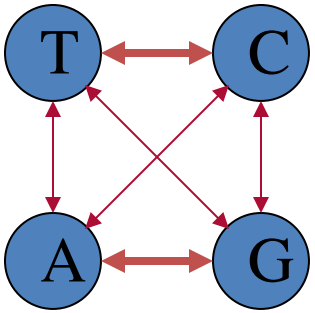
$$Q_{\mathrm{K80}} = \begin{pmatrix} \bullet & \alpha & \beta & \beta \\ \alpha & \bullet & \beta & \beta \\ \beta & \beta & \bullet & \alpha \\ \beta & \beta & \alpha & \bullet \end{pmatrix}$$

Kimura (1980)



pyrimidines

purines



transversions

transitions

Common models of nucleotide evolution

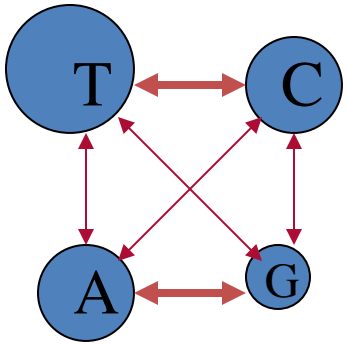
$$Q_{\text{HKY85}} = \begin{pmatrix} \bullet & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \bullet & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \bullet & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \bullet \end{pmatrix}$$

Hasegawa, Kishino, Yano (1984-85)

$$Q_{\text{TN93}} = \begin{pmatrix} \bullet & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & \bullet & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \bullet & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \bullet \end{pmatrix}$$

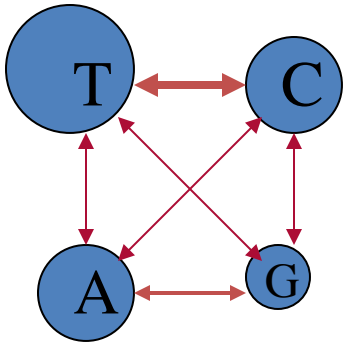
Tamura and Nei (1993)

Similar to F81 (Felsenstein 1981)



pyrimidines

purines



transversions

transitions

The probability of transition over time

$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

The probability of transition over time

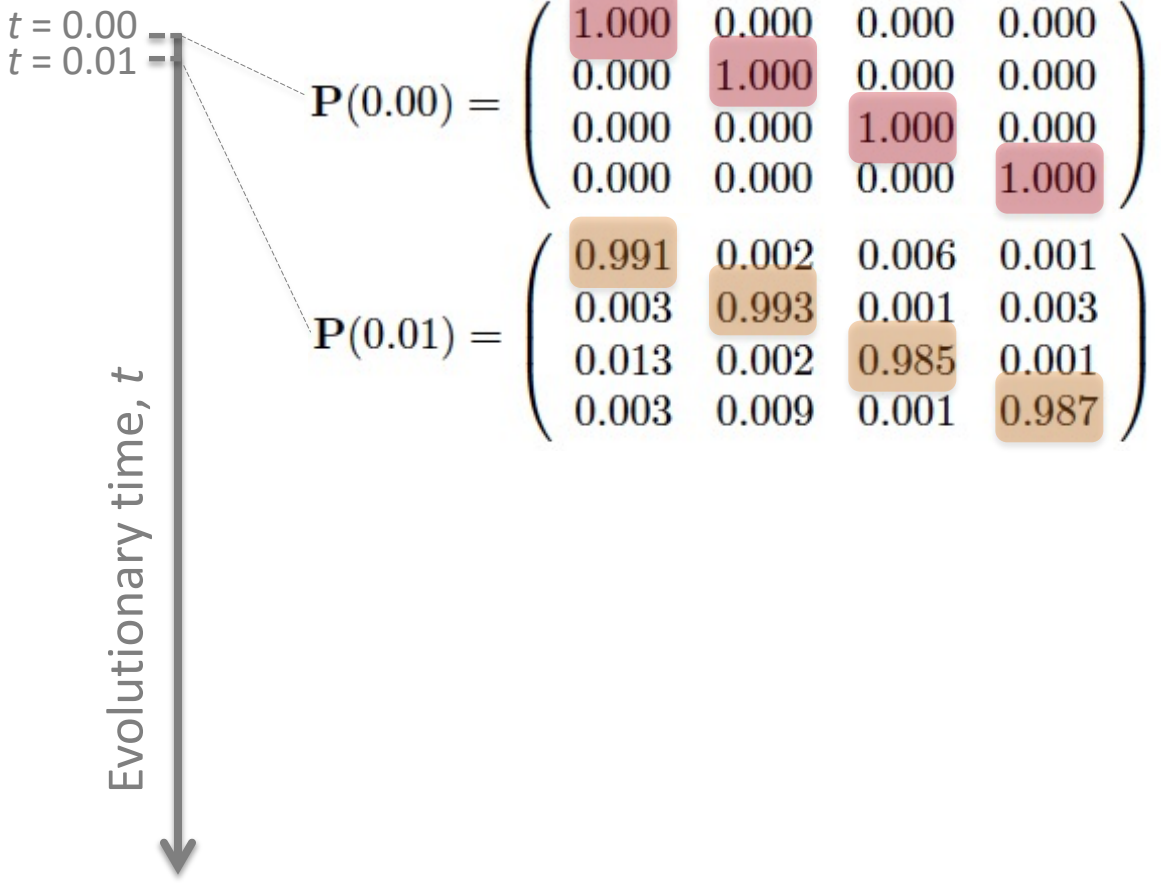
$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$

HKY model:

$\kappa = 5$

$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) =$

$(\mathbf{0.4, 0.3, 0.2, 0.1})$



The probability of transition over time

$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$

HKY model:

$\kappa = 5$

$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) =$

$(\mathbf{0.4, 0.3, 0.2, 0.1})$

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

$$\mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix}$$

The probability of transition over time

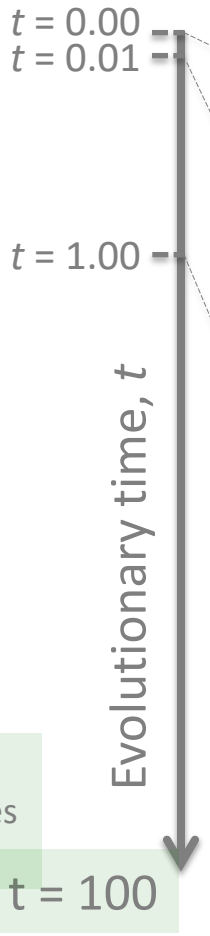
$$\frac{dP(t)}{dt} = P(t)Q \text{ et } P(0)=I \quad \Rightarrow \quad P(t) = \exp(Qt)$$

HKY model:

$\kappa = 5$

$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) =$

$(\mathbf{0.4, 0.3, 0.2, 0.1})$



$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

$$\mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix}$$

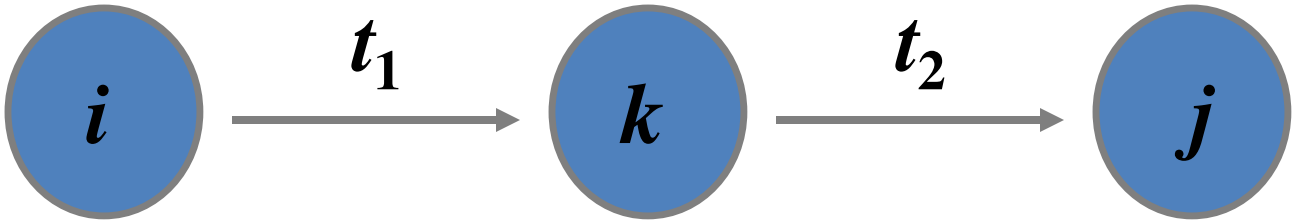
$$\mathbf{P}(100) = \begin{pmatrix} 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \end{pmatrix}$$

Convergence
to stationary frequencies
stationnaires:

Multiple substitutions

Markov process accounts for multiple hits and hidden changes.
By Chapman-Kolmogorov theorem:

$$p_{ij}(t_1+t_2) = \sum_k p_{ik}(t_1)\, p_{kj}(t_2) \quad \text{for } k \in \{\text{T, C, A, G}\}$$



Stationary distribution of states

Initial distribution of $X(0)$: $\pi(0) = (\pi_T(0), \pi_C(0), \pi_A(0), \pi_G(0))$

At time t: $\pi(t) = \pi(0)\, P(t)$

OR $\pi_i(t) = \pi_T(0)\, p_{Ti}(t) + \pi_C(0)\, p_{Ci}(t) + \pi_A(0)\, p_{Ai}(t) + \pi_G(0)\, p_{Gi}(t)$

The process is stationary if $\pi(t) = \pi(0)$ for any $t > 0$

Stationary distribution: $\pi = \pi P(t) \quad \Rightarrow \pi Q = 0$

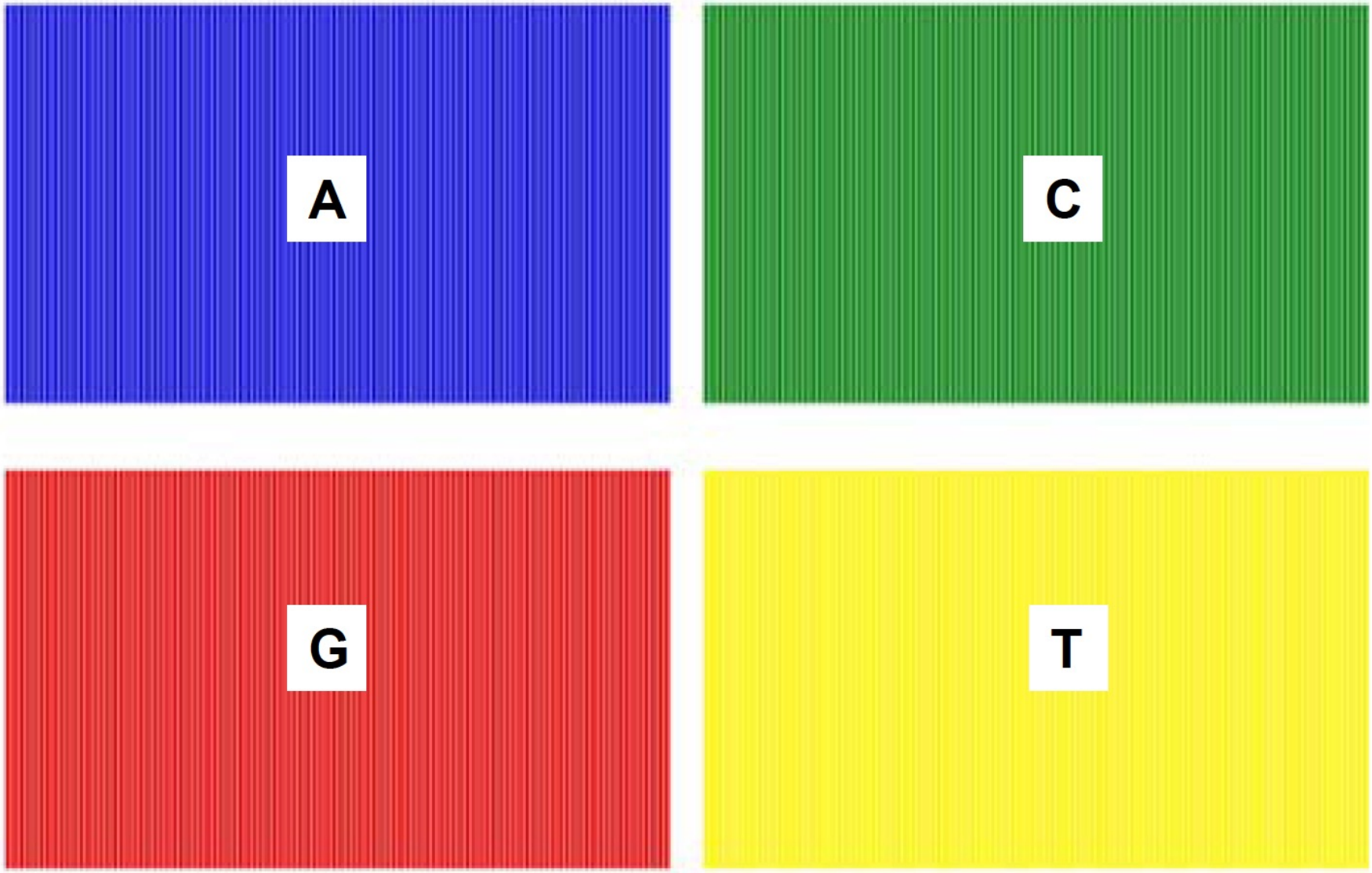
(π is an eigenvector for eigenvalue 0)

OR $\sum_i \pi_i q_{ij} = 0$ for any j

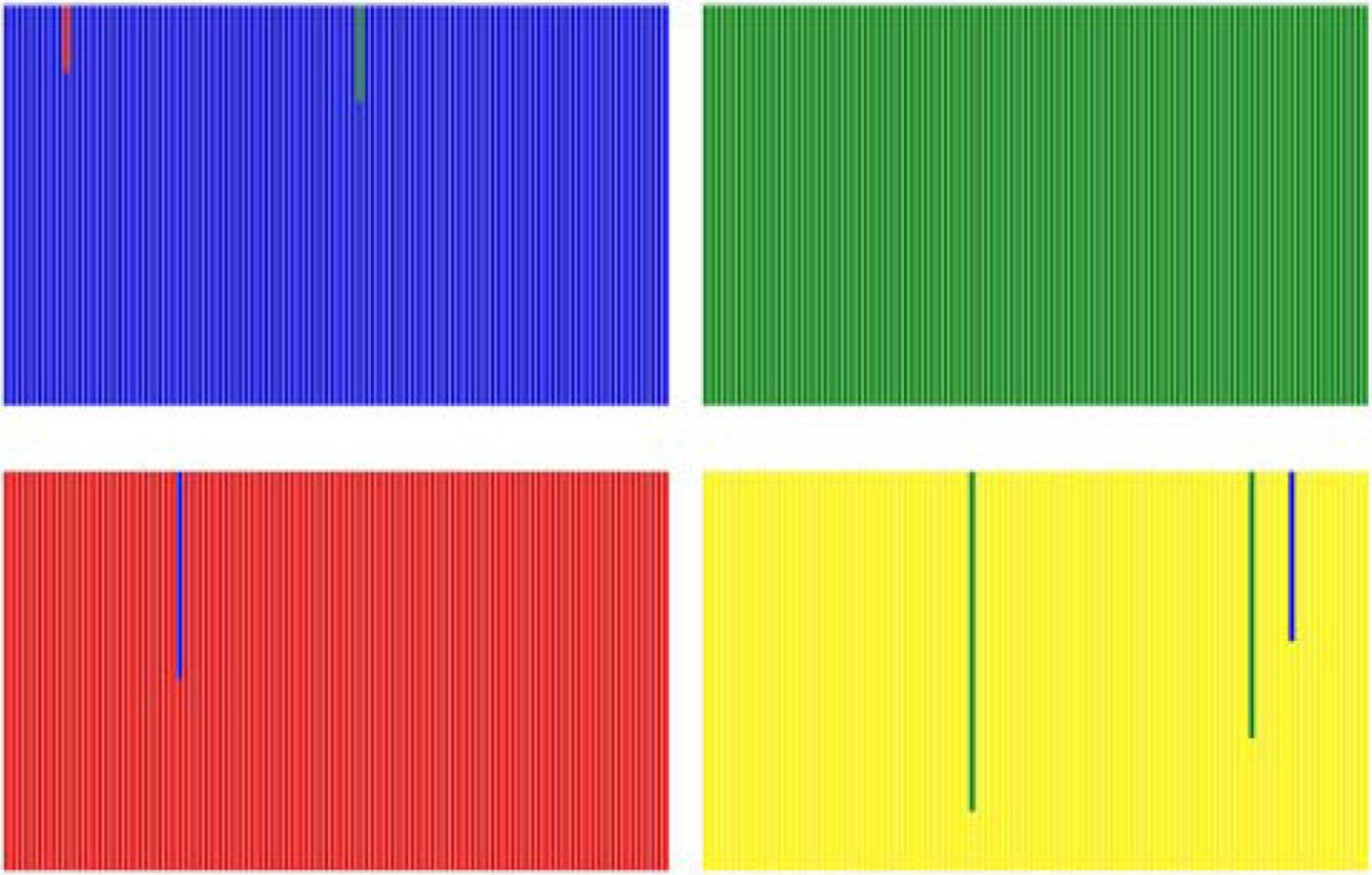
$-\pi_j q_{jj} = \sum_{i \neq j} \pi_i q_{ij}$

