Stochastic processes and Markov chains (part II)

Wessel van Wieringen
w.n.van.wieringen@vu.nl

Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

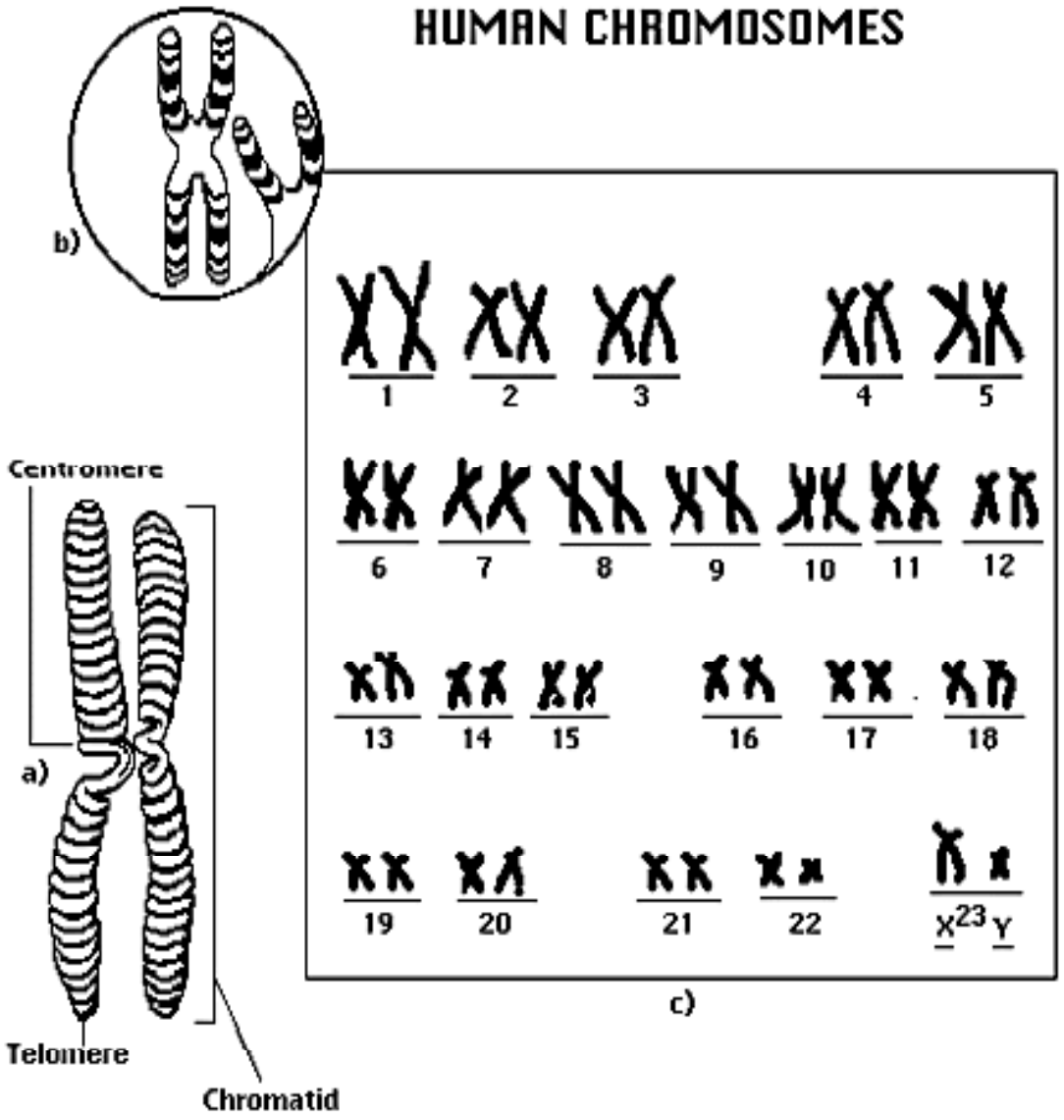
VU medisch centrum

Example

Example

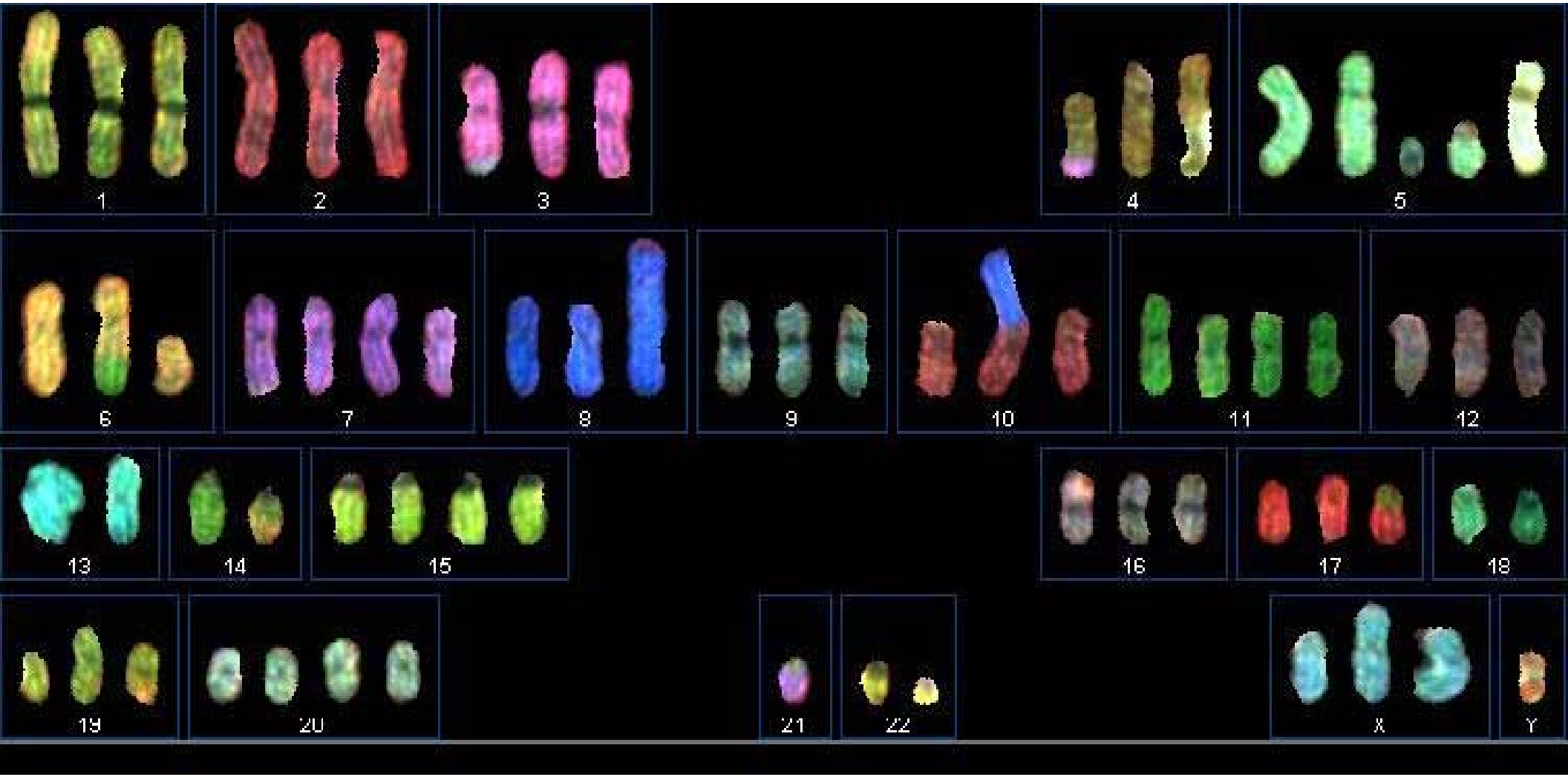
DNA copy number of a genomic segment is simply the number of copies of that segment present in the cell under study.

