Stochastic processes and Markov chains (part I)

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

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

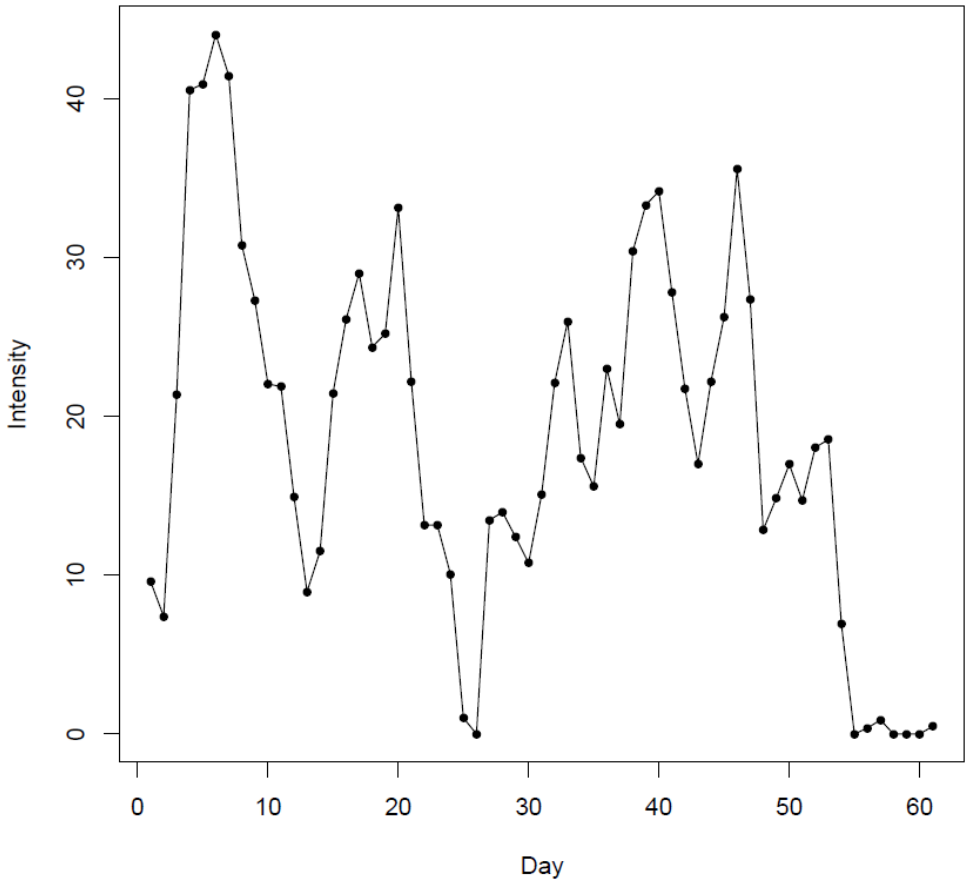
vrije Universiteit

VU medisch centrum

Stochastic processes

Example 1

- The intensity of the sun.
- Measured every day by the KNMI.
- Stochastic variable X_t represents the sun's intensity at day t, $0 \leq t \leq T$. Hence, X_t assumes values in $\mathbf{R}^+$ (positive values only).



Data from KNMI-website

Stochastic processes

Example 2

- DNA sequence of 11 bases long.
- At each base position there is an **A**, **C**, **G** or **T**.
- Stochastic variable X_i is the base at position i, i = 1,…,11.
- In case the sequence has been observed, say:
  $(x_1, x_2, \ldots, x_{11})$ = **ACCCGATAGCT**,
  then **A** is the realization of X_1, **C** that of X_2, et cetera.

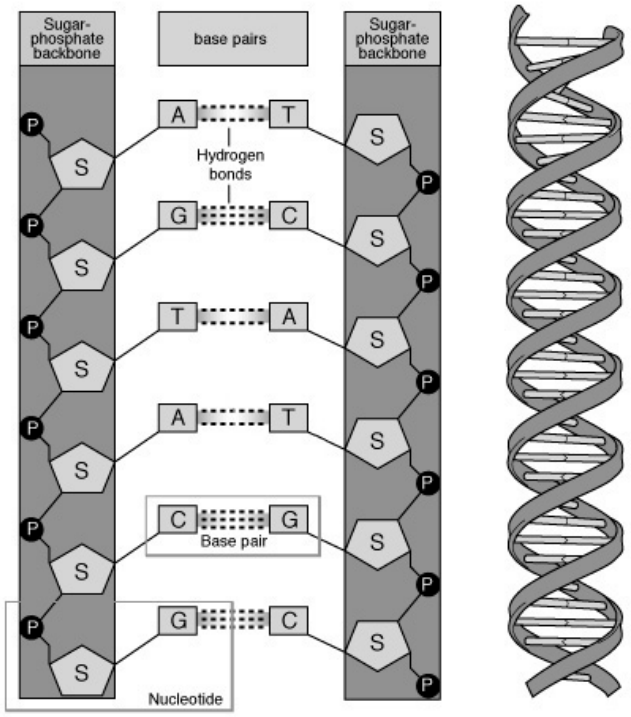
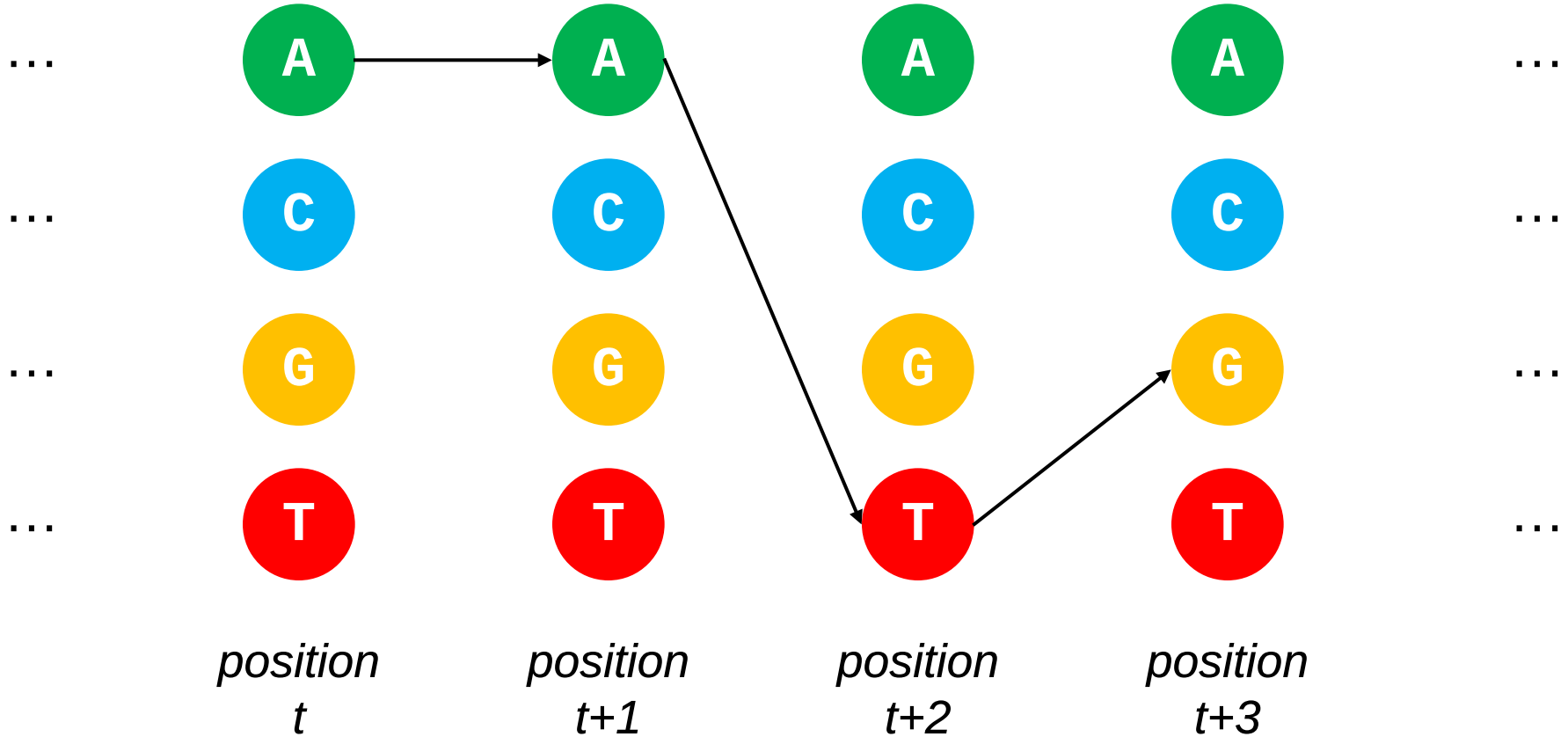


Figure: http://ircamera.as.arizona.edu/

Stochastic processes

position	…	t	t+1	t+2	t+3	…
base	…	A	A	T	C	…



Stochastic processes

Example 3

- A patient's heart pulse during surgery.
- Measured continuously during interval [0, T].
- Stochastic variable X_t represents the occurrence of a heartbeat at time t, $0 \leq t \leq T$. Hence, X_t assumes only the values 0 (no heartbeat) and 1 (heartbeat).