(Total flow out of j = Total flow into j)

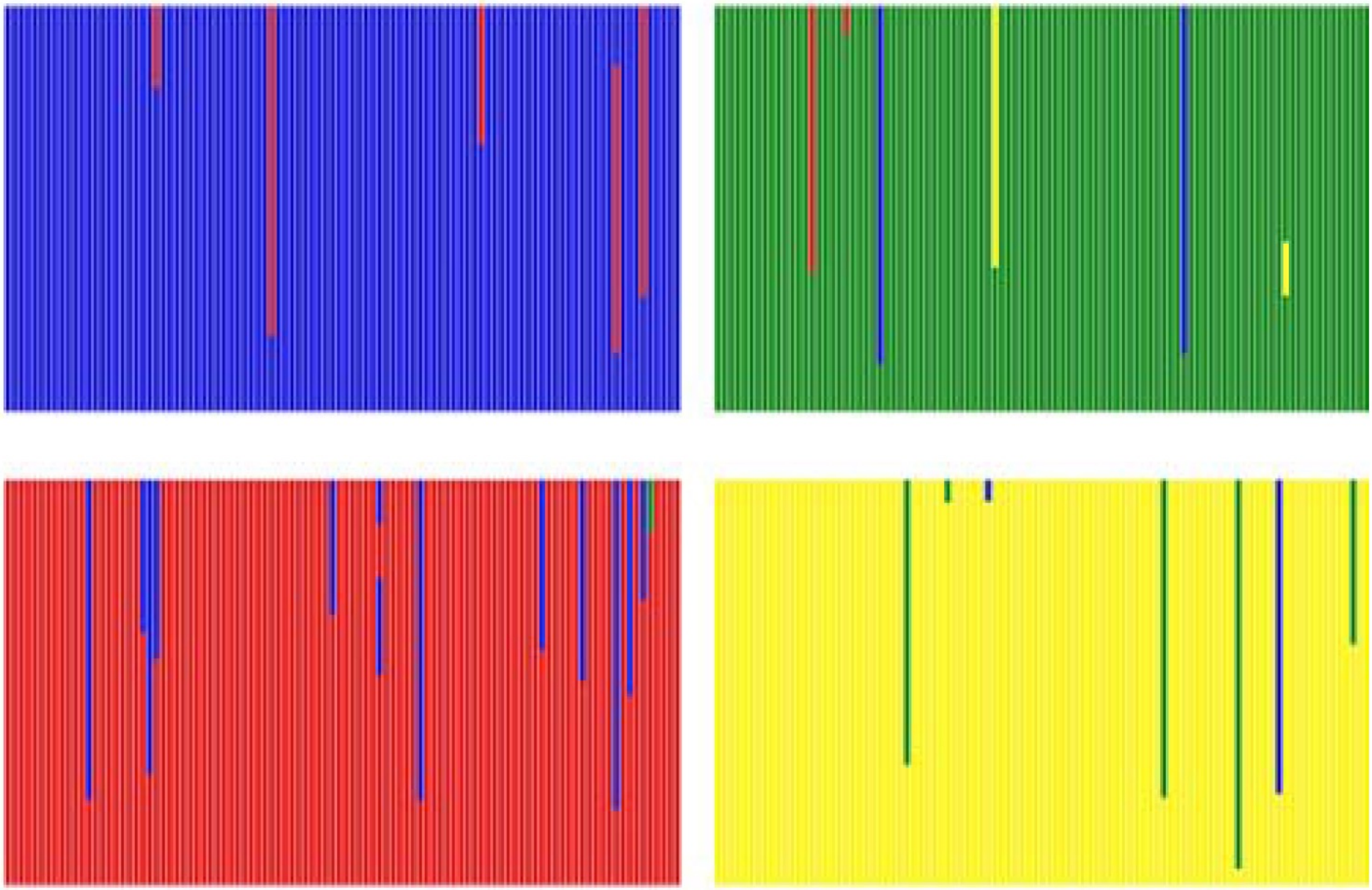
Example, DNA model: $t = 0$



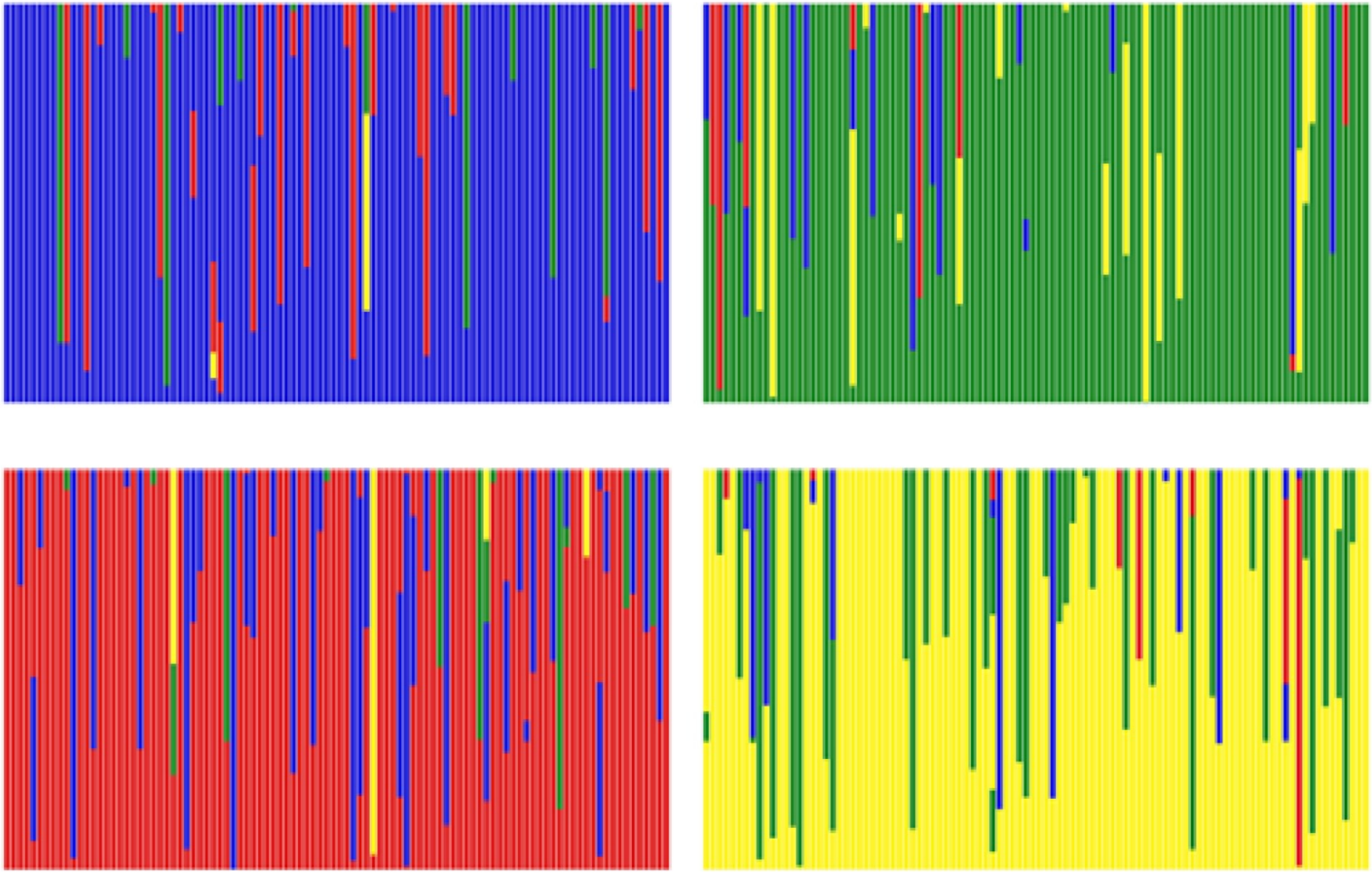
Example, DNA model: $t = 0.01$



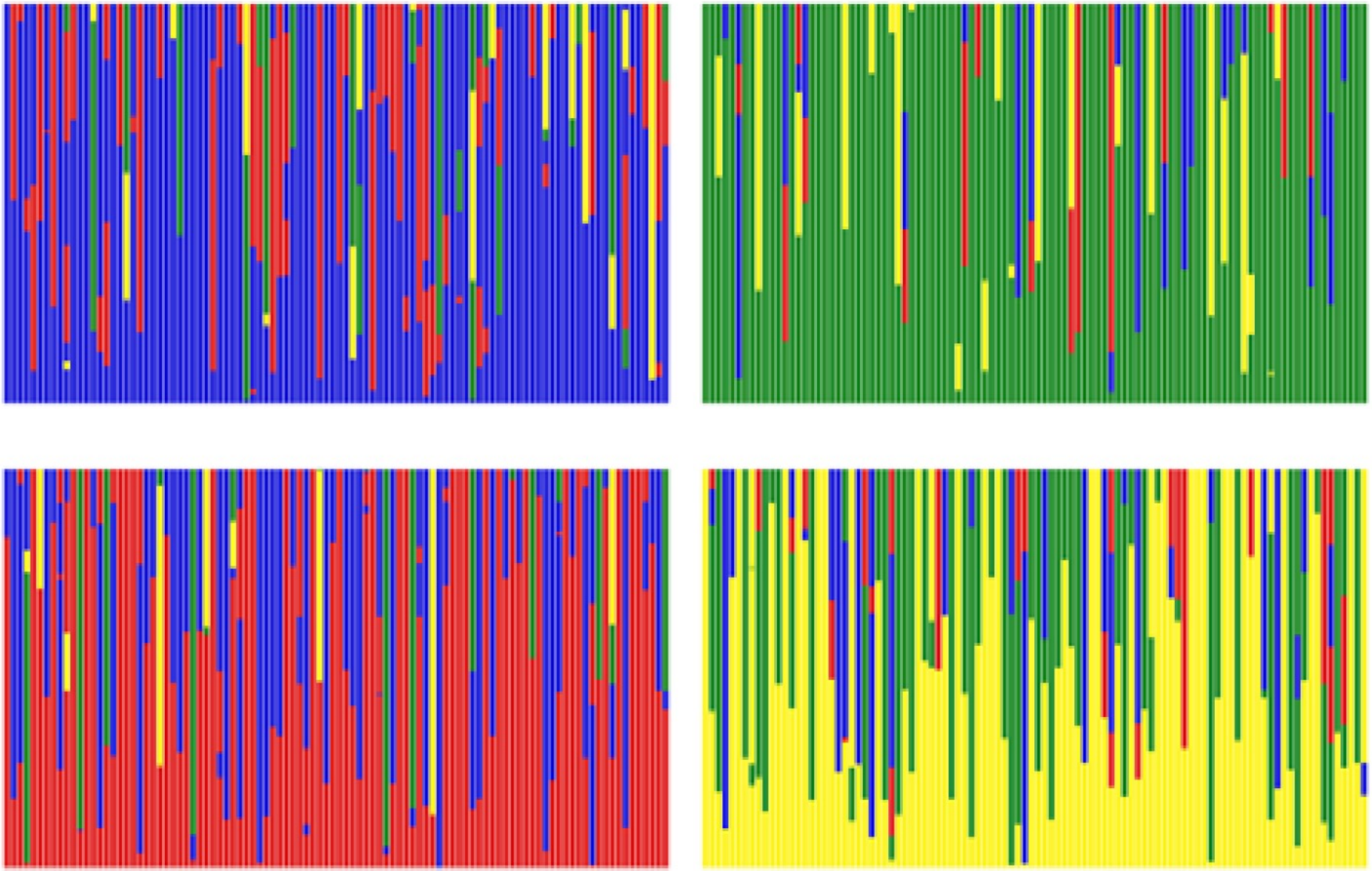