Healthy normal cell:
chr 1 : 2
…
chr 22 : 2
chr X : 1 or 2
chr Y : 0 or 1



Example

Chromosomes of a tumor cell



Technique: SKY

Example

The DNA copy number is often categorized into:

- L : *loss* : < 2 copies
- N : *normal* : 2 copies
- G : *gain* : > 2 copies

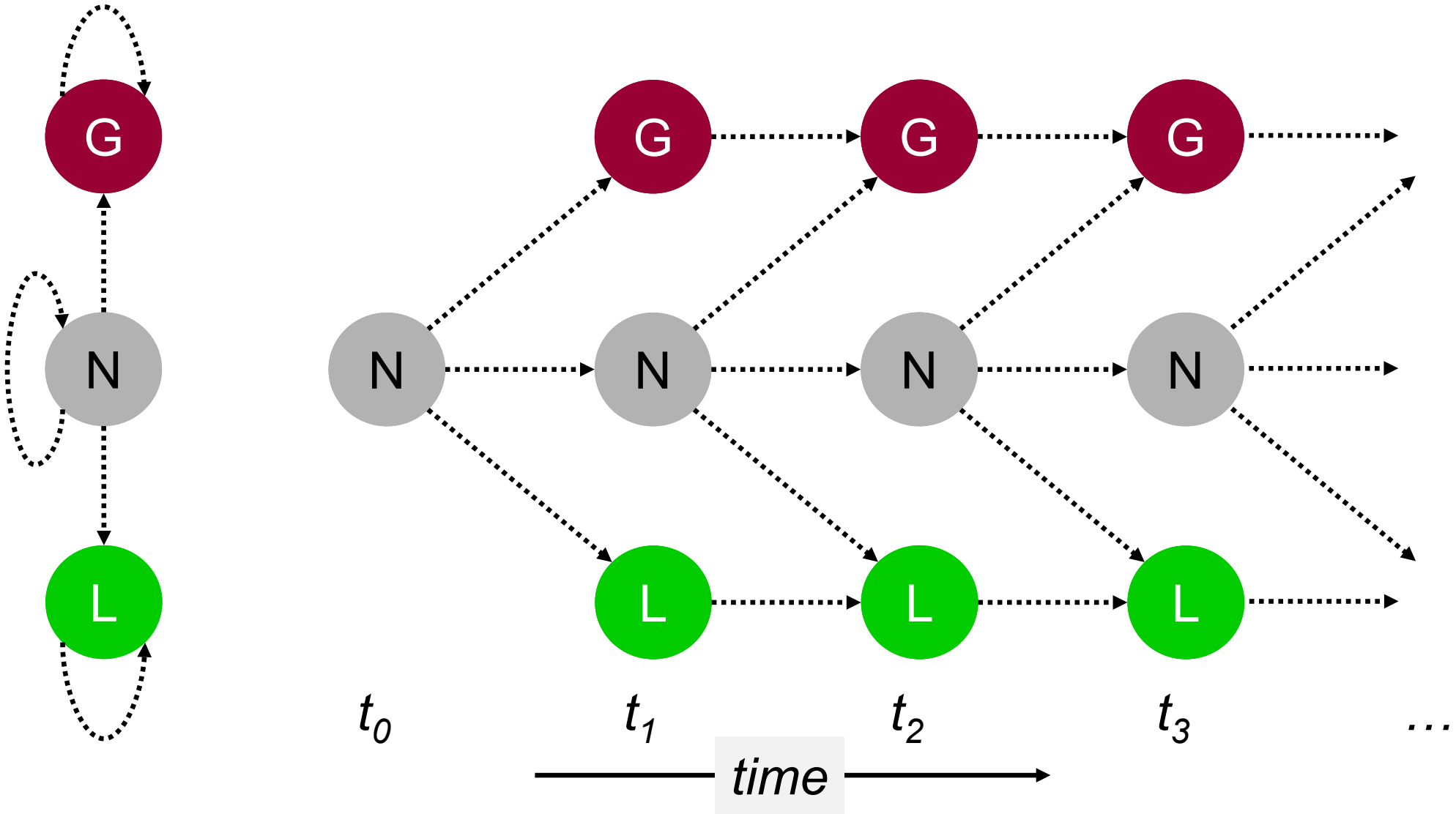
In cancer:

- The number of DNA copy number aberrations accumulates with the progression of the disease.
- DNA copy number aberrations are believed to be irreversible.

Let us model the accumulation process of DNA copy number aberrations.

Example

State diagram for the accumulation process of a locus.



Example

The associated initial distribution:

$$\boldsymbol{\pi} = (0, 1, 0)^T$$

and, associated transition matrix:

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1-\alpha-\beta & \beta \\ 0 & 0 & 1 \end{pmatrix}$$

with parameter constraints:

$$0 < \alpha + \beta < 1, \alpha > 0, \beta > 0$$

Example

Calculate the probability of a loss, normal and gain at this locus after *p* generations:

$$\begin{pmatrix} P(X_p = L) \\ P(X_p = N) \\ P(X_p = G) \end{pmatrix} = \boldsymbol{\pi}^T \mathbf{P}^p = \begin{pmatrix} (1-\alpha-\beta)^p \\ \alpha \sum_{t=0}^{p-1} (1-\alpha-\beta)^t \\ \beta \sum_{t=0}^{p-1} (1-\alpha-\beta)^t \end{pmatrix}$$

Using:

$$\sum_{t=0}^{\infty} (1-c)^t = \frac{1}{c}$$

These probabilities simplify to, e.g.:

$$P(X_p = L) = \frac{\alpha}{\alpha+\beta}[1-(1-\alpha-\beta)^p]$$

Example

In practice, a sample is only observed once the cancer has already developed. Hence, the number of generations *p* is unknown. This may be accommodated by modeling *p* as being Poisson distributed:

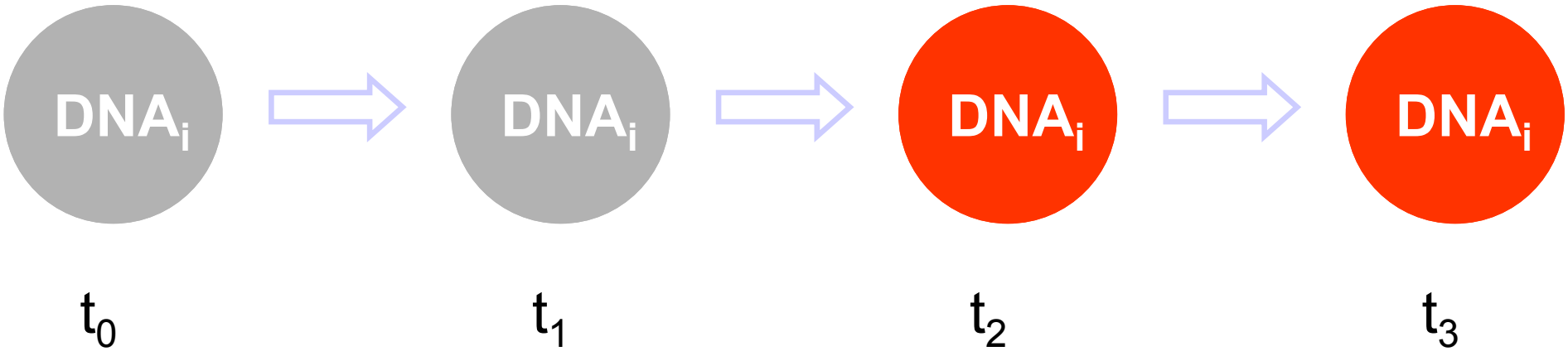
$$P(Y = p) = \lambda^p \exp(-\lambda)/p!$$

This yields, e.g.:

$$\begin{aligned} P(X = N) &= \sum_{p=0}^{\infty} P(X = L \,|\, Y = p)\, P(Y = p) \\ &= \sum_{p=0}^{\infty} (1 - \alpha - \beta)^p \, \lambda^p \, \exp(-\lambda)/p! \end{aligned}$$