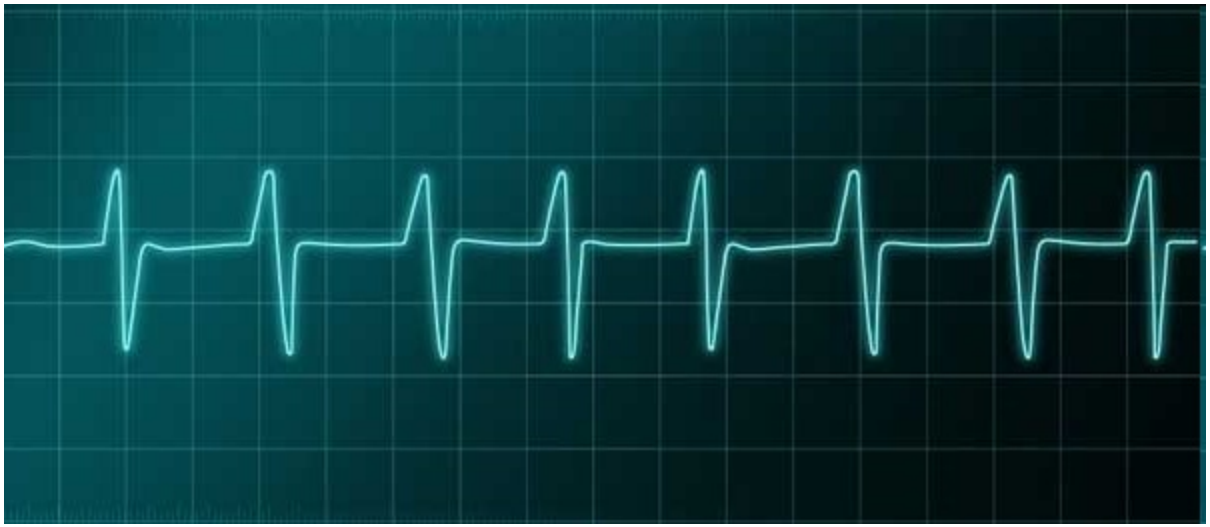
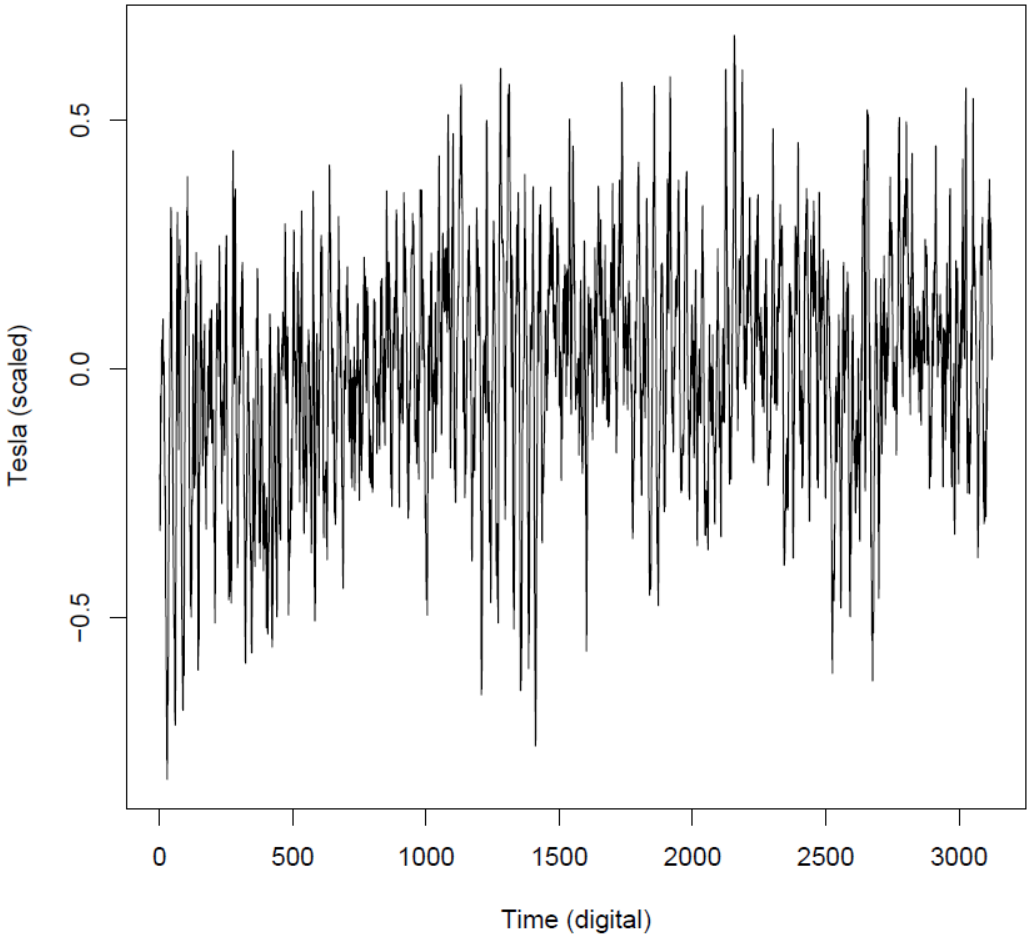


Figure: http://www.gnkaterini.gr/

Stochastic processes

Example 4

- Brain activity of a human under experimental conditions.
- Measured continuously during interval [0, T].
- Stochastic variable X_t represents the magnetic field at time t, $0 \leq t \leq T$. Hence, X_t assumes values on **R**.



Data kindly provided by R. Hindriks

Stochastic processes

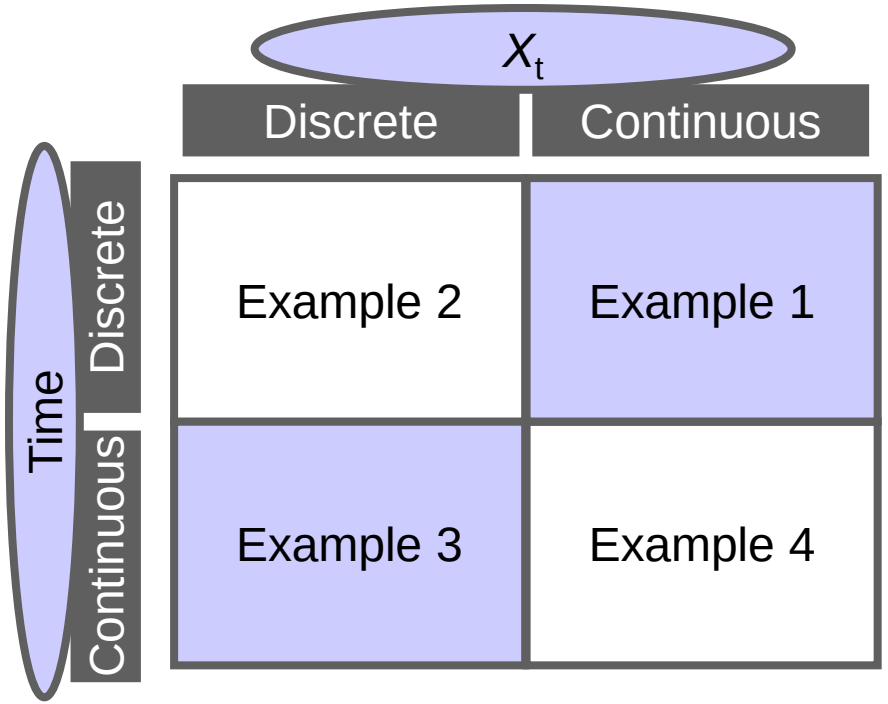
book Ewens, Grant (2005): Section 4.5.

The *state space* S is the collection of values that the random variables of the stochastic process may assume.

If $S = \{E_1, E_2, \ldots, E_s\}$, then X_t is a *discrete stochastic variable.*

If $S = [0, \infty)$, then X_t is a *continuous stochastic variable.*

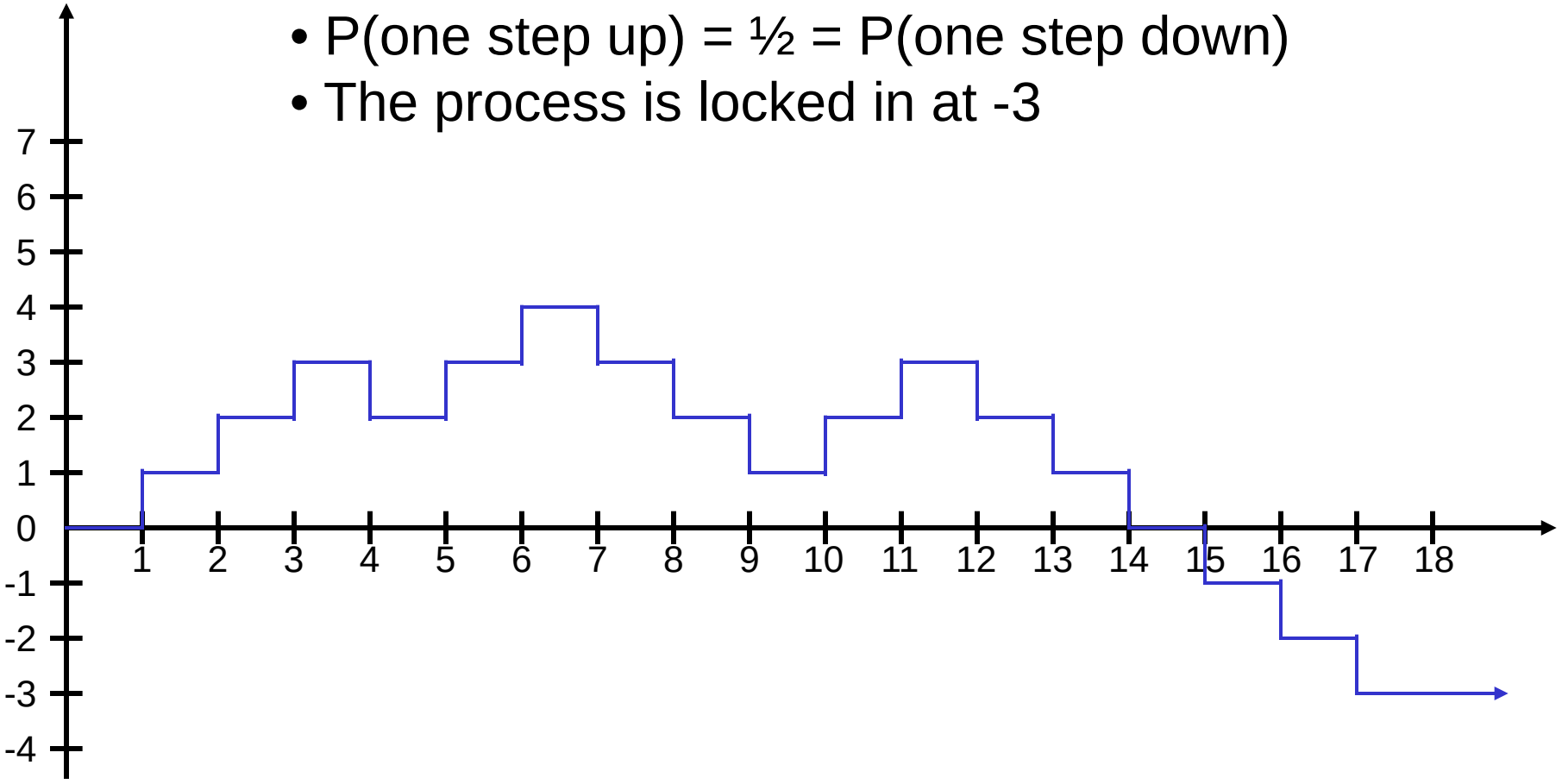
Time can also be either discrete or continuous.



Stochastic processes

Stochastic process:

- Discrete time: t = 0, 1, 2, 3, ….
- *S* = {-3, -2, -1, 0, 1, 2, …..}
- P(one step up) = ½ = P(one step down)
- The process is locked in at -3

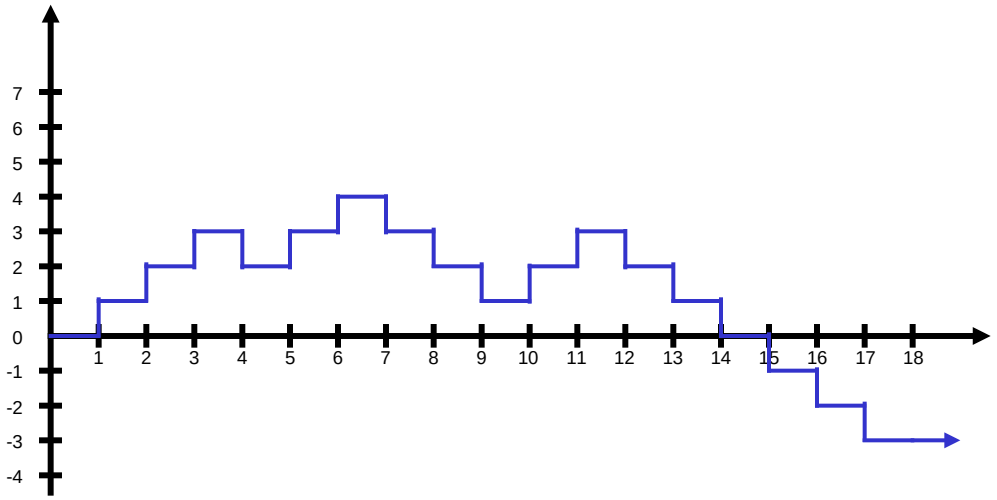


Stochastic processes

The *first passage time* of a certain state E_i in S is the time t at which $X_t = E_i$ for the first time since the start of the process.