Example, DNA model: $t = 0.1$



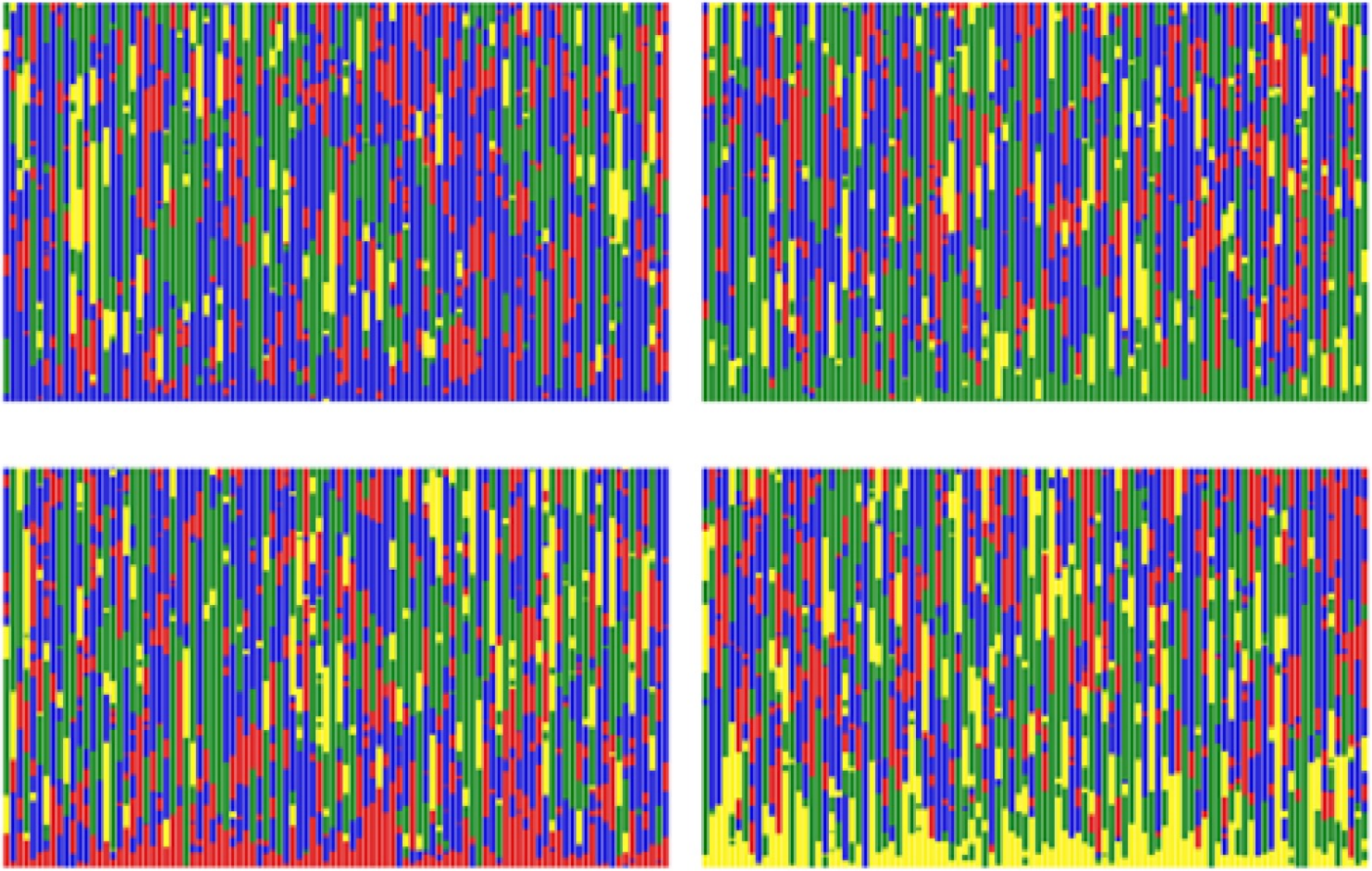
Example, DNA model: $t = 0.5$



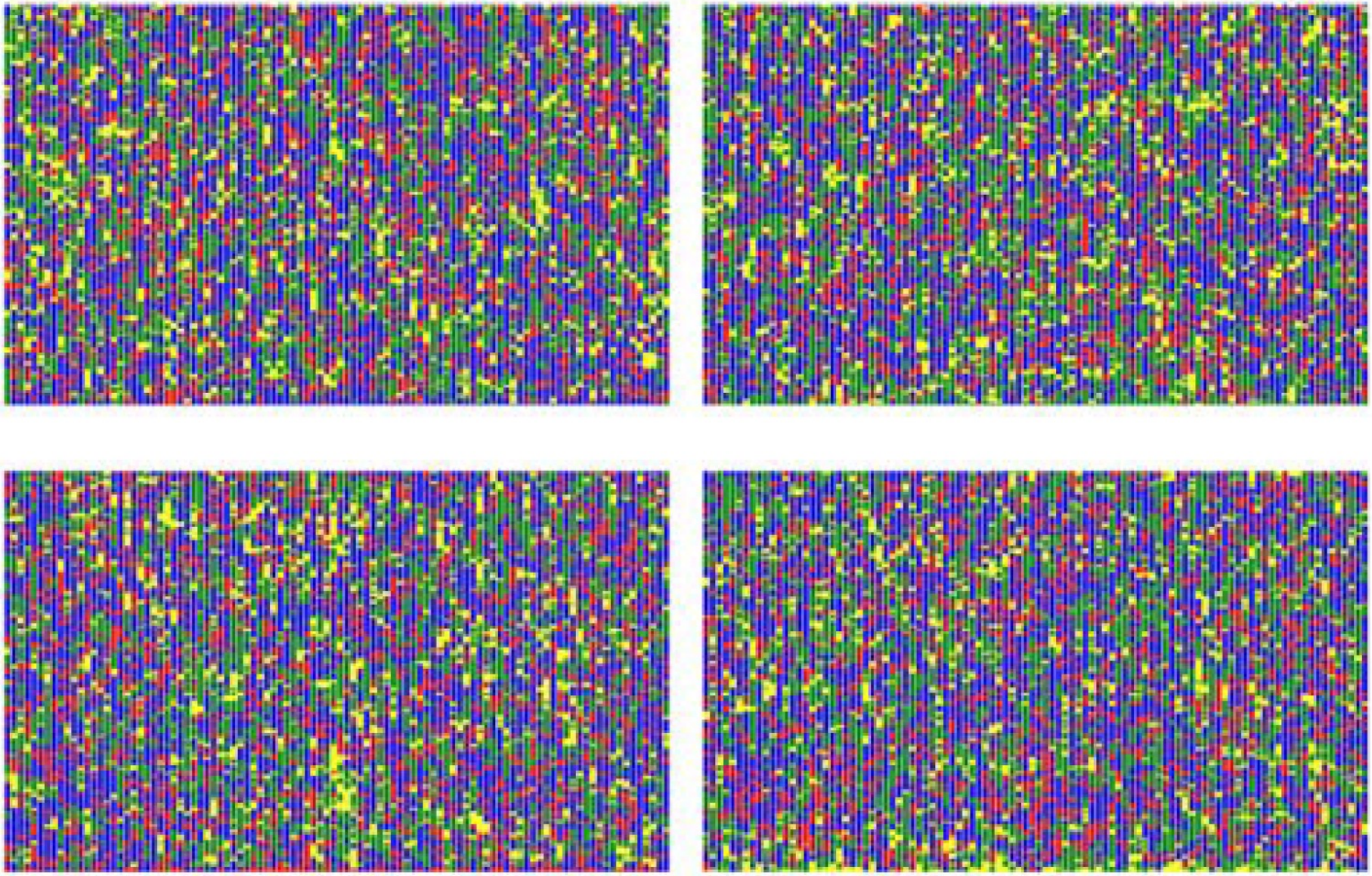
Example, DNA model: $t = 1$



Example, DNA model: $t = 10$