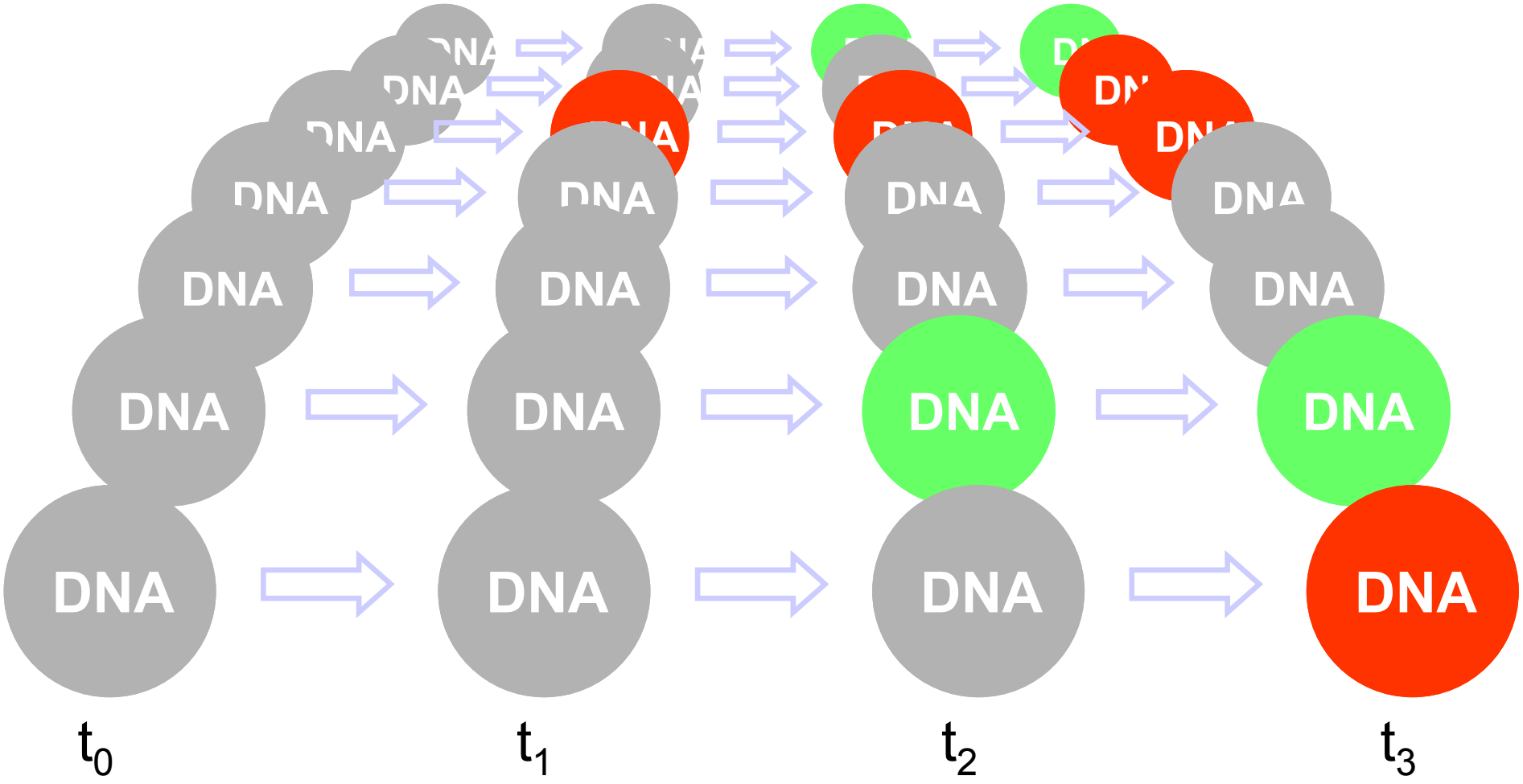
Example

So far, we only considered one locus. Hence:



Example

For multiple loci:



Example

Multiple loci → multivariate problem.

Complications:
- *p* unknown,
- loci not independent.

Solution:
- *p* random,
- assume particular dependence structure.

After likelihood formulation and parameter estimation:
- identify most aberrated loci,
- reconstruct time of onset of cancer.

Stationary distribution

Stationary distribution

We generated a DNA sequence of 10.000.000 bases long in accordance with a 1st order Markov chain.

For stretches DNA ever longer and ever farther away from the first base pair we calculated the nucleotide %.

bp	bp	%A	%C	%G	%T
1001	5e+03	0.014	0.069	0.262	0.655
5001	2e+04	0.045	0.239	0.228	0.488
20001	1e+05	0.144	0.319	0.211	0.327
100001	5e+05	0.158	0.295	0.205	0.342
500001	2e+06	0.142	0.284	0.216	0.357
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
		0.150	0.280	0.220	0.350

Stationary distribution

Hence, after a while an “equilibrium” sets in. Not necessarily a fixed state or pattern, but:

> the proportion of a time period that is spent in a particular state converges to a limit value.

The limit values of all states form the ***stationary distribution*** of the Markov process, denoted by:

$$\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_S)^T \qquad \text{with} \sum_{k=1}^{S} \varphi_k = 1$$

Stationary distribution

For a stationary process, it holds that

$$P(X_t=E_i) = \varphi_i$$

for all *t* and *i*.

In particular:

$$P(X_t=E_i) = \varphi_i = P(X_{t+1}=E_i).$$

Of course, this does *not* imply:

$$P(X_t=E_i, X_{t+1}=E_i) = \varphi_i \varphi_i$$

as this ignores the 1st order Markov dependency.

Stationary distribution

The stationary distribution is associated with the first-order Markov process, parameterized by ($\boldsymbol{\pi}$, $\mathbf{P}$).

Question

How do $\boldsymbol{\varphi}$ and ($\boldsymbol{\pi}$, $\mathbf{P}$) relate?

Hereto, recall:

$$\varphi_i = P(X_t = E_i)$$

Stationary distribution

We then obtain the following relation:

$$\varphi_i \quad = \quad P(X_{t+1} = E_i)$$

definition

Stationary distribution

We then obtain the following relation:

$$\begin{aligned} \varphi_i &= P(X_{t+1} = E_i) \\ &= \sum_{k=1}^{S} P(X_{t+1} = E_i, X_t = E_k) \end{aligned}$$

use the fact that

P(*A*, *B*) + P(*A*, B^C) = P(*A*)

Stationary distribution

We then obtain the following relation:

$$
\begin{aligned}
\varphi_i &= P(X_{t+1} = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+1} = E_i, X_t = E_k) \\
&= \sum_{k=1}^{S} P(X_{t+1} = E_i \mid X_t = E_k)\, P(X_t = E_k)
\end{aligned}
$$

use the definition of conditional probability:

$P(A, B) = P(A \mid B)\, P(B)$

Stationary distribution

We then obtain the following relation:

$$
\begin{aligned}
\varphi_i &= P(X_{t+1} = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+1} = E_i, X_t = E_k) \\
&= \sum_{k=1}^{S} P(X_{t+1} = E_i \,|\, X_t = E_k)\, P(X_t = E_k) \\
&= \sum_{k=1}^{S} p_{ki}\, \varphi_k
\end{aligned}
$$

Stationary distribution

We then obtain the following relation:

$$\begin{aligned}
\varphi_i &= P(X_{t+1} = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+1} = E_i, X_t = E_k) \\
&= \sum_{k=1}^{S} P(X_{t+1} = E_i \,|\, X_t = E_k)\, P(X_t = E_k) \\
&= \sum_{k=1}^{S} p_{ki}\, \varphi_k
\end{aligned}$$

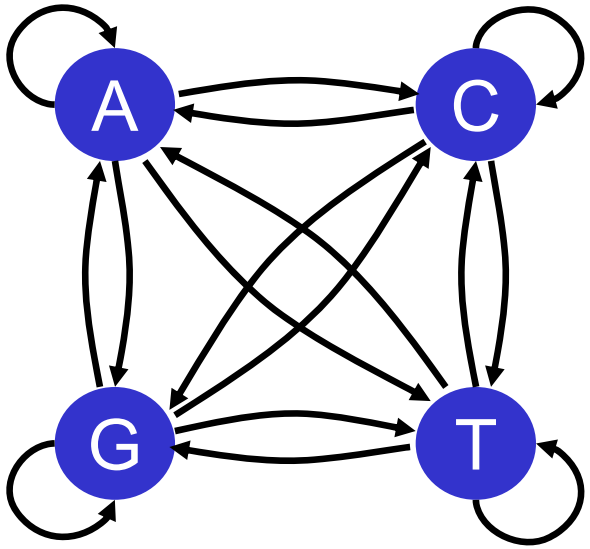
Thus:

$$\boldsymbol{\varphi}^T = \boldsymbol{\varphi}^T \mathbf{P} \longrightarrow \text{Eigenvectors!}$$

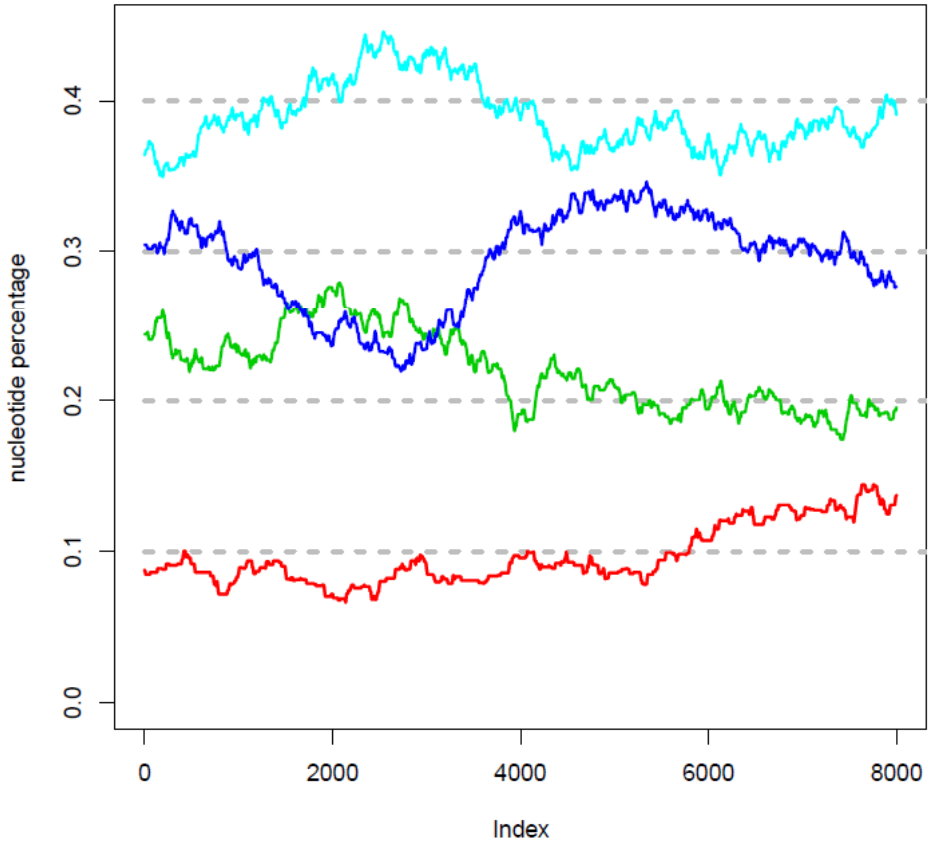
Stationary distribution

Theorem

Irreducible, aperiodic Markov chains with a finite state space *S* have a stationary distribution to which the chain converges as $t \to \infty$.