The *absorbing state* is a state E_i for which the following holds: if $X_t = E_i$ than $X_s = E_i$ for all $s \geq t$. The process will never leave state E_i.

The *time of absorption* of an absorbing state is the first passage time of that state.

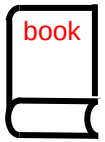


Stochastic processes

A *stochastic process* is described by a collection of time points, the state space and the simultaneous distribution of the variables X_t, i.e., the distributions of all X_t and their dependency.

There are two important types of processes:

- *Poisson process*: all variables are identically and independently distributed. Examples: queues for counters, call centers, servers, et cetera.
- *Markov process*: the variables are dependent in a simple manner.



Axelson-Fisk (2010):
Definition 2.1.

Markov processes

book Axelson-Fisk (2010): Definition 2.1.

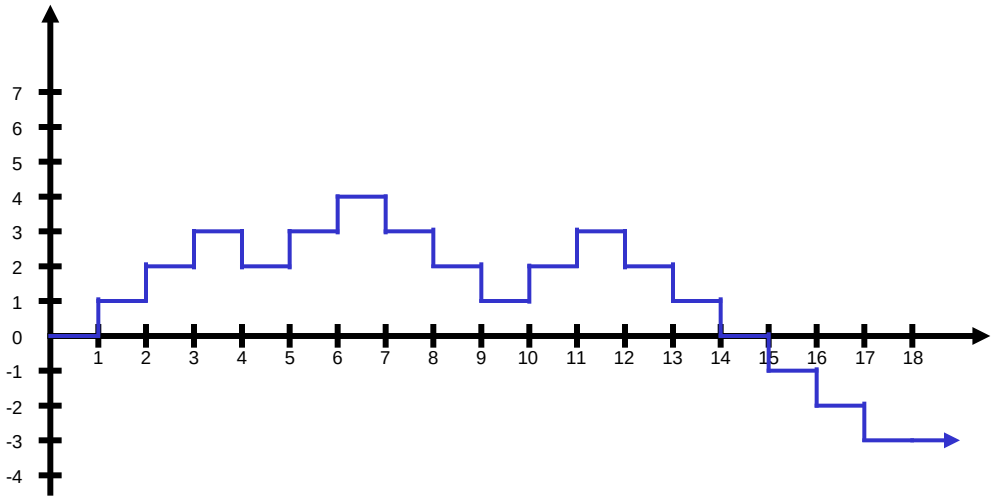
Markov processes

A *1st order Markov process* in discrete time is a stochastic process $\{X_t\}_{t=1,2,\ldots}$ for which the following holds:

$$P(X_{t+1}=x_{t+1} \mid X_t=x_t, \ldots, X_1=x_1) = P(X_{t+1}=x_{t+1} \mid X_t=x_t).$$

In other words, only the present determines the future, the past is irrelevant.

From every state with probability 0.5 one step up or down (except for state -3).



Markov processes

The *1st order Markov property*, i.e.

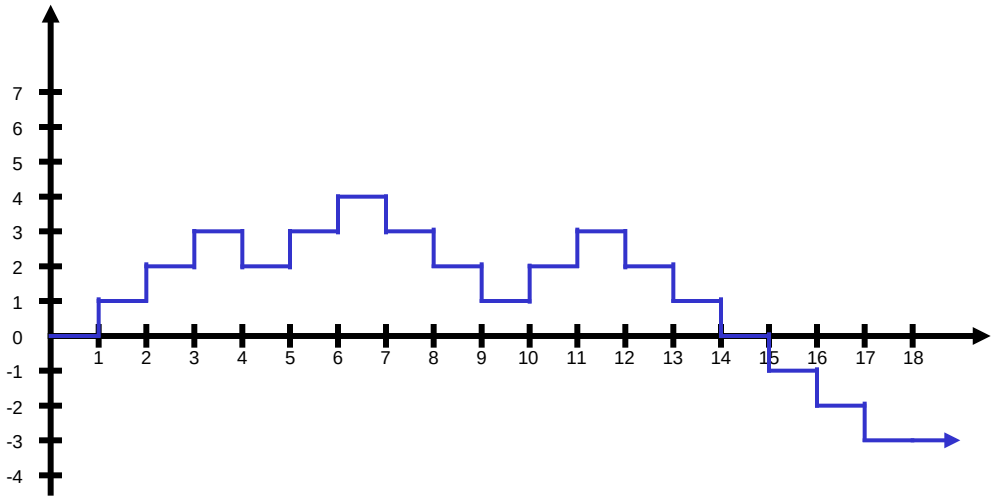
$$P(X_{t+1}=x_{t+1} \mid X_t=x_t, \ldots, X_1=x_1) = P(X_{t+1}=x_{t+1} \mid X_t=x_t),$$

does not imply independence between X_{t-1} and X_{t+1}.

Nor does it imply that $P(X_{t+1}=x_{t+1} \mid X_{t-1}=x_{t-1})$ equals zero.

For instance,

$P(X_{t+1}=2 \mid X_{t-1}=0) =$
$P(X_{t+1}=2 \mid X_t=1)$
$* P(X_t=1 \mid X_{t-1}=0) = ¼.$



Markov processes

The *0^{th} order Markov property*, i.e.

$$P(X_{t+1}=x_{t+1} \mid X_t=x_t, \ldots, X_1=x_1) = P(X_{t+1}=x_{t+1}).$$

Thus, X_t and X_{t+1} are independent. So are X_{t-1} and X_{t+1}, and X_t and X_{t+100}, and …

Or:

$$P(X_{t+1}=x_{t+1}, X_t=x_t) = P(X_{t+1}=x_{t+1}) * P(X_t=x_t)$$

Classic example:

- Tossing a coin : $P(X_{t+1} = \text{head} \mid X_t = \text{tail}) = P(X_{t+1} = \text{head})$
- Rolling dice : $P(X_{t+1} = 6 \mid X_t = 6) = P(X_{t+1} = 6)$

Markov processes

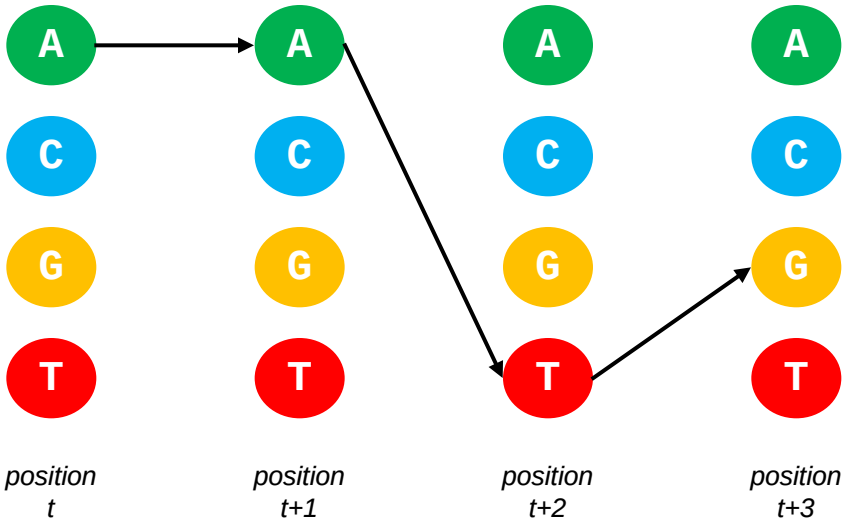
Difference between 0^{th} and 1^{st} order Markov process

→ DNA example

→ state space: {**A**, **C**, **G**, **T**}

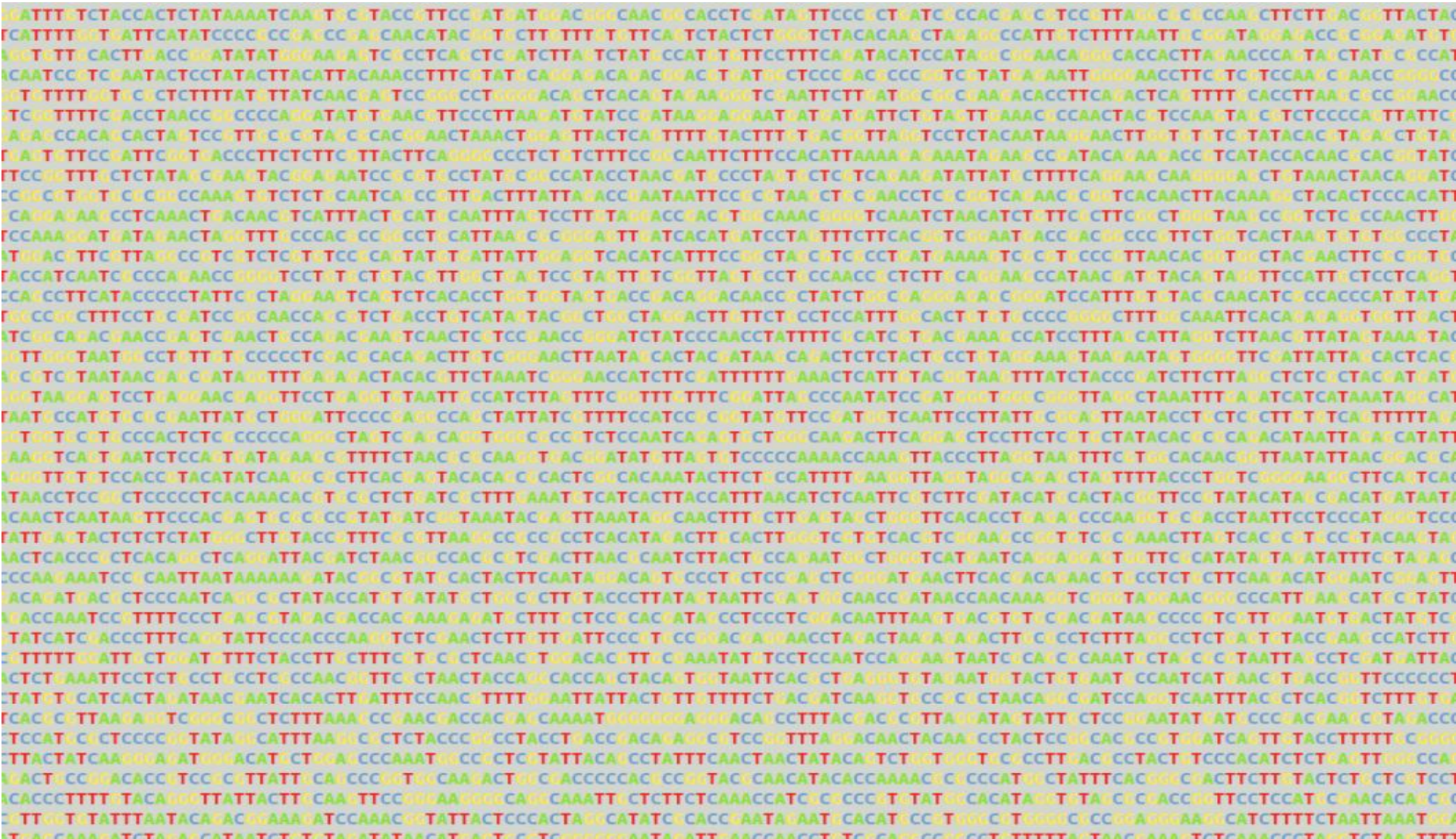
→ sample in accordance with a 0^{th} and 1^{st} Markov chain

Recall: 1^{st} order takes into account the previous state visited, 0^{th} order does not.