Example, DNA model: $t = 100$



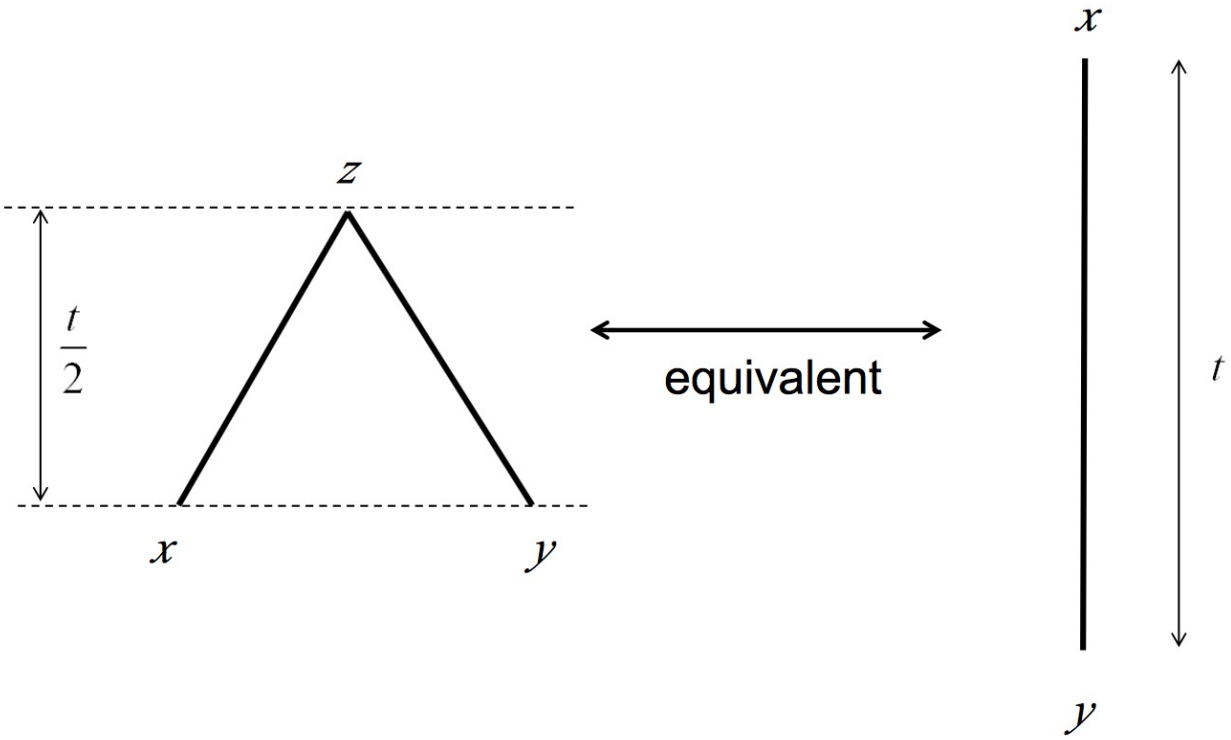
Time reversibility

Markov process is *time-reversible* if and only if

$\pi_i q_{ij} = \pi_j q_{ji}$ $(i \neq j)$

(In steady state: flow $i \to j$ = flow $j \to i$)