$\varphi_A=0.1$, $\varphi_C=0.2$, $\varphi_G=0.3$, $\varphi_T=0.4$

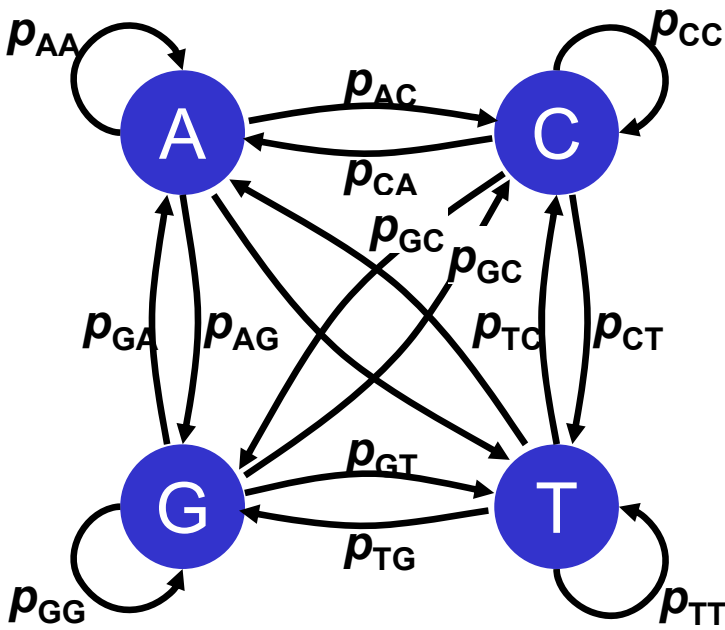


Stationary distribution

A Markov chain is ***aperiodic*** if there is no state that can only be reached at multiples of a certain period. E.g., state E_i only at t = 0, 3, 6, 9, et cetera.

Example of an aperiodic Markov chain

$$\mathbf{P} = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix} \end{array}$$

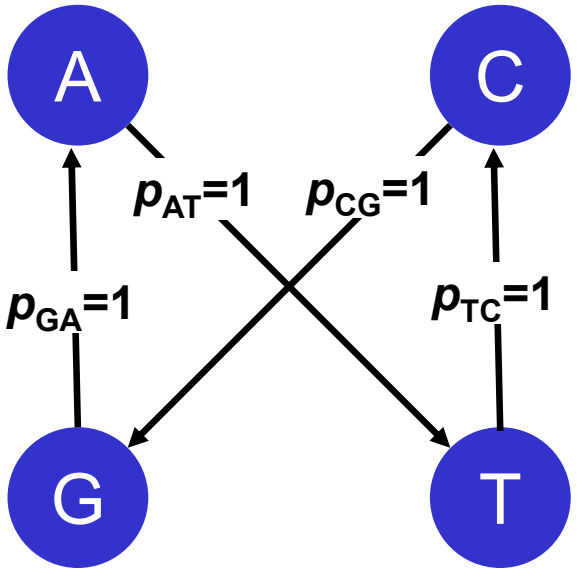


...**ATCGATCGATCGATCG**...

Stationary distribution

Example of a periodic Markov chain

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$



…**ATCGATCGATCGATCG**…

The fraction of time spent in **A** (roughly $\varphi_{\mathbf{A}}$):

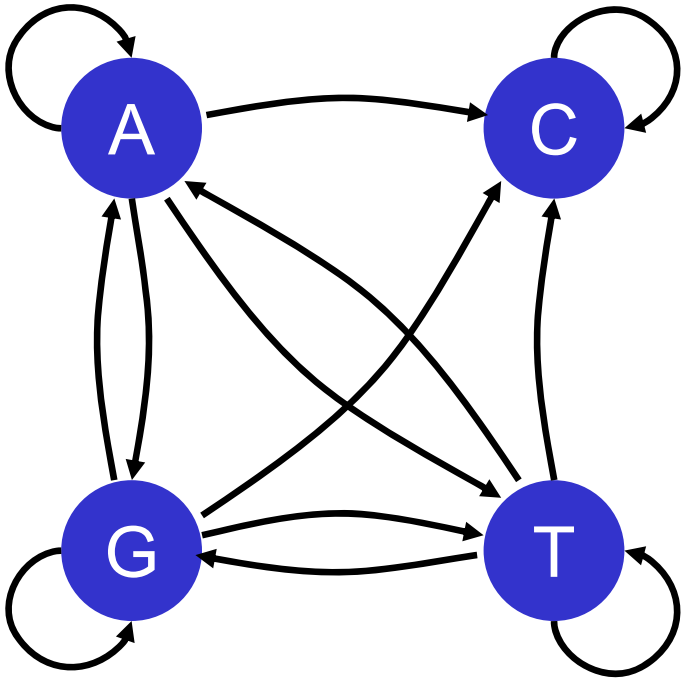
$$P(X_{t+1000} = \mathbf{A}) = ¼$$

whereas $P(X_{t+1000} = \mathbf{A} \mid X_4 = \mathbf{A}) = 1$.

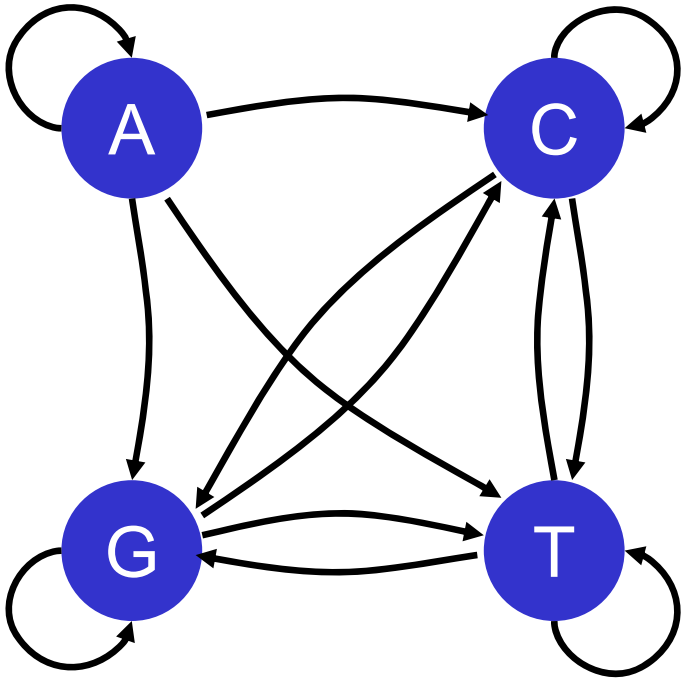
Stationary distribution

A Markov chain is ***irreducible*** if every state can (in principle) be reached (after enough time has passed) from every other state.

Examples of a reducible Markov chain



C is an absorbing state.



A will not be reached.

Stationary distribution

How do we find the stationary distribution?

We know the stationary distribution satisfies:

$$\boldsymbol{\varphi}^T = \boldsymbol{\varphi}^T \mathbf{P} \qquad (*)$$

and

$$\varphi_1 + \varphi_2 + \ldots + \varphi_S = 1$$

We thus have *S*+1 equations for *S* unknowns: one of the equations in (*) can be dropped (which is irrelevant), and the system of *S* remaining equations needs to be solved.

Stationary distribution

Consider the transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.35 & 0.65 \\ 0.81 & 0.19 \end{pmatrix}$$

In order to find the stationary distribution we need to solve the following system of equations:

$$\begin{aligned} 0.35\,\varphi_1 + 0.81\,\varphi_2 &= \varphi_1 \\ \varphi_1 + \varphi_2 &= 1 \end{aligned}$$

This yields: $(\varphi_1, \varphi_2)^T = (0.5547945, 0.4452055)^T$

Stationary distribution

On the other hand, for *n* large:

$$P(X_{t+n}=E_j \mid X_t=E_i) = \varphi_j$$

is independent of *i*. Or, $p_{ij}^{(n)} = \varphi_j$.

Hence, the *n*-step transition matrix $\mathbf{P}^{(n)}$ has identical rows:

$$\lim_{n\to\infty} \mathbf{P}^{(n)} = \lim_{n\to\infty} \mathbf{P}^{n} = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_S \\ \varphi_1 & \varphi_2 & \cdots & \varphi_S \\ \vdots & \vdots & & \vdots \\ \varphi_1 & \varphi_2 & \cdots & \varphi_S \end{pmatrix}$$

This motivates a numerical way to find the stationary distribution.