Markov processes

Question: what is the order of the processes?



Markov processes

Question: what is the order of the processes?

book Axelson-Fisk (2010): Formula 2.9.

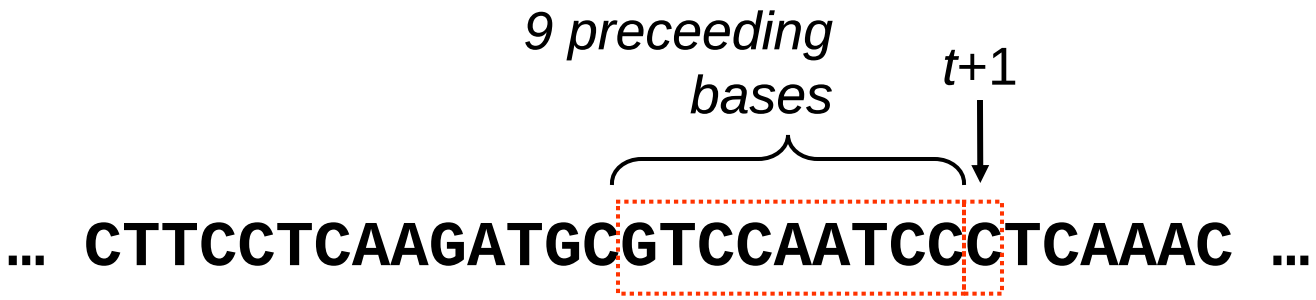
Markov processes

An *m-order Markov process* in discrete time is a stochastic process $\{X_t\}_{t=1,2,\ldots}$ for which the following holds:

$$P(X_{t+1}=x_{t+1} \mid X_t=x_t, \ldots, X_1=x_1)$$
$$= P(X_{t+1}=x_{t+1} \mid X_t=x_t,\ldots,X_{t-m+1}= x_{t-m+1}).$$

Loosely, the future depends on the most recent past.

Suppose *m=9:*



Distribution of the *t*+1-th base depends on the 9 preceeding ones.

Markov processes

A Markov process is called a *Markov chain* if the state space is discrete, i.e., is finite or countable. We consider Markov chains in discrete time.

The *transition probabilities* are the $P(X_{t+1}=x_{t+1} \mid X_t=x_t)$, but also the

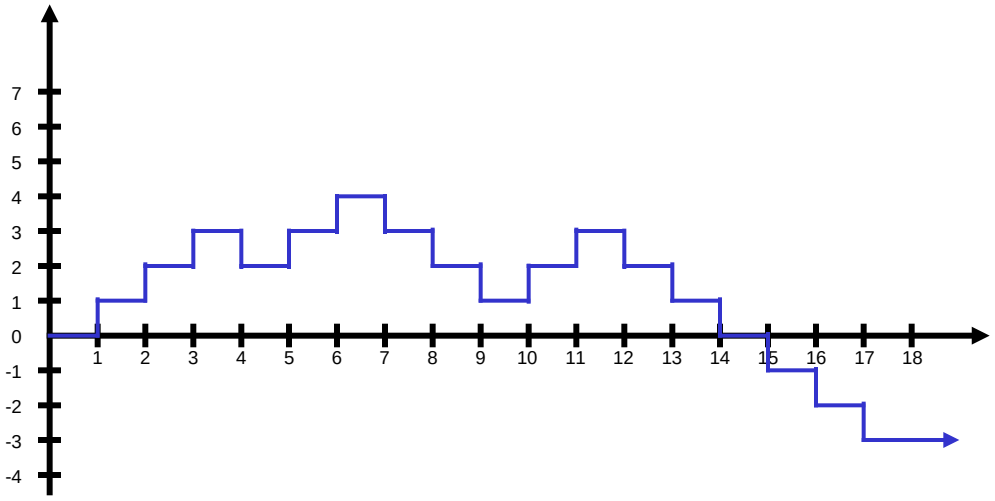
$$P(X_{t+1}=x_{t+1} \mid X_s=x_s) \quad \text{for} \quad s < t,$$

where x_{t+1}, x_t in $\{E_1, \ldots, E_s\} = S$.

$P(X_{t+1}=3 \mid X_t=2) = 0.5$

$P(X_{t+1}=0 \mid X_t=1) = 0.5$

$P(X_{t+1}=5 \mid X_t=7) = 0$

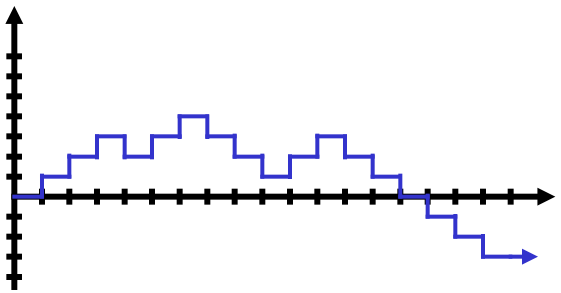


book Axelson-Fisk (2010): Definition 2.3.

Markov processes

A Markov process is called *time homogeneous* if the transition probabilities are independent of t:

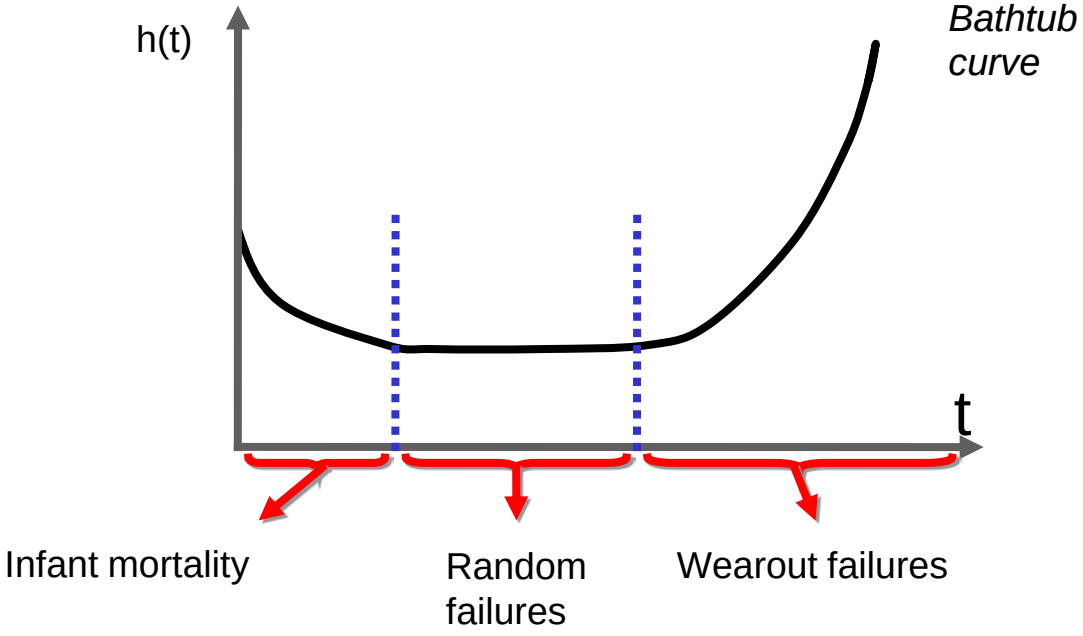
$$P(X_{t+1}=x_1 \mid X_t=x_2) = P(X_{s+1}=x_1 \mid X_s=x_2).$$



For example:

$$P(X_2=-1 \mid X_1=2) = P(X_6=-1 \mid X_5=2)$$

Example of a *time inhomogeneous* process (not in lecture).

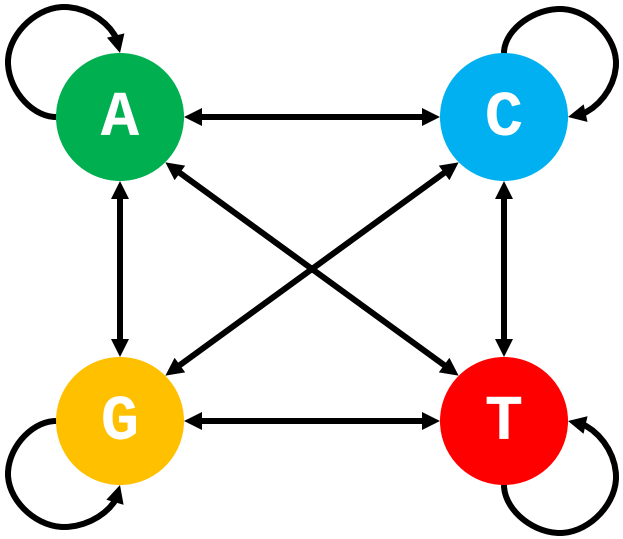


book Ewens, Grant (2005): Section 4.9.

Markov processes

Consider a DNA sequence of 11 bases. Then, S={**A, C, G, T**}, X_t is the base of position t, and $\{X_t\}_{t=1, \ldots, 11}$ is a Markov chain if the base of position t only depends on the base of position t-1, and not on those before t-1.

If this is plausible, a 1st Markov chain may be an acceptable model for base ordering in DNA sequences.



state diagram

book Ewens, Grant (2005): Section 4.6.

Markov processes

Denote the transition probabilities of a finite, time homogeneous Markov chain in discrete time $\{X_t\}_{t=1,2,\ldots}$ with $S=\{E_1, \ldots, E_s\}$ as:

$$P(X_{t+1}=E_j \mid X_t=E_i) = p_{ij} \qquad \text{(does not depend on } t\text{)}.$$

Putting the p_{ij} in a matrix yields the transition matrix:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & & p_{2s} \\ \vdots & & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{pmatrix}$$

The rows of this matrix sum to one.