Equivalently: $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$ $(i \neq j)$



Time reversibility

If reversibility is assumed:

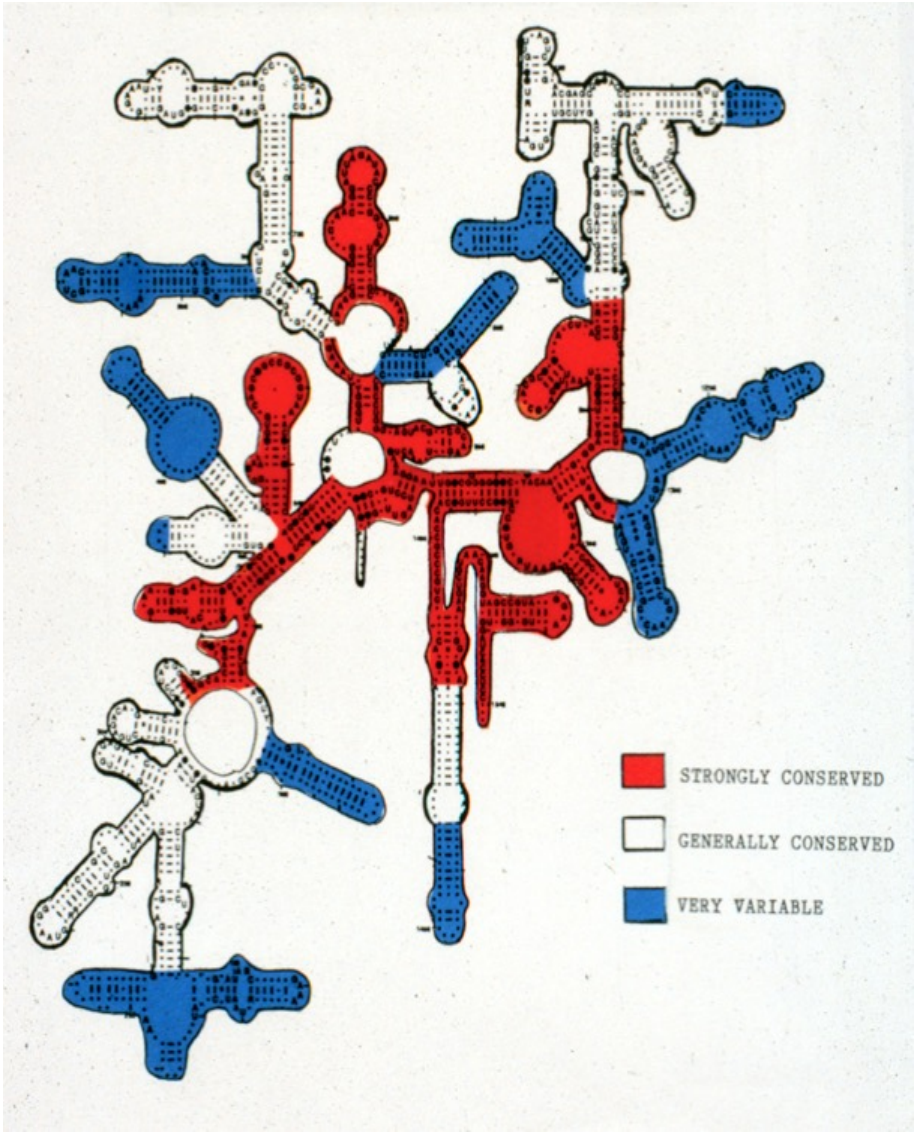
$q_{ij} = s_{ij}\pi_j$, where $s_{ij} = s_{ji}$ is exchangeability between *i* and *j*

Q is described by 9 independent parameters (GTR, Tavare 1986):

$$Q = \begin{pmatrix} \bullet & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \bullet & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \bullet & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \bullet \end{pmatrix} = \begin{pmatrix} \bullet & a & b & c \\ a & \bullet & d & e \\ b & d & \bullet & f \\ c & e & f & \bullet \end{pmatrix} \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

Model with no reversibility constraint: UNREST (Yang 1994)

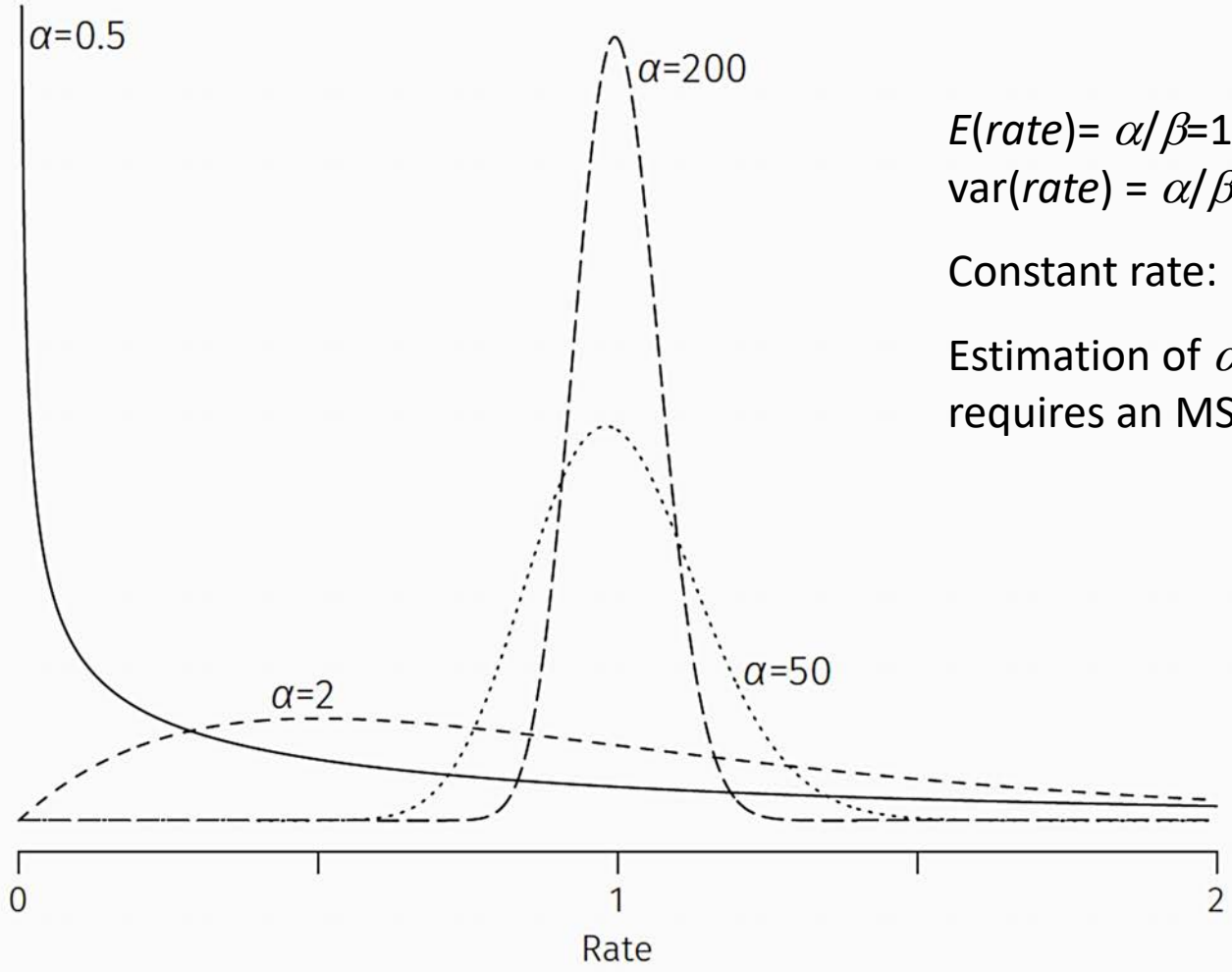
Across-sites rate variability



Small subunit ribosomal RNA (18S or 16S)