Stationary distribution

Same example as before:

$$\mathbf{P} = \begin{pmatrix} 0.35 & 0.65 \\ 0.81 & 0.19 \end{pmatrix}$$

with stationary distribution: $(\varphi_1, \varphi_2)^T = (0.5547945, 0.4452055)^T$

Then:

$$\mathbf{P}^{(2)} = \begin{pmatrix} 0.35 & 0.65 \\ 0.81 & 0.19 \end{pmatrix} \begin{pmatrix} 0.35 & 0.65 \\ 0.81 & 0.19 \end{pmatrix}$$

$$= \begin{pmatrix} 0.35 \times 0.35 + 0.65 \times 0.81 & 0.35 \times 0.65 + 0.65 \times 0.19 \\ 0.81 \times 0.35 + 0.19 \times 0.81 & 0.81 \times 0.65 + 0.19 \times 0.19 \end{pmatrix}$$

matrix multiplication (“rows times columns”)

Stationary distribution

Thus:

$$\mathbf{P}^{(2)} = \begin{pmatrix} 0.6490 & 0.3510 \\ 0.4374 & 0.5626 \end{pmatrix}$$

In similar fashion we obtain:

$$\mathbf{P}^{(5)} = \begin{pmatrix} 0.5456249 & 0.4543751 \\ 0.5662212 & 0.4337788 \end{pmatrix}$$

$$\mathbf{P}^{(20)} = \begin{pmatrix} 0.5547946 & 0.4452054 \\ 0.5547944 & 0.4452056 \end{pmatrix}$$

Convergence of the stationary distribution

Convergence to the stationary distribution

Define the vector $\mathbf{1} = (1,\ldots,1)^T$.

We have already seen:

$$\mathbf{P}^n = \mathbf{1}\,\boldsymbol{\varphi}^T \qquad \text{for large } n$$

Question

How fast does $\mathbf{P}^n$ go to $\mathbf{1}\,\boldsymbol{\varphi}^T$ as $n \to \infty$?

Answer

1) Use linear algebra
2) Calculate numerically

Convergence to the stationary distribution

Fact

The transition matrix **P** of a finite, aperiodic, irreducible Markov chain has an eigenvalue equal to 1 ($\lambda_1=1$), while all other eigenvalues are (in the absolute sense) smaller than one: $|\lambda_k| < 1$, k=2,…,3.

Focus on $\lambda_1=1$

We know $\boldsymbol{\varphi}^T \mathbf{P} = \boldsymbol{\varphi}^T$ for the stationary distribution. Hence, $\boldsymbol{\varphi}$ is the left eigenvector of eigenvalue $\lambda=1$.

Also, row sums of **P** equal **1**: $\mathbf{P}\,\mathbf{1} = \mathbf{1}$. Hence, **1** is a right eigenvector of eigenvalue λ=1.

Convergence to the stationary distribution

The ***spectral decomposition*** of a square matrix **P** is given by:

$$\mathbf{P} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^{-1}$$

where:

$\mathbf{D}$ diagonal matrix containing the eigenvalues,

$\mathbf{V}$ columns contain the corresponding eigenvectors.

In case of **P** is symmetric, $\mathbf{V}$ is orthogonal: $\mathbf{V}^{-1} = \mathbf{V}^{T}$

Then:

$$\mathbf{P} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^{T}$$

Convergence to the stationary distribution

The ***spectral decomposition*** of **P**, reformulated:

$$\mathbf{P} = \sum_{k=1}^{S} \lambda_k \, \mathbf{r}_k \, \boldsymbol{\ell}_k^T$$

eigenvalues — λ_k

right eigenvectors — $\mathbf{r}_k$

left eigenvectors — $\boldsymbol{\ell}_k^T$

The eigenvectors are normalized:

$$\begin{aligned} \mathbf{r}_k^T \, \boldsymbol{\ell}_k &= 1, && \text{and} \\ \mathbf{r}_{k_1}^T \, \boldsymbol{\ell}_{k_2} &= 0 && \text{if } k_1 \neq k_2 \end{aligned}$$

Convergence to the stationary distribution

We now obtain the spectral decomposition of the *n*-step transition matrix $\mathbf{P}^n$. Hereto observe that:

$$\mathbf{P}^n \mathbf{r}_k = \mathbf{P}^{n-1} \left(\sum_{k_0=1}^{S} \lambda_{k_0} \mathbf{r}_{k_0} \boldsymbol{\ell}_{k_0}^T \right) \mathbf{r}_k$$

plug in the spectral decomposition of **P**

Convergence to the stationary distribution

We now obtain the spectral decomposition of the *n*-step transition matrix $\mathbf{P}^n$. Hereto observe that:

$$\mathbf{P}^n \mathbf{r}_k = \mathbf{P}^{n-1} \left(\sum_{k_0=1}^{S} \lambda_{k_0} \mathbf{r}_{k_0} \boldsymbol{\ell}_{k_0}^T \right) \mathbf{r}_k$$

$$= \mathbf{P}^{n-1} \left(\sum_{k_0=1}^{S} \lambda_{k_0} \mathbf{r}_{k_0} [\mathbf{r}_k^T \boldsymbol{\ell}_{k_0}]^T \right)$$

bring the right eigenvector in the sum,
and use the properties of the transpose operator

Convergence to the stationary distribution

We now obtain the spectral decomposition of the *n*-step transition matrix $\mathbf{P}^n$. Hereto observe that:

$$
\begin{aligned}
\mathbf{P}^n \mathbf{r}_k &= \mathbf{P}^{n-1} \Big(\sum_{k_0=1}^{S} \lambda_{k_0} \mathbf{r}_{k_0} \boldsymbol{\ell}_{k_0}^T \Big) \mathbf{r}_k \\
&= \mathbf{P}^{n-1} \Big(\sum_{k_0=1}^{S} \lambda_{k_0} \mathbf{r}_{k_0} [\mathbf{r}_k^T \boldsymbol{\ell}_{k_0}]^T \Big) \\
&= \mathbf{P}^{n-1} \lambda_k \mathbf{r}_k
\end{aligned}
$$

the eigenvectors are normalized

Convergence to the stationary distribution

We now obtain the spectral decomposition of the *n*-step transition matrix $\mathbf{P}^n$. Hereto observe that:

$$
\begin{aligned}
\mathbf{P}^n \, \mathbf{r}_k &= \mathbf{P}^{n-1} \Big(\sum_{k_0=1}^{S} \lambda_{k_0} \, \mathbf{r}_{k_0} \, \boldsymbol{\ell}_{k_0}^T \Big) \mathbf{r}_k \\
&= \mathbf{P}^{n-1} \Big(\sum_{k_0=1}^{S} \lambda_{k_0} \, \mathbf{r}_{k_0} \, [\mathbf{r}_k^T \, \boldsymbol{\ell}_{k_0}]^T \Big) \\
&= \mathbf{P}^{n-1} \, \lambda_k \, \mathbf{r}_k \\
&= \lambda_k \, \mathbf{P}^{n-1} \, \mathbf{r}_k
\end{aligned}
$$

Convergence to the stationary distribution

Repeating this argument *n* times yields:

$$\begin{aligned} \mathbf{P}^n \, \mathbf{r}_k &= \lambda_k^n \, \mathbf{r}_k \\ \boldsymbol{\ell}_k^T \, \mathbf{P}^n &= \lambda_k^n \, \boldsymbol{\ell}_k^T \end{aligned}$$

Hence, $\boldsymbol{\ell}_k$ and $\mathbf{r}_k$ are left and right eigenvector with eigenvalue λ_k^n of $\mathbf{P}^n$. Thus:

$$\mathbf{P}^n = \sum_{k=1}^{S} \lambda_k^n \, \mathbf{r}_k \, \boldsymbol{\ell}_k^T$$

Convergence to the stationary distribution

Verify the spectral decomposition for $\mathbf{P}^2$:

$$
\begin{aligned}
\mathbf{P}^2 &= \left(\sum_{k=1}^{S} \lambda_k \, \mathbf{r}_k \, \boldsymbol{\ell}_k^T\right)\left(\sum_{k=1}^{S} \lambda_k \, \mathbf{r}_k \, \boldsymbol{\ell}_k^T\right) \\
&= \sum_{k_1=1}^{S}\sum_{k_2=1}^{S} \lambda_{k_1} \, \mathbf{r}_{k_1} \, \boldsymbol{\ell}_{k_1}^T \, \lambda_{k_2} \, \mathbf{r}_{k_2} \, \boldsymbol{\ell}_{k_2}^T \\
&= \sum_{k_1=1}^{S}\sum_{k_2=1}^{S} \lambda_{k_1} \, \lambda_{k_2} \, \mathbf{r}_{k_1} \, \boldsymbol{\ell}_{k_1}^T \, \mathbf{r}_{k_2} \, \boldsymbol{\ell}_{k_2}^T \\
&= \sum_{k_1=1}^{S}\sum_{k_2=1}^{S} \lambda_{k_1} \, \lambda_{k_2} \, \mathbf{r}_{k_1} \, (\mathbf{r}_{k_2}^T \, \boldsymbol{\ell}_{k_1})^T \, \boldsymbol{\ell}_{k_2}^T \\
&= \sum_{k=1}^{S} \lambda_k^2 \, \mathbf{r}_k \, \boldsymbol{\ell}_k^T
\end{aligned}
$$

Convergence to the stationary distribution

Use the spectral decomposition of $\mathbf{P}^n$ to show how fast $\mathbf{P}^n$ converges to $\mathbf{1}\,\boldsymbol{\varphi}^T$ as $n \to \infty$.