Coin: P(X=**head**) = ½ → P(X=**tail**) = 1 - P(X=**head**)

Markov processes

DNA:

$$\mathbf{P} = \begin{array}{c} \\ \text{from} \end{array} \begin{array}{c} \begin{array}{cccc} \multicolumn{4}{c}{\overbrace{\qquad\qquad\qquad\qquad}^{\textit{to}}} \\ \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{array}{c} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \end{array} \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$

where

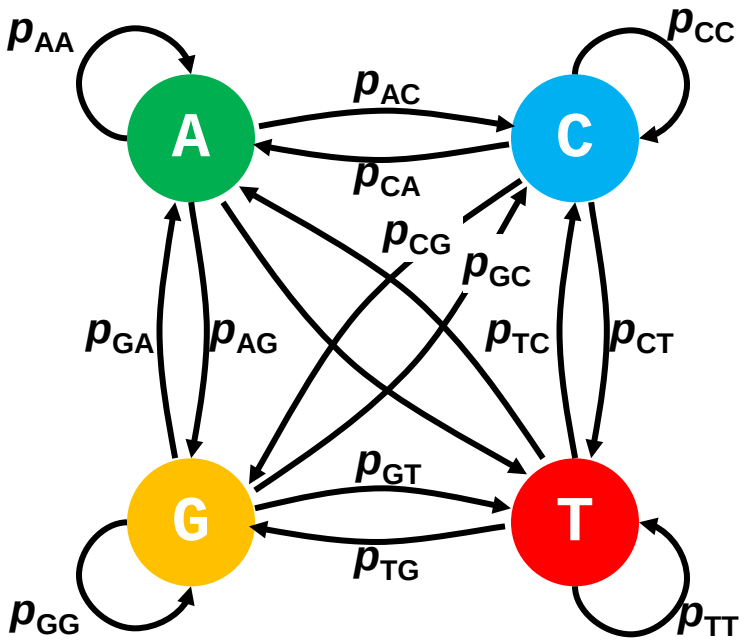
$p_{AA} = P(X_{t+1} = \mathbf{A} \mid X_t = \mathbf{A})$

and:

$p_{AA} + p_{AC} + p_{AG} + p_{AT} = 1$

$p_{CA} + p_{CC} + p_{CG} + p_{CT} = 1$

et cetera

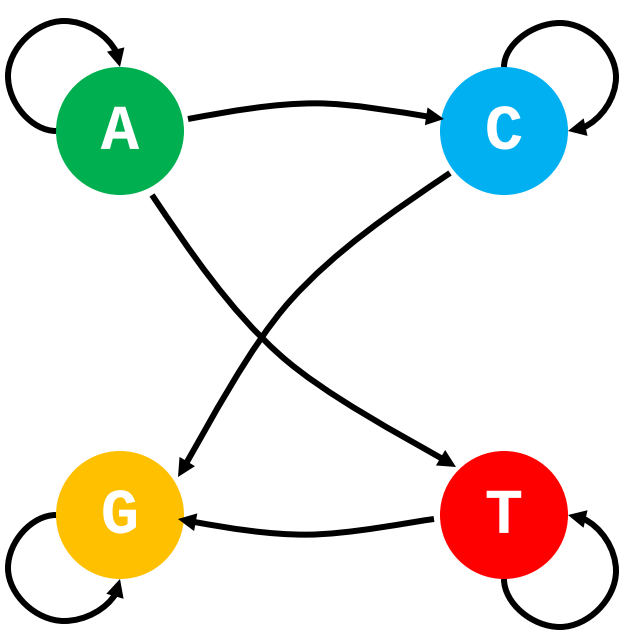


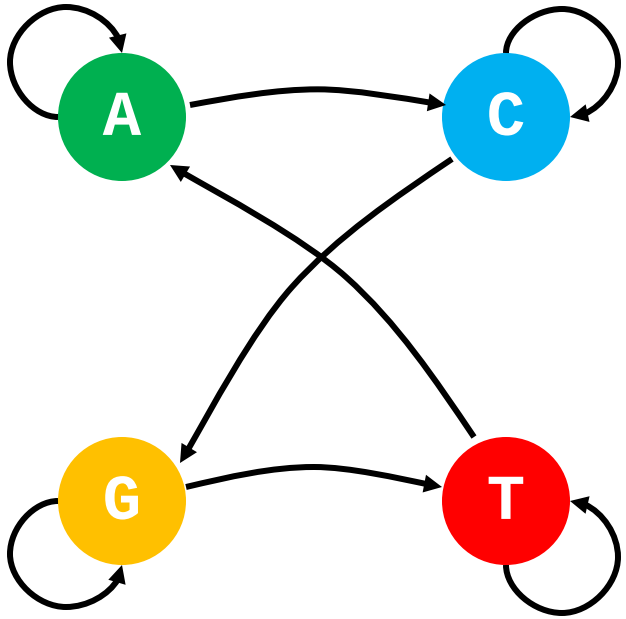
Markov processes

Question

Which state diagram corresponds to the transition matrix **P**?

$$\mathbf{P} = \begin{array}{c} \\ \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \begin{array}{c} \overbrace{\qquad\qquad\qquad\qquad} \\ \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} 0.1 & 0.9 & 0 & 0 \\ 0 & 0.1 & 0.9 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.9 & 0 & 0 & 0.1 \end{pmatrix} \end{array}$$ *to*





Markov processes

The *initial distribution* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_s)^T$ gives the probabilities of the initial state:

$$\pi_i = P(X_1 = E_i) \qquad \text{for } i = 1, \ldots, s,$$

and $\pi_1 + \ldots + \pi_s = 1$.

Together the initial distribution $\boldsymbol{\pi}$ and the transition matrix $\mathbf{P}$ determine the probability distribution of the process, the Markov chain $\{X_t\}_{t=1,2,\ldots}$

Markov processes

Question

Consider the transition matrix **P** (to the right) and initial distribution **π** = (0, 0.5, 0.5, 0).

$$\mathbf{P} = \begin{array}{c} \\ \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \begin{array}{c} \overbrace{\qquad\qquad\qquad\qquad} \\ \begin{pmatrix} 0.1 & 0.9 & 0 & 0 \\ 0 & 0.1 & 0.9 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.9 & 0 & 0 & 0.1 \end{pmatrix} \end{array}$$

to

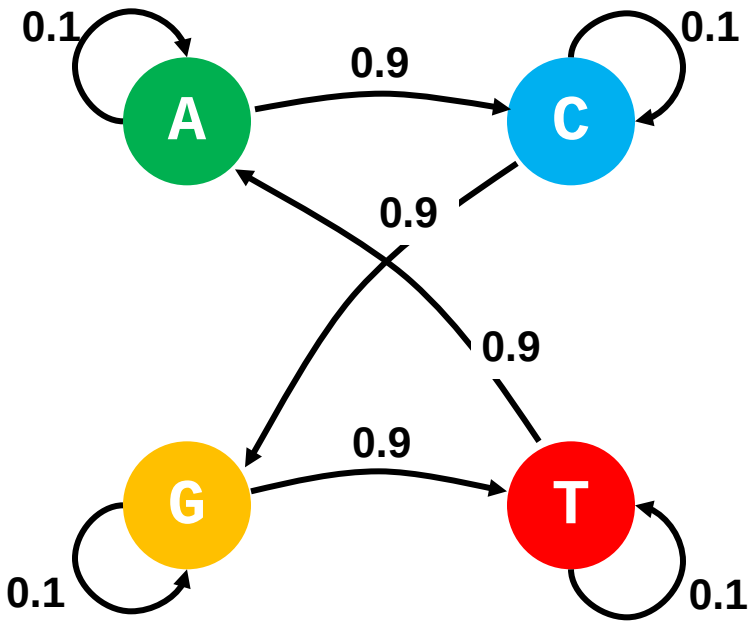
A **C** **G** **T**

Calculate the probabilities:

→ P(X_1=**A**, X_2=**C**),

→ P(X_1=**C**, X_2=**G**),

→ P(X_1=**C**, X_2=**T**).



book Ewens, Grant (2005): Section 4.6.

Markov processes

We now show how the couple (**π**, **P**) determines the probability distribution (transition probabilities) of time steps larger than one.

Hereto define for

n = 2 : $p_{ij}^{(2)} = P(X_{t+2} = E_j \mid X_t = E_i)$

general *n* $p_{ij}^{(n)} = P(X_{t+n} = E_j \mid X_t = E_i)$

Now express $p_{ij}^{(2)}$ in terms of (**π**, **P**).

Markov processes

For $n = 2$:

$$p_{ij}^{(2)} = P(X_{t+2} = E_j \mid X_t = E_i)$$

just the definition

Markov processes

For $n = 2$:

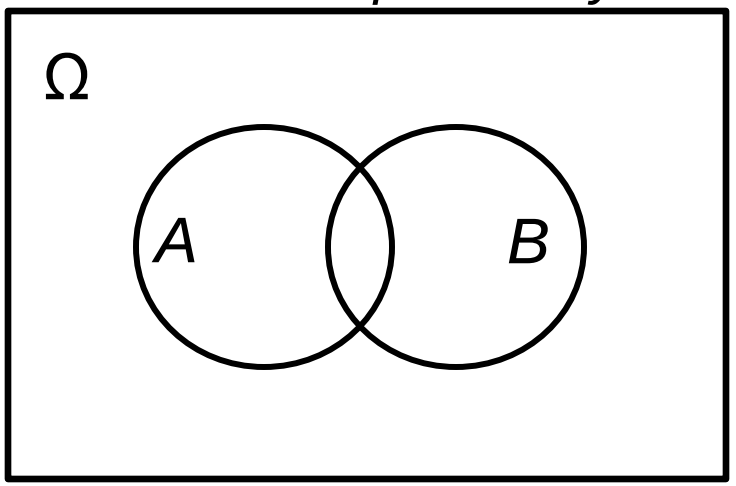
$$p_{ij}^{(2)} = P(X_{t+2} = E_j \mid X_t = E_i)$$

$$= \sum_{k=1}^{S} P(X_{t+2} = E_j, X_{t+1} = E_k \mid X_t = E_i)$$

use the fact that

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

Total probability law



Markov processes

For $n = 2$:

$$
\begin{aligned}
p_{ij}^{(2)} &= P(X_{t+2} = E_j \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j, X_{t+1} = E_k \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j \mid X_{t+1} = E_k, X_t = E_i) \\
&\qquad \times P(X_{t+1} = E_k \mid X_t = E_i)
\end{aligned}
$$

use the definition of conditional probability:

P(*A*, *B* | *C*) = P(*A*, *B*, *C*) / P(*C*)

= P(*A*, *B*, *C*) / P(*B*, *C*) * P(*B*, *C*) / P(*C*)

Markov processes

For $n = 2$:

$$
\begin{aligned}
p_{ij}^{(2)} &= P(X_{t+2} = E_j \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j, X_{t+1} = E_k \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j \mid X_{t+1} = E_k, X_t = E_i) \\
&\qquad \times P(X_{t+1} = E_k \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j \mid X_{t+1} = E_k)\, P(X_{t+1} = E_k \mid X_t = E_i)
\end{aligned}
$$

use the Markov property

Markov processes

For $n = 2$:

$$
\begin{aligned}
p_{ij}^{(2)} &= P(X_{t+2} = E_j \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j, X_{t+1} = E_k \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j \mid X_{t+1} = E_k, X_t = E_i) \\
&\qquad \times P(X_{t+1} = E_k \mid X_t = E_i) \\
&= \sum_{k=1}^{S} P(X_{t+2} = E_j \mid X_{t+1} = E_k)\, P(X_{t+1} = E_k \mid X_t = E_i) \\
&= \sum_{k=1}^{S} p_{kj}\, p_{ik} \;=\; \sum_{k=1}^{S} p_{ik}\, p_{kj} \;=\; (\mathbf{P}^2)_{ij}
\end{aligned}
$$

Markov processes

In summary, we have shown that:

$$\mathbf{P}^{(2)} = \mathbf{P}^2$$

In a similar fashion we can show that:

$$\mathbf{P}^{(n)} = \mathbf{P}^n \qquad \text{for all } n \geq 2.$$

In words: the transition matrix for *n* steps is the one-step transition matrix raised to the power *n*.