Across-sites rate variability

Can be modeled using the Γ-distribution with α = β



$E(rate) = \alpha/\beta = 1$

$var(rate) = \alpha/\beta^2 = 1/\alpha$

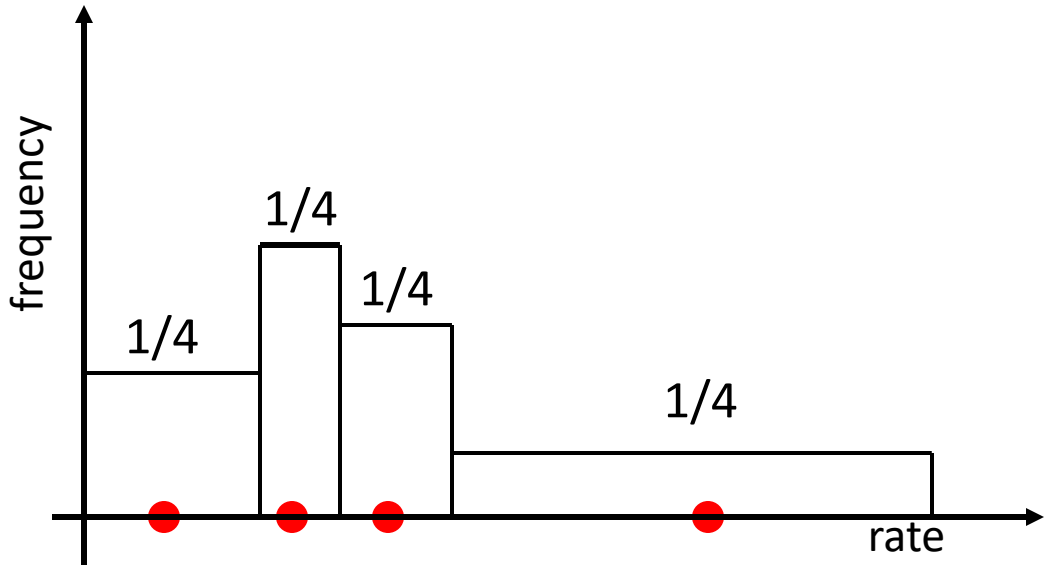
Constant rate: $\alpha \to \infty$

Estimation of α requires an MSA

The gamma distribution has no biological justification, it was chosen for its convenience.

Across-sites rate variability

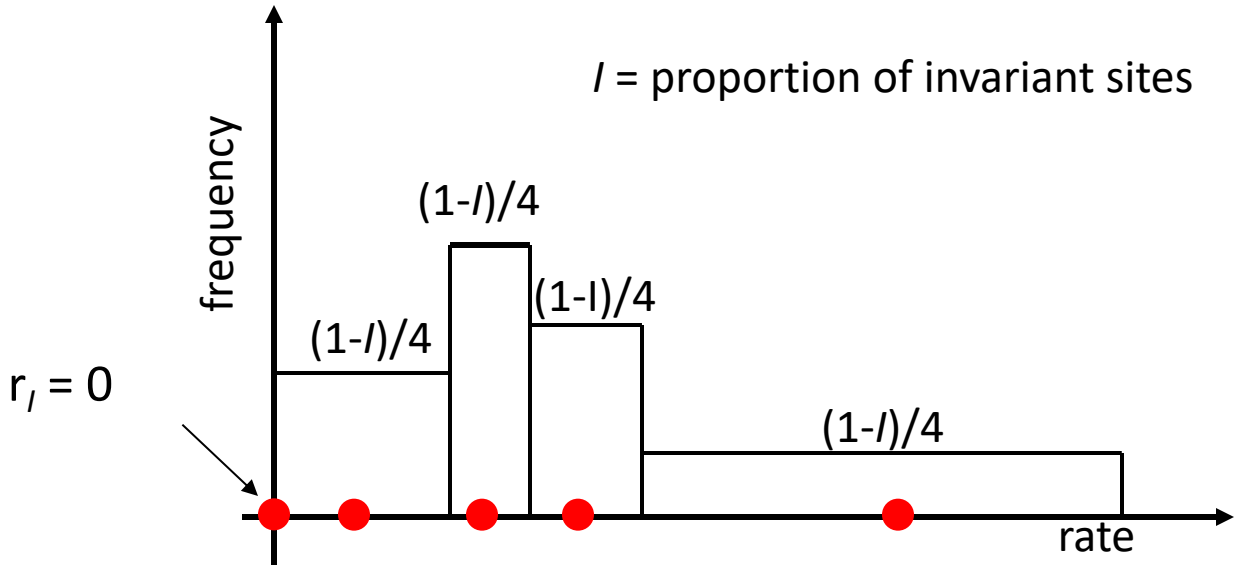
The Γ-distribution is simplified by discretization, for example with 4 classes of equal weight:



Across-sites rate variability

Γ + I model allows a proportion of invariable sites

I should be estimated from the data



How many discrete categories?

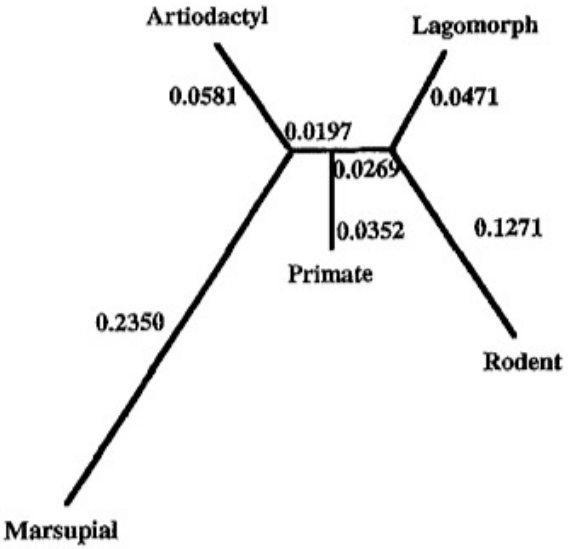


Fig. 3. The maximum likelihood tree for the five orders of mammals from the α and β globin genes (570 bp). The F84 + Γ model was assumed. Branch lengths are measured by the average numbers of nucleotide substitutions per site.