We know:

$\lambda_1=1$, $|\lambda_k| < 1$ for k=2, …,S, $\boldsymbol{\ell}_1 = \boldsymbol{\varphi}$ and $\mathbf{r}_1 = \mathbf{1}$

Then:

$$\mathbf{P}^n = 1^n\,\mathbf{1}\,\boldsymbol{\varphi}^T + \sum_{k=2}^{S} \lambda_k^n\,\mathbf{r}_k\,\boldsymbol{\ell}_k^T$$

Convergence to the stationary distribution

Expanding this:

$$\mathbf{P}^n = 1^n \mathbf{1}\boldsymbol{\varphi}^T + \sum_{k=2}^{S} \lambda_k^n \mathbf{r}_k \boldsymbol{\ell}_k^T$$

$$= \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (\varphi_1, \varphi_2, \ldots, \varphi_S) + \lambda_2^n \mathbf{r}_2 \boldsymbol{\ell}_2^T + \ldots + \lambda_S^n \mathbf{r}_S \boldsymbol{\ell}_S^T$$

$$= \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_S \\ \varphi_1 & \varphi_2 & \cdots & \varphi_S \\ \vdots & \vdots & & \vdots \\ \varphi_1 & \varphi_2 & \cdots & \varphi_S \end{pmatrix} + \lambda_2^n \mathbf{r}_2 \boldsymbol{\ell}_2^T + \ldots + \lambda_S^n \mathbf{r}_S \boldsymbol{\ell}_S^T$$

$\lambda_2^n \mathbf{r}_2 \boldsymbol{\ell}_2^T \to 0$ as $n \to \infty$

$\lambda_S^n \mathbf{r}_S \boldsymbol{\ell}_S^T \to 0$ as $n \to \infty$

Convergence to the stationary distribution

Clearly:

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_S \\ \varphi_1 & \varphi_2 & \cdots & \varphi_S \\ \vdots & \vdots & & \vdots \\ \varphi_1 & \varphi_2 & \cdots & \varphi_S \end{pmatrix}$$

Furthermore, as:

$$0 \leq |\lambda_{k_1}| < |\lambda_{k_2}| < 1 \qquad \Rightarrow \qquad 0 \leq |\lambda_{k_1}^n| < |\lambda_{k_2}^n| < 1$$

It is the second largest (in absolute sense) eigenvalue that dominates, and thus determines the convergence speed to the stationary distribution.

Convergence to the stationary distribution

Fact

A Markov chain with a symmetric **P** has a uniform stationary distribution.

Proof

- Symmetry of **P** implies that left- and right eigenvectors are identical (up to a constant).
- First right eigenvector corresponds vector of ones, **1**.
- Hence, the left eigenvector equals c**1**.
- The left eigenvector is the stationary distribution and should sum to one: c = 1 / (number of states).

Convergence to the stationary distribution

Question

Suppose the DNA may reasonably described by a first order Markov model with transition matrix **P**:

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0.3 & 0.2 & 0.2 & 0.3 \\ 0.4 & 0.1 & 0.1 & 0.4 \end{pmatrix}$$

and stationary distribution:

$$\boldsymbol{\varphi}^T = (\varphi_A, \varphi_C, \varphi_G, \varphi_T) = (1/4, 1/4, 1/4, 1/4)$$

and eigenvalues:

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T = (1, 0.2, 0, 0)^T$$

Convergence to the stationary distribution

Question

What is the probability of a **G** at position 2, 3, 4, 5, 10?
And how does this depend on the starting nucleotide?

In other words, give:

$$P(X_2 = \mathbf{G} \mid X_1 = \mathbf{A}) = \ldots$$
$$P(X_2 = \mathbf{G} \mid X_1 = \mathbf{C}) = \ldots$$
$$P(X_2 = \mathbf{G} \mid X_1 = \mathbf{G}) = \ldots$$
$$P(X_2 = \mathbf{G} \mid X_1 = \mathbf{T}) = \ldots$$

But also:

$$P(X_3 = \mathbf{G} \mid X_1 = \mathbf{A}) = \ldots$$

et cetera.

Convergence to the stationary distribution

Thus, calculate P(X_t = **G** | X_1 = x_1) with t=2, 3, 4, 5, 10, and x_1 = **A**, **C**, **G**, **T**.

	x1=A	x1=C	x1=G	x1=T
t= 2	0.3000000	0.4000000	0.2000000	0.1000000
t= 3	0.2600000	0.2800000	0.2400000	0.2200000
t= 4	0.2520000	0.2560000	0.2480000	0.2440000
t= 5	0.2504000	0.2512000	0.2496000	0.2488000
t=10	0.2500001	0.2500004	0.2499999	0.2499996

Study the influence of the first nucleotide on the calculated probability for increasing t.

Convergence to the stationary distribution

```
> # define π and P
> pi <- matrix(c(1, 0, 0, 0), ncol=1)
> P <- matrix(c(2, 3, 3, 2, 1, 4, 4, 1, 3, 2, 2, 3,
      4, 1, 1, 4), ncol=4, byrow=TRUE)/10

> # define function that calculates the powers of a
> # matrix (inefficiently though)
> matrixPower <- function(X, power){
      Xpower <- X
      for (i in 2:power){
            Xpower <- Xpower %*% X
      }
      return(Xpower)
  }

> # calculate P to the power 100
> matrixPower(P, 100)
```

Convergence to the stationary distribution

Question

Suppose the DNA may reasonably described by a first order Markov model with transition matrix **P**:

$$\mathbf{P} = \begin{pmatrix} 0.77450 & 0.22500 & 0.00025 & 0.00025 \\ 0.22500 & 0.77450 & 0.00025 & 0.00025 \\ 0.00025 & 0.00025 & 0.77450 & 0.22500 \\ 0.00025 & 0.00025 & 0.22500 & 0.77450 \end{pmatrix}$$

and stationary distribution:

$$\boldsymbol{\varphi}^T = (\varphi_A, \varphi_C, \varphi_G, \varphi_T) = (1/4, 1/4, 1/4, 1/4)$$

and eigenvalues:

$$\boldsymbol{\lambda} = (1, 0.9990, 0.5495, 0.5495)^T$$

Convergence to the stationary distribution

Again calculate $P(X_t = \mathtt{G} \mid X_1 = x_1)$ with t=2, 3, 4, 5, 10, and x_1 = **A**, **C**, **G**, **T**.

	x1=A	x1=C	x1=G	x1=T
t= 2	0.0002500	0.0002500	0.7745000	0.2250000
t= 3	0.0004998	0.0004998	0.6504754	0.3485251
t= 4	0.0007493	0.0007493	0.5822116	0.4162899
t= 5	0.0009985	0.0009985	0.5445885	0.4534145
t=10	0.0022410	0.0022410	0.5000429	0.4954751
...	...	...	...	...
t=∞	0.2500000	0.2500000	0.2500000	0.2500000

Now the influence of the first nucleotide fades slowly.
This can be explained by the large 2nd eigenvalue.

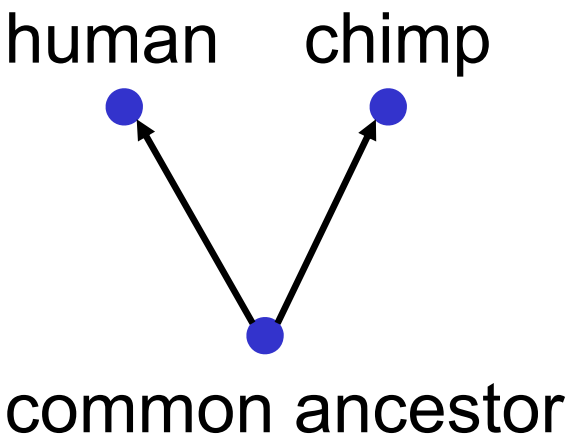
Processes
back in time

Processes back in time

So far, we have studied Markov chains forward in time. In practice, we may wish to study processes back in time.