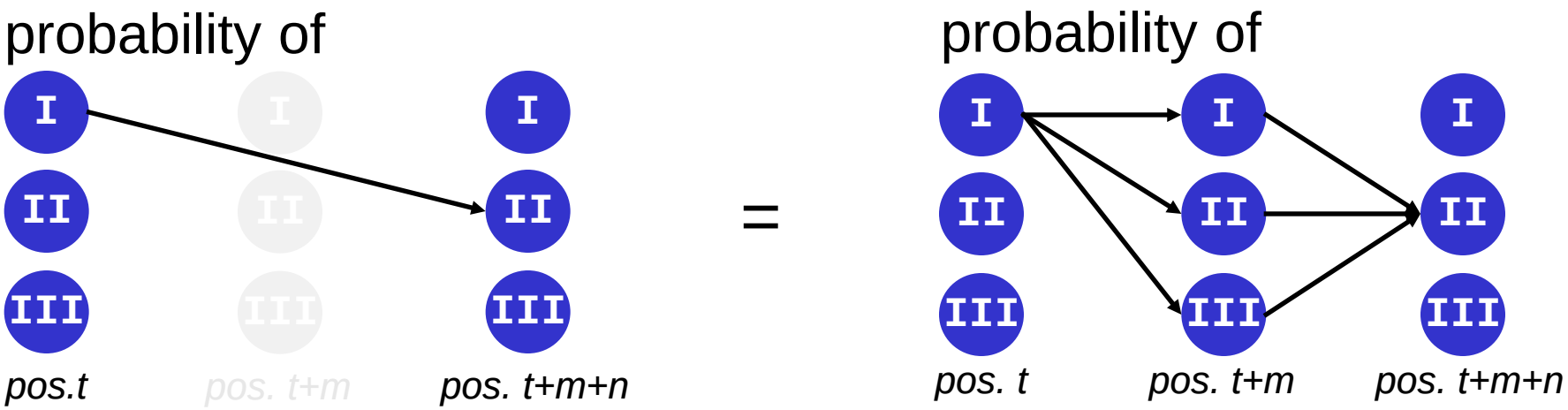
Markov processes

The general case is proven (see SM) by induction to *n*.
This requires the *Kolmogorov-Chapman equations*:

$$(\mathbf{P}^{n+m})_{ij} = \sum_{k=1}^{S} (\mathbf{P}^{n})_{ik} (\mathbf{P}^{m})_{kj}$$

for all n, m ≥ 0 and i, j =1,…,S.

Kolmogorov-Chapman equations illustrated:



Markov processes

A numerical example:

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

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”)

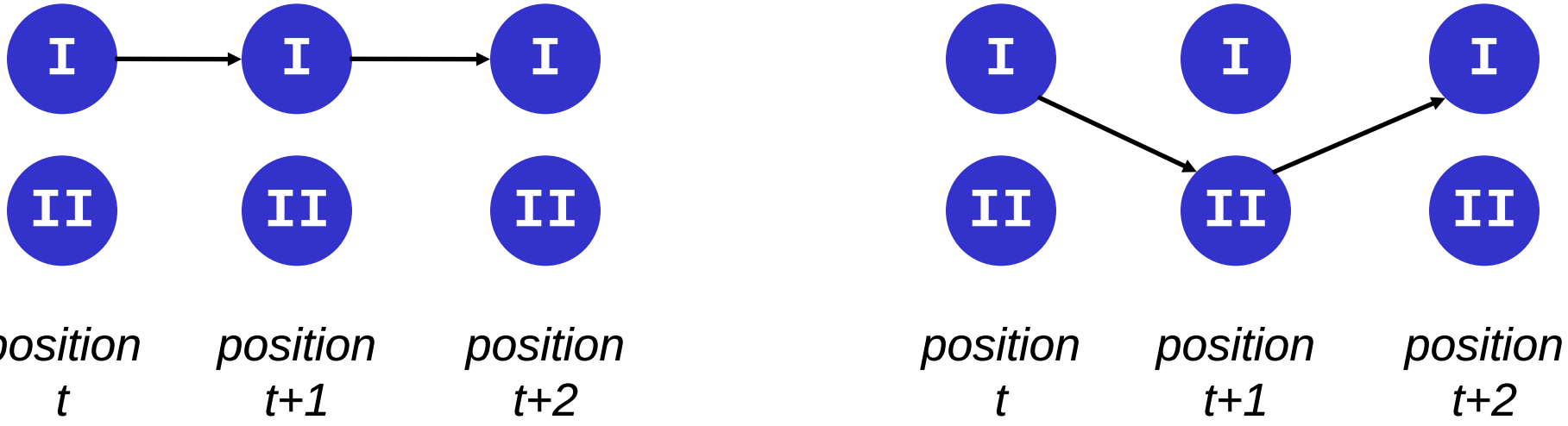
Markov processes

Thus:

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

0.6490 = P(X_{t+2} = **I** | X_t = **I**) is composed of two probabilities:

probability of this route *plus* probability of this route

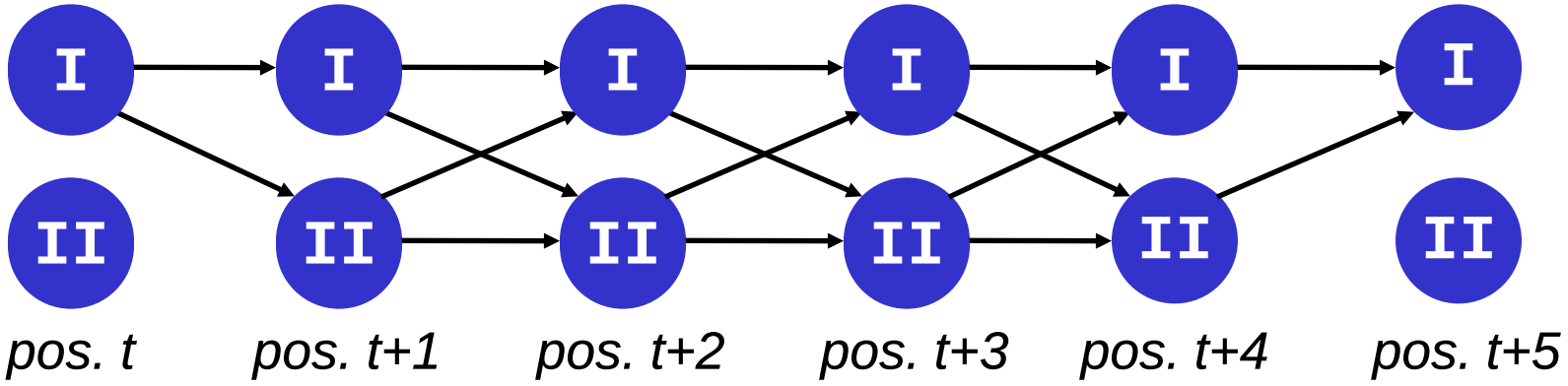


Markov processes

In similar fashion we may obtain:

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

Sum over probs. of all possible paths between 2 states:



Similarly:

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

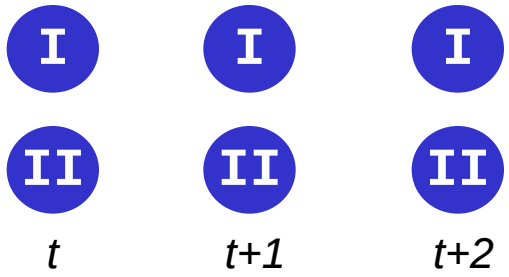
Markov processes

Question: consider the transition matrix:

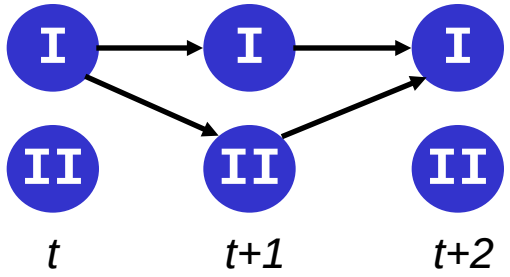
$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Then $(\mathbf{P}^2)_{1,1}$ corresponds to the probability over paths:

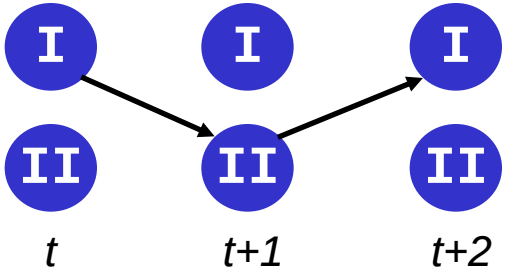
A)



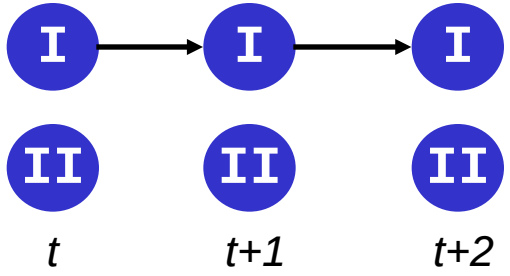
B)



C)



D)



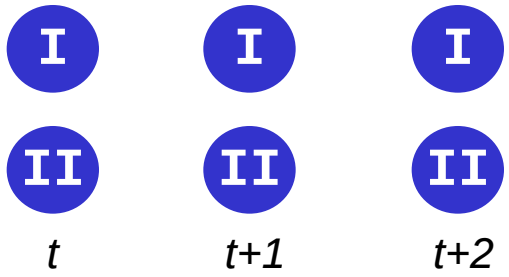
Markov processes

Question: consider the transition matrix:

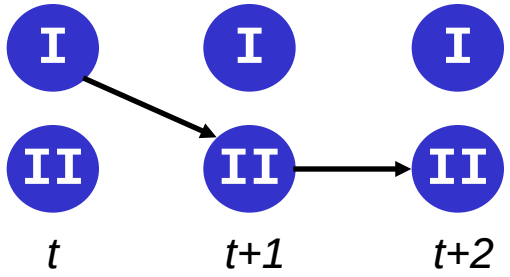
$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Then $(\mathbf{P}^2)_{1,2}$ corresponds to the probability over paths:

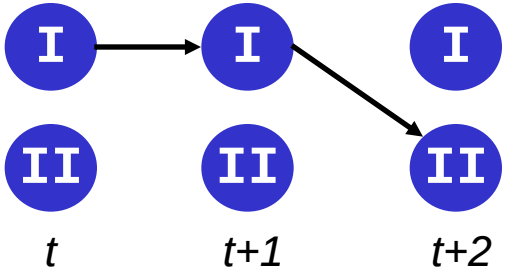
A)



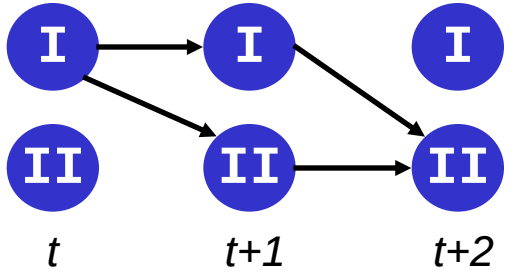
B)



C)



D)

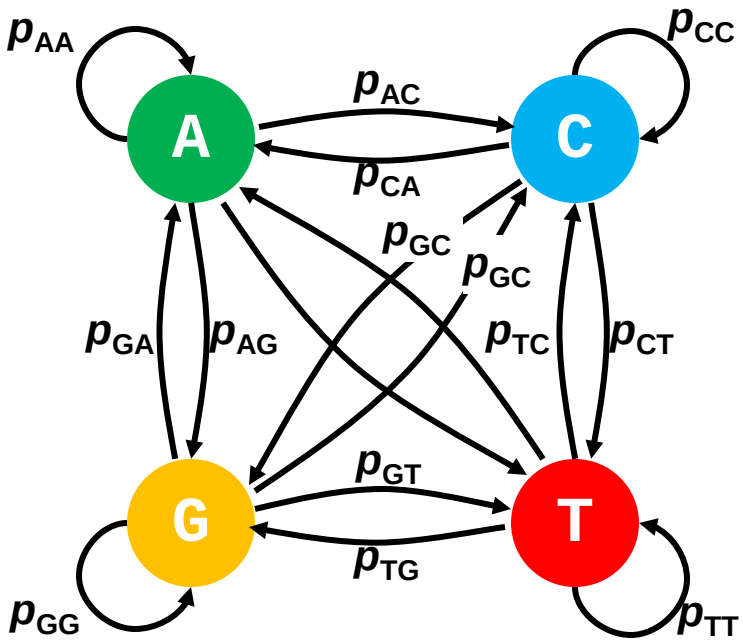


Markov process

How to sample from a Markov process?