as the F84 + Γ and F84 + dG4 models: the other two

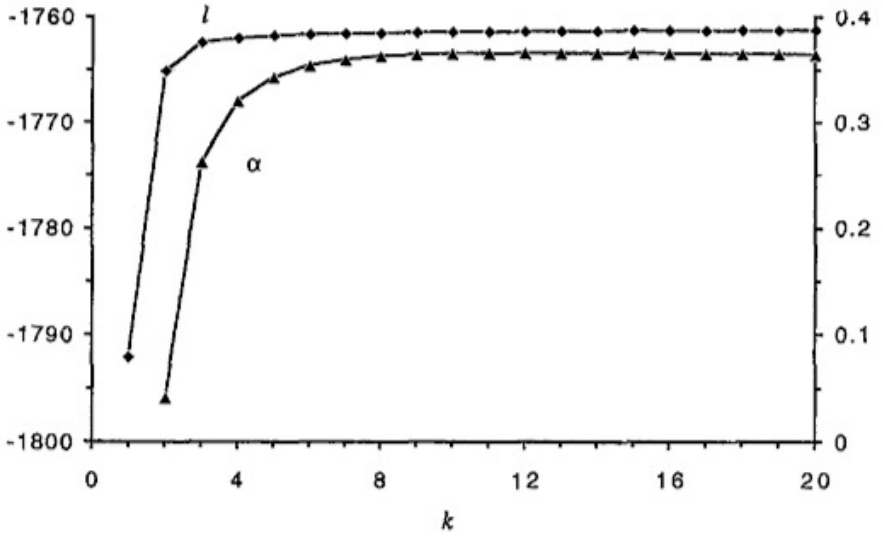


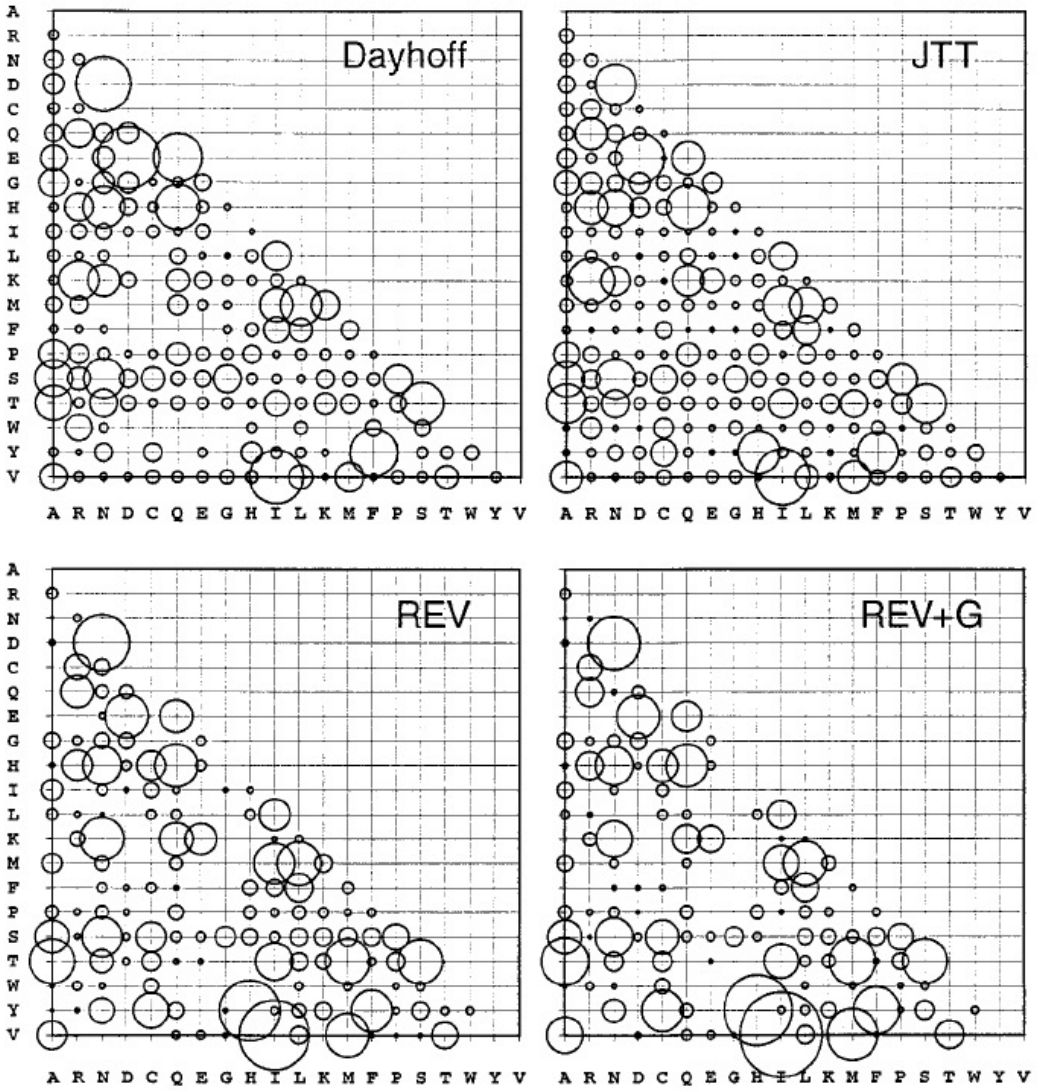
Fig. 4. Likelihood values and estimates of the α parameter as functions of k, the number of categories in the discrete gamma model. The α and β globin genes for the five mammalian orders (570 bp) are analyzed, assuming the best tree (Fig. 3) and the F84 + dG model. The average nucleotide frequencies are $\pi_T = 0.2200$, $\pi_C = 0.2449$, $\pi_A = 0.2761$, and $\pi_G = 0.2590$, with $\ell_{max} = -1{,}579.76$. When $k = \infty$, that is, with the F84 + Γ model, $\ell = -1{,}761.17$ and $\hat{\alpha} = 0.360$.

From Yang 1994

Table 1. Maximum likelihood estimates of the α parameter[a]

Sequences	Species	$\hat{\alpha}$	Refs
Nuclear genes			
α- and β-globin genes, positions 1 and 2	5 mammals	0.36	10,23
Albumin genes, all positions	5 vertebrates	1.05	44
Insulin genes, all positions	5 vertebrates	0.40	44
c-myc genes, all positions	5 vertebrates	0.47	44
Prolactin genes, all positions	5 vertebrates	1.37	44
16S-like rRNAs, stem region	5 species	0.29	45
16S-like rRNAs, loop region	5 species	0.58	45
$\psi\eta$-globin pseudogenes	6 primates	0.66	23
Viral genes			
Hepatitis B virus genomes	13 variants	0.26	46
Mitochondrial genes			
12S rRNAs	9 rodents	0.16	22
895-bp mtDNAs	9 primates	0.43	10
Positions 1 and 2 of 13 genes[b]	11 vertebrates	0.13–0.95	28
Position 1 of four genes	6 primates	0.18	19
Position 2 of four genes	6 primates	0.08	19
Position 3 of four genes	6 primates	1.58	19
D-loop region of mtDNAs[c]	25 humans	0.17	12
Protein sequences			
Mitochondrial cytochrome *b*	16 deuterostomes	0.44	12

Empirical models for proteins



Empirical models for proteins

JTT

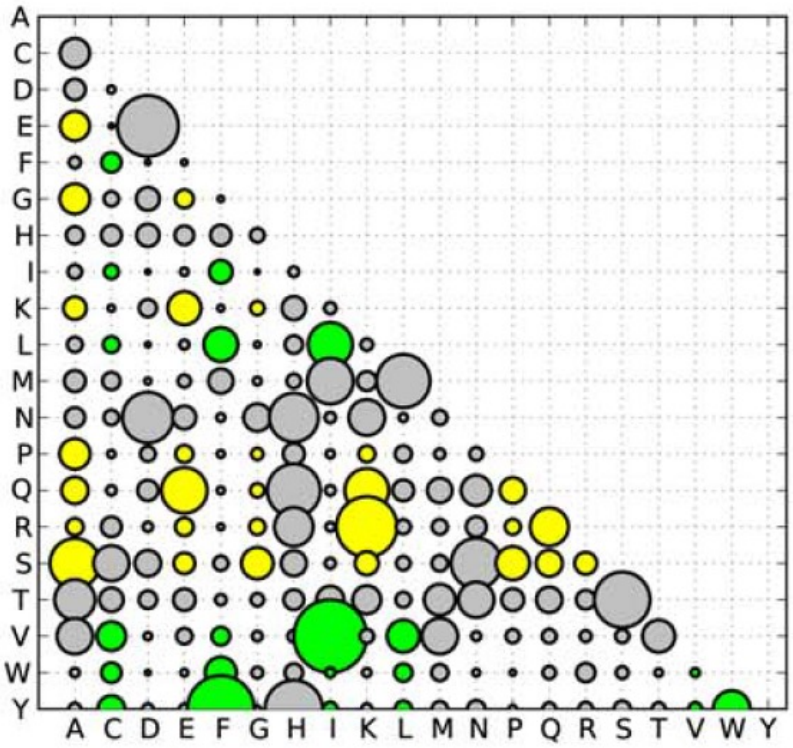
WAG

LG

HIV

Order/IDP

α/β



Q-matrix

AA stationary frequencies: π_i

+ F option: estimate frequencies from data