Example

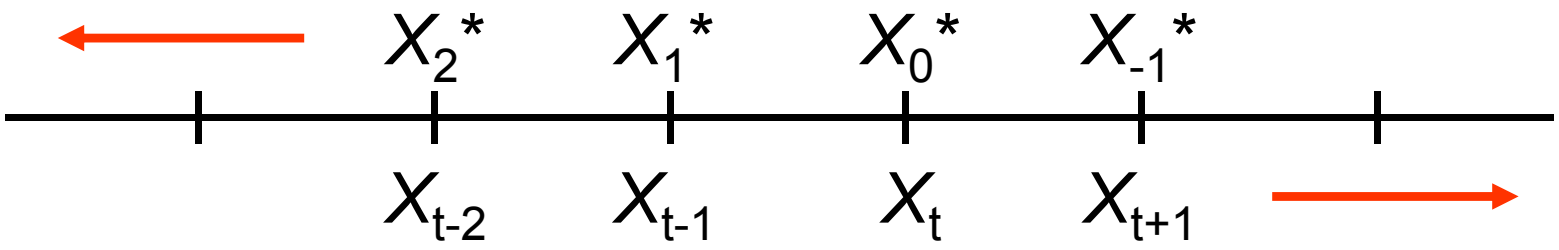
Evolutionary models that describe occurrence of SNPs in DNA sequences. We aim to attribute two DNA sequences to a common ancestor.



Processes back in time

Consider a Markov chain $\{X_t\}_{t=1,2,\ldots}$. The reverse Markov chain $\{X_r^*\}_{r=1,2,\ldots}$ is then defined by:

$$X_r^* = X_{t-r}$$



With transition probabilities:

$$p_{ij} = P(X_t=E_j \mid X_{t-1}=E_i)$$

$$p_{ij}^* = P(X_r^*=E_j \mid X_{r-1}^*=E_i)$$

Processes back in time

Show how the transition probabilities p_{ii}^* relate to p_{ii}:

$$p_{ij}^* = P(X_r^* = E_j \mid X_{r-1}^* = E_i)$$

just the definition

Processes back in time

Show how the transition probabilities p_{ii}^* relate to p_{ii}:

$$\begin{aligned} p_{ij}^* &= P(X_r^* = E_j \mid X_{r-1}^* = E_i) \\ &= P(X_{t-r} = E_j \mid X_{t-r+1} = E_i) \end{aligned}$$

express this in terms of the original Markov chain using that $X_r^* = X_{t-r}$

Processes back in time

Show how the transition probabilities p_{ii}^* relate to p_{ii}:

$$
\begin{aligned}
p_{ij}^* &= P(X_r^* = E_j \mid X_{r-1}^* = E_i) \\
&= P(X_{t-r} = E_j \mid X_{t-r+1} = E_i) \\
&= P(X_{t-r+1} = E_i \mid X_{t-r} = E_j) \, \frac{P(X_{t-r} = E_j)}{P(X_{t-r+1} = E_i)}
\end{aligned}
$$

apply definition of conditional probability twice (Bayes):
P(A | B) = P(B | A) P(A) / P(B)

Processes back in time

Show how the transition probabilities p_{ii}^* relate to p_{ii}:

$$\begin{aligned} p_{ij}^* &= P(X_r^* = E_j \mid X_{r-1}^* = E_i) \\ &= P(X_{t-r} = E_j \mid X_{t-r+1} = E_i) \\ &= P(X_{t-r+1} = E_i \mid X_{t-r} = E_j) \, \frac{P(X_{t-r} = E_j)}{P(X_{t-r+1} = E_i)} \\ &= p_{ji} \, \varphi_j \, / \, \varphi_i \end{aligned}$$

Hence:

$$p_{ij}^* = p_{ji} \, \frac{\varphi_j}{\varphi_i}$$

Processes back in time

Check that rows of the transition matrix **P*** sum to one, i.e.:

$$p_{i1}^* + p_{i2}^* + \ldots + p_{jS}^* = 1$$

Hereto:

$$\sum_{j=1}^{S} p_{ij}^* = \sum_{j=1}^{S} p_{ji} \frac{\varphi_j}{\varphi_i}$$

$$= \frac{1}{\varphi_i} \sum_{j=1}^{S} p_{ji} \varphi_j = \frac{1}{\varphi_i} \varphi_i = 1$$

use the fact that $\boldsymbol{\varphi}^T = \boldsymbol{\varphi}^T \mathbf{P}$

Processes back in time

The two Markov chains defined by **P** and **P*** have the same stationary distribution. Indeed, as:

$$\begin{aligned}\sum_{k=1}^{S} \varphi_k \, p^*_{kj} &= \sum_{k=1}^{S} p_{ik} \, \varphi_j \\ &= \varphi_j \sum_{k=1}^{S} p_{ik} \\ &= \varphi_j\end{aligned}$$

we have:

$$\boldsymbol{\varphi}^T = \boldsymbol{\varphi}^T \mathbf{P}^*$$

Processes back in time

Definition

A Markov chain is called ***reversible*** if $p_{ij}^* = p_{ij}$.
In that case:

$$p_{ij}^* = p_{ij} = p_{ji}\, \varphi_j / \varphi_i$$

Or,

$$\varphi_i\, p_{ij} = \varphi_j\, p_{ji} \qquad \text{for all i and j.}$$

These are the so-called ***detailed balance equations***.

Theorem

A Markov chain is reversible if and only if the detailed balance equations hold.

Processes back in time

Example 1

The 1st order Markov chain with transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

is irreversible. Check that this (deterministic) Markov chain does not satisfy the detailed balance equations.

Irreversibility can be seen from a sample of this chain:

... **ABCABCABCABCABCABC**...

In the reverse direction transitions from **B** to **C** do not occur!

Processes back in time

Example 2

The 1st order Markov chain with transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \end{pmatrix}$$

is irreversible. Again, a uniform stationary distribution:

$$\boldsymbol{\varphi}^T = (\varphi_A, \varphi_B, \varphi_C) = (1/3, 1/3, 1/3)$$

As **P** is not symmetric, the detailed balance equations are not satisfies:

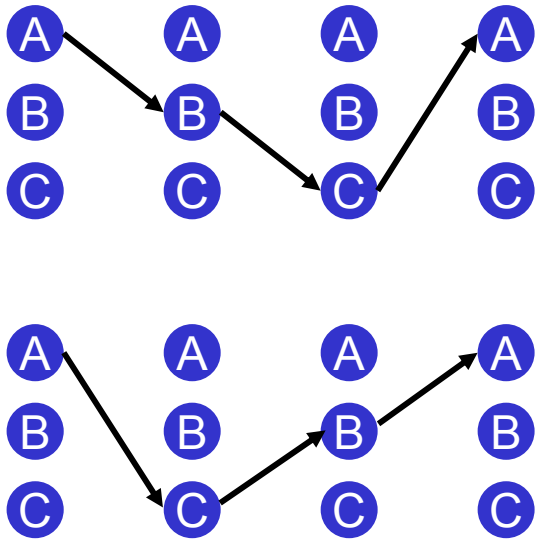
$$p_{ij} / 3 \neq p_{ji} / 3 \quad \text{for all i and j.}$$

Processes back in time

Example 2 (continued)

The irreversibility of this chain implies:

$$P(\mathbf{A} \to \mathbf{B} \to \mathbf{C} \to \mathbf{A})$$
$$= P(\mathbf{A} \to \mathbf{B})\, P(\mathbf{B} \to \mathbf{C})\, P(\mathbf{C} \to \mathbf{A})$$
$$= 0.8 * 0.8 * 0.8$$
$$\neq 0.1 * 0.1 * 0.1$$
$$= P(\mathbf{A} \to \mathbf{C})\, P(\mathbf{C} \to \mathbf{B})\, P(\mathbf{B} \to \mathbf{A})$$
$$= P(\mathbf{A} \to \mathbf{C} \to \mathbf{B} \to \mathbf{A}).$$



It matters how one walks from **A** to **A**.

Or, it matters whether one walks forward or backward.

Processes back in time

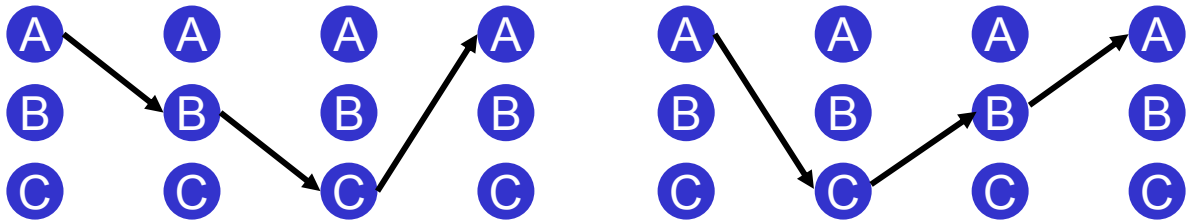
Kolmogorov condition for reversibility

A stationary Markov chain is reversible if and only if any path from state E_i to state E_i has the same probability as the path in the opposite direction. Or, the Markov chain is reversible if and only if:

$$p_{i,i_1} \cdot p_{i_1,i_2} \cdot \ldots \cdot p_{i_k,i} \quad = \quad p_{i,i_k} \cdot \ldots \cdot p_{i_2,i_1} \cdot p_{i_1,i}$$

for all i, i_1, i_2, …, i_k.