Consider the DNA example.

$$\pi = \begin{matrix} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{matrix} \\ (0.45, 0.05, 0.25, 0.25)^T$$



$$\mathbf{P} = \begin{matrix} \text{from} \\ \\ \end{matrix} \begin{matrix} \\ \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{matrix} \overbrace{\begin{matrix} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{matrix}}^{\textit{to}} \\ \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.7 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.5 & 0.1 \end{pmatrix}$$

Markov process

1. Sample the first base from the initial distribution $\boldsymbol{\pi}$:

   $P(X_1=\mathbf{A}) = 0.45$, $P(X_1=\mathbf{C}) = 0.10$, et cetera

   Suppose we draw an **A**. Our DNA sequence now consists of one base, namely **A**.

2. Sample the next base using **P** given the previous base. Here $X_1=\mathbf{A}$, thus we sample from:

   $P(X_2=\mathbf{A} \mid X_1=\mathbf{A}) = 0.1$, $P(X_2=\mathbf{C} \mid X_1=\mathbf{A}) = 0.1$, et cetera.

   Our DNA sequence after the second step: **AT**

3. The last step is iterated until the last base.
   DNA sequence: **ATCCGATGC**

Andrey Markov

In the famous first application of Markov chains, Andrey Markov studied the sequence of 20,000 letters in Pushkin's poem "Eugeny Onegin", discovering that
- the stationary vowel probability is 0.432,
- the probability of a vowel following a vowel is 0.128, and
- the probability of a vowel following a consonant is 0.663.

Markov also studied the sequence of **100,000** letters in Aksakov's novel "The Childhood of Bagrov, the Grandson".

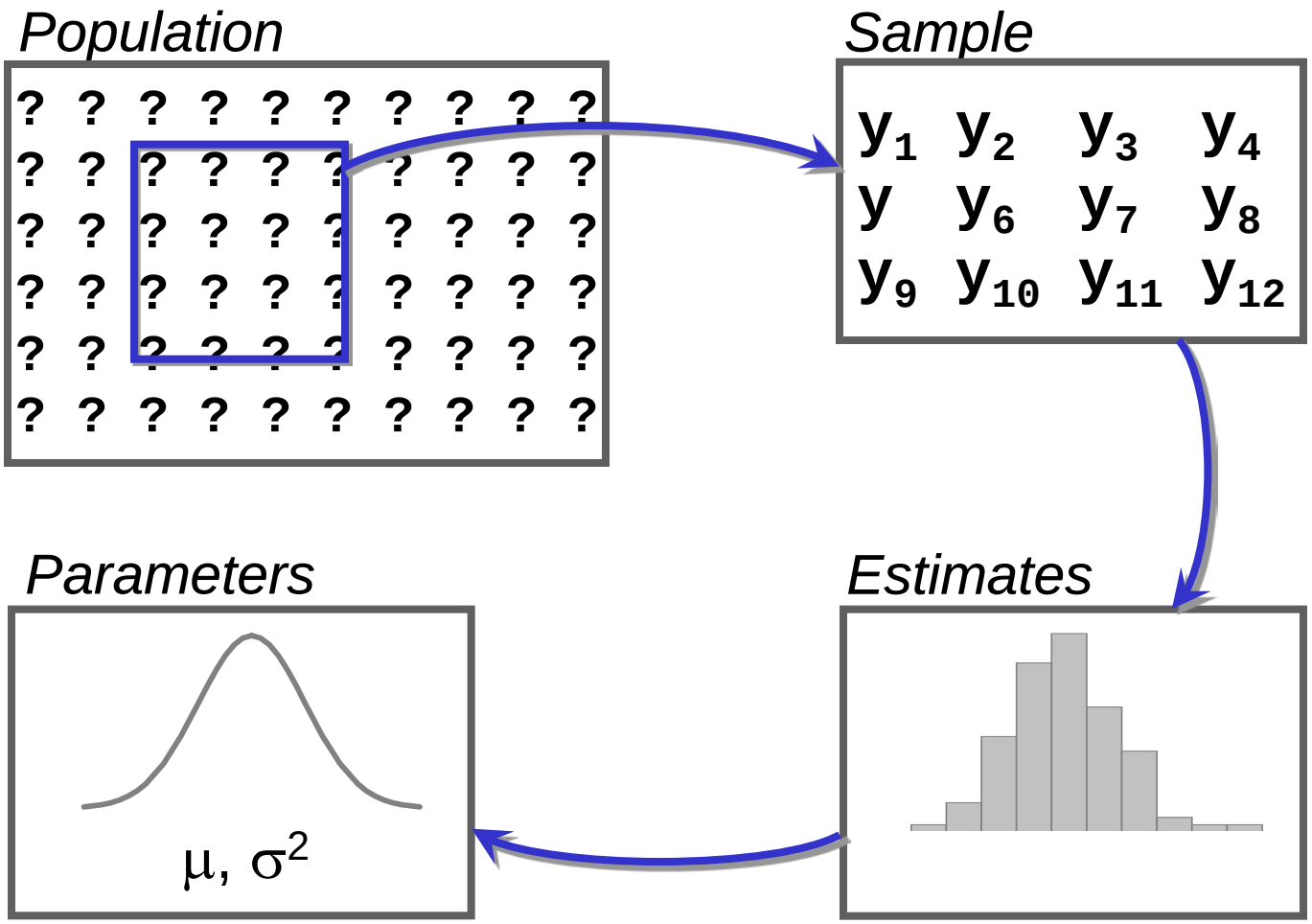
This was in 1913, long before the computer age!

Basharinl et al. (2004); Picture: http://en.wikipedia.org/

Maximum likelihood estimation

Maximum likelihood estimation

Estimation (from a sample)



Maximum likelihood estimation

Likelihood

The *likelihood* is a function from the model parameter space, where the parameter (say) $\boldsymbol{\theta}$ lives, to the probability space:

$$L : \boldsymbol{\theta} \rightarrow [0, 1]$$

The likelihood yields the probability of the observed data **X** for any parameter choice:

$$L(\mathbf{X}; \boldsymbol{\theta}) = P(\mathbf{X}_1, \ldots, \mathbf{X}_n; \boldsymbol{\theta})$$

If the observations are independent, this factorizes to:

$$L(\mathbf{X}; \boldsymbol{\theta}) = \prod_{i=1}^{n} P(\mathbf{X}_i; \boldsymbol{\theta})$$

Maximum likelihood estimation

Likelihood (example)

Let X_i be a random variable representing the outcome of tossing of a coin, either head (**H**) or tail (**T**). The typical distributional assumption for X_i is:

$$X \sim \text{Bernoulli}(\theta)$$

where $\theta = P(X_i = \mathtt{T})$.

Assume four realizations of X_i are available:

$$\mathbf{x} = (x_1, x_2, x_3, x_3) = (\mathtt{H}, \mathtt{H}, \mathtt{T}, \mathtt{H})$$

Construct the likelihood of for these data.

Maximum likelihood estimation

Likelihood (example)

The likelihood is:

$$
\begin{aligned}
& L(\mathbf{X} = \mathbf{x}; \theta) \\
= \quad & P(\mathbf{X} = \mathbf{x}; \theta) \quad = \quad \prod_{i=1}^{4} P(X_i = x_i; \theta) \\
= \quad & P(X_1 = \mathtt{H}; \theta) \times P(X_2 = \mathtt{H}; \theta) \\
& \times P(X_3 = \mathtt{T}; \theta) \times P(X_4 = \mathtt{H}; \theta) \\
= \quad & (1-\theta) \times (1-\theta) \times \theta \times (1-\theta) \\
= \quad & \theta\,(1-\theta)^3.
\end{aligned}
$$

Maximum likelihood estimation

Maximum likelihood

The parameter value for which the likelihood attains its maximum is referred to as the *maximum likelihood estimate*:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} L(\mathbf{X}; \boldsymbol{\theta})$$

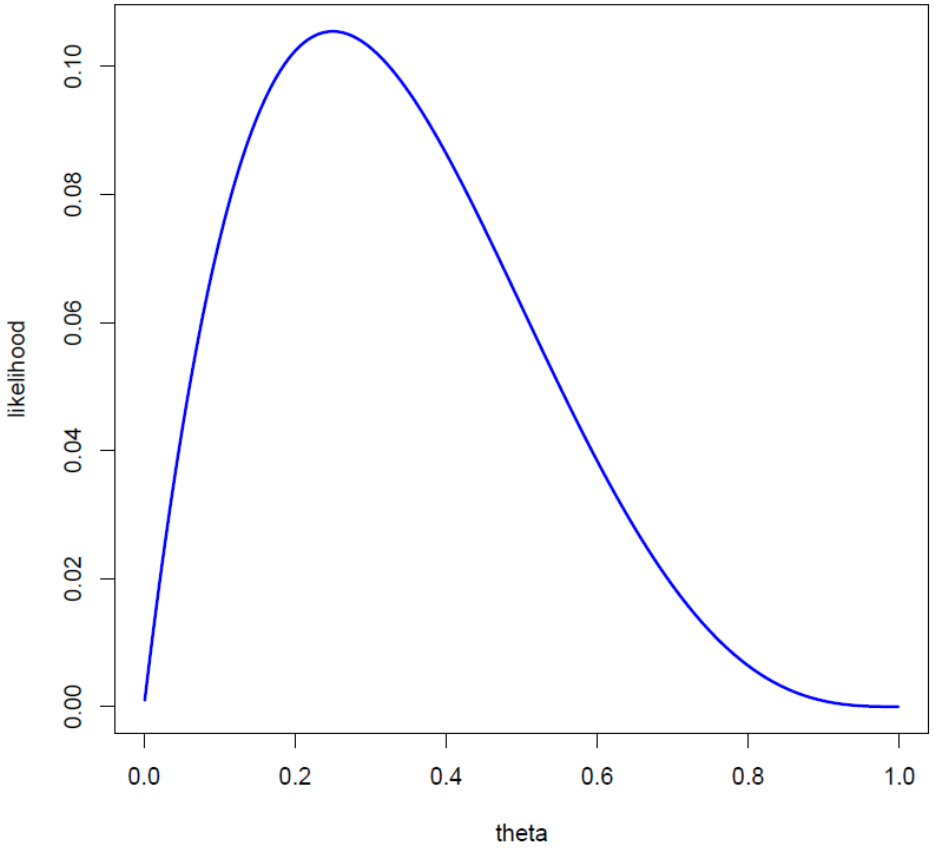
The maximum likelihood estimate is the parameter value that, given the model, is most likely to have given rise to the observed data.