E.g.: P(**A** →**B**→**C**→**A**) = P(**A** →**C**→**B**→**A**).



Processes back in time

Interpretation

For a reversible Markov chain it is not possible to determine the direction of the process from the observed state sequence alone.

- Molecular phylogenetics aims to reconstruct evolutionary relationships between present day species from their DNA sequences. Reversibility is then an essential assumption.
- Genes are transcribed in one direction only (from the 3' end to the 5' end). The promoter is only on the 3' end. This suggests irreversibility.

Processes back in time

E. Coli

For a gene in the E. Coli genome, we estimate:

```
Transition matrix
          [,1]      [,2]      [,3]      [,4]
[1,] 0.2296984 0.3155452 0.2273782 0.2273782
[2,] 0.1929134 0.2421260 0.2933071 0.2716535
[3,] 0.1979167 0.2854167 0.2166667 0.3000000
[4,] 0.2522686 0.2032668 0.2341198 0.3103448

Stationary distribution
[1] 0.2187817 0.2578680 0.2436548 0.2796954
```

Then, the detailed balance equation do not hold, e.g.:

$\pi_1 \, p_{12} \neq \pi_2 \, p_{21.}$

Processes back in time

Note

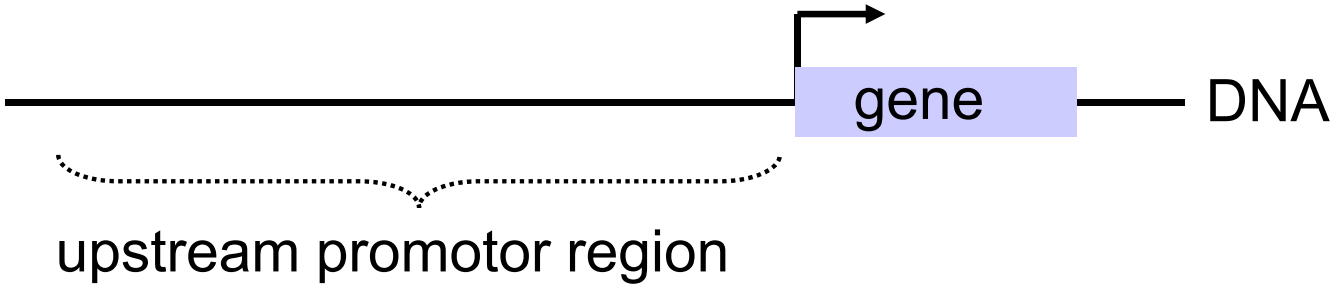
Within evolution theory the notion of irreversibility refers to the presumption that complex organisms once lost evolution will not appear in the same form.

Indeed, the likelihood of reconstructing a particular phylogenic system is infinitesimal small.

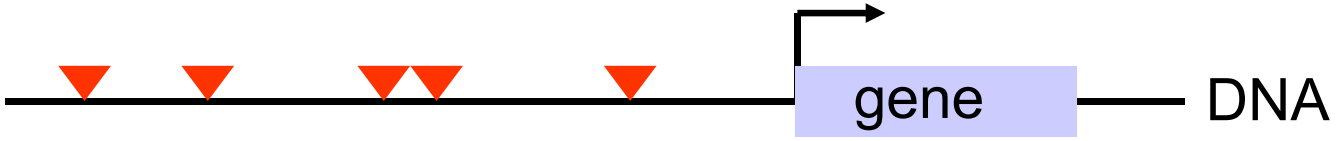
Application: motifs

Application: motifs

Study the sequence of the promotor region up-stream of a gene.



This region contains binding sites for the transcription factors that regulate the transcription of the gene.



Application: motifs

The binding sites of a transciption factor (that may regulate multiple genes) share certain sequence patterns, ***motifs***.

Not all transcription factors and motifs are known. Hence, a high occurrence of a particular sequence pattern in the upstream regions of a gene may indicate that it has a regulatory function (e.g., binding site).

Problem

Determine the probability of observing *m* motifs in a background generated by a 1^{st} order stationary Markov chain.

Application: motifs

An *h*-letter ***word*** $W = w_1\ w_2\ \ldots\ w_h$ is a map from {1, …, *h*} to $\mathcal{A}^h$, where $\mathcal{A}$ some non-empty set, called the ***alphabet***.

In the DNA example:

$$\mathcal{A} = \{\mathtt{A},\ \mathtt{C},\ \mathtt{G},\ \mathtt{T}\}$$

and, e.g.:

$$W = \mathtt{CAGTACGACT}$$

$$W = \mathtt{TACGACTGCATATGCGTA}$$

Application: motifs

A word *W* is ***p*-periodic** if

$$w_{t_1} = w_{t_2} \quad \text{for all } t_1 \equiv t_2 \bmod p,\ \ 1 \le t_1, t_2 \le p$$

The lag between two overlapping occurrences of the word.

The set of all periods of *W* (less than *h*) is the ***period set***, denoted by $\mathcal{P}(W)$. In other words, the set of integers 0 < *p* < *h* such that a new occurrence of *W* can start *p* letters after an occurrence of *W*.

Application: motifs

If W_1 = **CGATCGATCG**, then $\mathcal{P}(W) = \{5, 9\}$

For:

```
123456789
CGATCGATC
    CGATCGATC
        CGATCGATC
```

If W_2 = **CGAACTG**, then $\mathcal{P}(W) = \emptyset$

Application: motifs

Let $N(W)$ be the number of (overlapping) occurrences of an h-letter word W in a random sequence n on A.

If Y_t is the random variable defined by:

$$Y_t = I_{\{\text{an occurence of } W \text{ starts at position } t\}}$$

then

$$N(W) = \sum_{t=1}^{n-h+1} Y_t$$

Application: motifs

Also, denote the number of occurences in *W* by:

$$n(a) \quad = \quad \{\#t \mid a \in W\}$$

and

$$n(a\bullet) \quad = \quad \sum_{b \in \mathcal{A}} n(ab)$$

If *W* = **CATGA**, then:

n(**A**) = 2 and n(**A**•) = 1

Application: motifs

Assume a stationary 1st order Markov model for the random sequence of length *n*. The probability of an occurrence of *W* in the random sequence is given by:

$$\mu(W) \quad = \quad \varphi_{w_1} \prod_{t=1}^{h-1} (\mathbf{P})_{w_j, w_{j+1}}$$

Application: motifs

In the 1st order Markov model, the expected number of occurrences of *W* is approximated by:

$$\hat{N}(W) = \frac{N(w_1 w_2) \cdot N(w_2 w_3) \cdot \ldots \cdot N(w_{h-1} w_h)}{N(w_2) \cdot N(w_3) \cdot \ldots \cdot N(w_{h-1})}$$

and its variance by:

$$\sigma^2(W) = \mu(W) + \sum_{p \in \mathcal{P}(W)} \mu(w_1 \ldots w_p \, w_1 \ldots w_h)$$
$$+ \mu(W)^2 \Big(\sum_{a \in \mathcal{A}} \frac{n(a\bullet)^2}{\mu(a)} - \sum_{a_1, a_2 \in \mathcal{A}} \frac{n(a_1 a_2)^2}{\mu(a_1 a_2)} + \frac{1 - n(w_1\bullet)^2}{\mu(w_1)} \Big)$$

Application: motifs

To find words with exceptionally frequency in the DNA, the following (asymptotically) standard normal statistic is used:

$$Z(W) = \frac{N(W) - \hat{N}(W)}{\sqrt{n\,\hat{\sigma}^2(W)}} \sim N(0,1)$$

The *p*-value of word *W* is then given by:

$$\begin{aligned} p(W) &= P(Z \geq Z(W)) \\ &= 1 - \Phi_{0,1}(Z(W)) \end{aligned}$$

Application: motifs

Note

Robin, Daudon (1999) provide exact probabilities of word occurences in random sequences.

However, Robin, Schbath (2001) point out that calculation of the exact probabilities is computationally intensive. Hence, the use of an approximation here.

References & further reading

References and further reading

Ewens, W.J, Grant, G (2006), *Statistical Methods for Bioinformatics*, Springer, New York.

Reinert, G., Schbath, S., Waterman, M.S. (2000), “Probabilistic and statistical properties of words: an overview”, *Journal of Computational Biology*, **7**, 1-46.

Robin, S., Daudin, J.-J. (1999), “Exact distribution of word occurrences in a random sequence of letters”, *Journal of Applied Probability,* **36**, 179-193.

Robin, S., Schbath, S. (2001), “Numerical comparison of several approximations of the word count distribution in random sequences”, *Journal of Computational Biology*, **8**(4), 349-359.

Schbath, S. (2000), “An overview on the distribution of word counts in Markov chains”, *Journal of Computational. Biology,* **7**, 193-202.

Schbath, S., Robin, R. (2009), “How can pattern statistics be useful for DNA motif discovery?”. In *Scan Statistics: Methods and Applications* by Glaz, J. *et al*. (eds.).