Or loosely, that explains the observations best.

Maximum likelihood estimation

Likelihood (example continued)

Return to the ‘coin example’. Its likelihood:



Question: How does one find its maximum?

Maximum likelihood estimation

Likelihood (example continued)

Return to the 'coin example'. The maximum likelihood estimate is found by maximizing the likelihood. Due to the concavity of the likelihood, this is equivalent to finding the maximum of the log-likelihood:

$$\log[L(\mathbf{X} = \mathbf{x}; \theta)] \quad = \quad \log(\theta) + 3\log(1-\theta).$$

Equate the derivative w.r.t. θ to zero:

$$\frac{\partial}{\partial\theta}\log[L(\mathbf{X} = \mathbf{x}; \theta)] \quad = \quad \frac{1}{\theta} - \frac{3}{1-\theta} \quad = \quad 0.$$

And solve for θ to arrive at:

$$\hat{\boldsymbol{\theta}}_{ML} = \frac{1}{4}.$$

Maximum likelihood estimation

Maximum likelihood estimation

A general procedure, based on an appealing principle, to derive estimators.

ML estimation comprises:

→ specification of the distribution of the random variable.

→ formulation of the likelihood.

→ taking the logarithm of the likelihood (for convenience)

→ search for the (location of the) maximum:

- take derivative with respect to parameters,
- equate derivative to zero,
- find zeros of this equation.

Maximum likelihood estimation

Likelihood (example 2)

Let Y_i be a continuous random variable following a normal distribution:

$$Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

Obtain the likelihood for continuous random variables, the density (instead of probability) is used. That of the normal distribution is well-known.

Assuming the independence assumption, the likelihood for a collection of *n* samples is:

$$L(\mathbf{Y} = \mathbf{y}; \mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp[-(y_i - \mu)^2/(2\sigma^2)]$$

Maximum likelihood estimation

Likelihood (example 2)

Take the logarithm and obtain the log-likelihood:

$$\sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - (y_i - \mu)^2/(2\sigma^2) \right\}$$

The log-likelihood is maximized at:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

average estimates the mean

sum of squared deviations from the mean estimates the variance

Parameter estimation

Parameter estimation

Likelihood:

Given the 1st order Markov model, we may write down the likelihood of an observed sequence:

$$P(\mathtt{GCTAATTCAGCT}; \mathbf{P})$$

The likelihood is a function from the model parameter space, where transition matrix **P** lives, to the probability space [0,1]. It yields the probability of the observed data for any parameter choice.

Note

The likelihood does not factorize as before as contiguous DNA bases are not independent.

Parameter estimation

The likelihood for the sequence **AG**:

$$P(X_1 = \mathtt{A}, X_2 = \mathtt{G})$$
$$= \quad P(X_2 = \mathtt{G} \,|\, X_1 = \mathtt{A})\, P(X_1 = \mathtt{A})$$

The likelihood for the sequence **AGT**:

$$P(X_1 = \mathtt{A}, X_2 = \mathtt{G}, X_3 = \mathtt{T})$$
$$= \quad P(X_3 = \mathtt{T} \,|\, X_1 = \mathtt{A}, X_2 = \mathtt{G})$$
$$\times P(X_1 = \mathtt{A}, X_2 = \mathtt{G})$$
$$= \quad P(X_3 = \mathtt{T} \,|\, X_2 = \mathtt{G})$$
$$\times P(X_2 = \mathtt{G} \,|\, X_1 = \mathtt{A})\, P(X_1 = \mathtt{A})$$

Parameter estimation

More general, using the definition of conditional probability, the likelihood can be decomposed as follows:

$$
\begin{aligned}
L(\mathbf{X}) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T) \\
&= P(X_T = x_T \,|\, X_1 = x_1, \ldots, X_{T-1} = x_{T-1}) \\
&\quad \times P(X_{T-1} = x_{T-1} \,|\, X_1 = x_1, \ldots, X_{T-2} = x_{T-2}) \\
&\quad \ldots \\
&\quad \times P(X_2 = x_2 \,|\, X_1 = x_1) \\
&\quad \times P(X_1 = x_1)
\end{aligned}
$$

where x_1, …, x_T in $S = \{E_1, \ldots, E_S\}$.

Parameter estimation

Using the Markov property, we obtain:

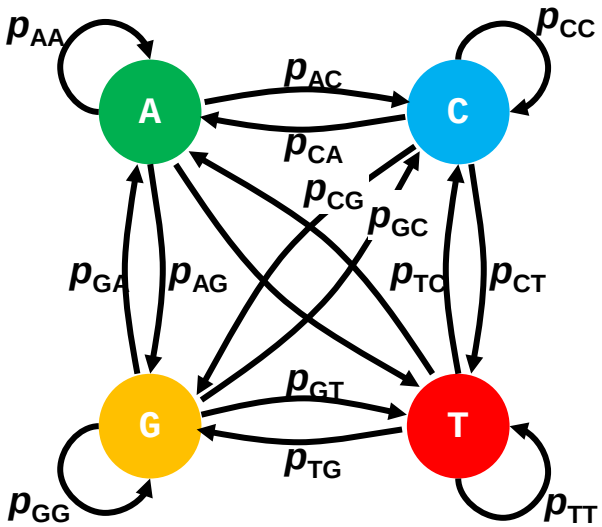
$$
\begin{aligned}
L(\mathbf{X}) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T) \\
&= P(X_T = x_T \,|\, X_{T-1} = x_{T-1}) \\
&\quad \times P(X_{T-1} = x_{T-1} \,|\, X_{T-2} = x_{T-2}) \\
&\quad \ldots \\
&\quad \times P(X_2 = x_2 \,|\, X_1 = x_1) \\
&\quad \times P(X_1 = x_1) \\
&= P(X_1 = x_1) \prod_{t=2}^{T} P(X_t = x_t \,|\, X_{t-1} = x_{t-1})
\end{aligned}
$$

Parameter estimation

Question

Consider the 1st order Markov model with initial distribution $\boldsymbol{\pi}$ = (0, 0.5, 0.5, 0) and transition matrix:

$$\mathbf{P} = \begin{array}{c} \\ \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \overbrace{\begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \\ \hline \end{array}}^{\textit{to}} \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$



Express the likelihood of the sequence:

GCATA

in terms of the parameters of the 1st order Markov chain.

Parameter estimation

The indicator function:

$$I_{\{X=\mathtt{A}\}} = \begin{cases} 1 & \text{if} \quad X = \mathtt{A} \\ 0 & \text{if} \quad X \neq \mathtt{A} \end{cases}$$

Recall:

$$\begin{aligned} \prod_{k=1}^{5} x^{a_k} &= x^{a_1}\, x^{a_2}\, x^{a_3}\, x^{a_4}\, x^{a_5} \\ &= x^{a_1+a_2+a_3+a_4+a_5} \\ &= x^{\sum_{k=1}^{5} a_k} \end{aligned}$$

Parameter estimation

Furthermore, e.g.:

$$P(X_t = x_t \mid X_{t-1} = x_{t-1})$$

$$= p_{AA}^{I_{\{X_t=A, X_{t-1}=A\}}} p_{AC}^{I_{\{X_t=C, X_{t-1}=A\}}} p_{AG}^{I_{\{X_t=T, X_{t-1}=A\}}} p_{AT}^{I_{\{X_t=T, X_{t-1}=A\}}}$$

$$p_{CA}^{I_{\{X_t=A, X_{t-1}=C\}}} p_{CC}^{I_{\{X_t=C, X_{t-1}=C\}}} p_{CG}^{I_{\{X_t=T, X_{t-1}=C\}}} p_{CT}^{I_{\{X_t=T, X_{t-1}=C\}}}$$

$$p_{GA}^{I_{\{X_t=A, X_{t-1}=G\}}} p_{GC}^{I_{\{X_t=C, X_{t-1}=G\}}} p_{GG}^{I_{\{X_t=T, X_{t-1}=G\}}} p_{GT}^{I_{\{X_t=T, X_{t-1}=G\}}}$$

$$p_{TA}^{I_{\{X_t=A, X_{t-1}=T\}}} p_{TC}^{I_{\{X_t=C, X_{t-1}=T\}}} p_{TG}^{I_{\{X_t=T, X_{t-1}=T\}}} p_{TT}^{I_{\{X_t=T, X_{t-1}=T\}}}$$

where only one transition probability at the time enters the likelihood, due to the indicator function.

Parameter estimation

The likelihood then becomes:

$$
\begin{aligned}
L(\mathbf{X}) &= P(X_1 = x_1) \\
&\quad \times p_{AA}^{\sum_{t=2}^{T} I_{\{X_t = A, X_{t-1} = A\}}} \cdot \ldots \cdot p_{TT}^{\sum_{t=2}^{T} I_{\{X_t = T, X_{t-1} = T\}}} \\
&= P(X_1 = x_1)\, p_{AA}^{n_{AA}} \cdot \ldots \cdot p_{TT}^{n_{TT}}
\end{aligned}
$$

where, e.g.,

$$
n_{AA} = \sum_{t=2}^{T} I_{\{X_t = A, X_{t-1} = A\}}
$$

Parameter estimation

Recall that, e.g.,

$$p_{AA} + p_{AC} + p_{AG} + p_{AT} = 1.$$

Or,

$$p_{AT} = 1 - p_{AA} - p_{AC} - p_{AG}.$$

Substitute this in the likelihood, and take the logarithm to arrive at the log-likelihood.

Note: it is irrelevant which transition probability is substituted.

Parameter estimation

The log-likelihood:

$$
\begin{aligned}
&\log[L(\mathbf{X})] \\
&\quad = \quad \log[P(X_1 = x_1)] \\
&\qquad + n_{AA}\,\log(p_{AA}) + n_{AC}\,\log(p_{AC}) \\
&\qquad + n_{AG}\,\log(p_{AG}) + n_{AT}\,\log(1 - p_{AA} - p_{AC} - p_{AG}) \\
&\qquad\quad \ldots \\
&\qquad + n_{TA}\,\log(p_{TA}) + n_{TC}\,\log(p_{TC}) \\
&\qquad + n_{TG}\,\log(p_{TG}) + n_{TT}\,\log(1 - p_{TA} - p_{TC} - p_{TG})
\end{aligned}
$$

Parameter estimation

Differentiation the log-likelihood yields, e.g.:

$$\frac{\partial \log[L(\mathbf{X})]}{\partial p_{AA}} = \frac{n_{AA}}{p_{AA}} - \frac{n_{AT}}{1 - p_{AA} - p_{AC} - p_{AG}}$$

Equate the derivatives to zero. This yields four systems of equations to solve, e.g.:

$$\begin{cases} n_{AT}\, p_{AA} &= n_{AA}\,(1 - p_{AA} - p_{AC} - p_{AG}) \\ n_{AT}\, p_{AC} &= n_{AC}\,(1 - p_{AA} - p_{AC} - p_{AG}) \\ n_{AT}\, p_{AG} &= n_{AG}\,(1 - p_{AA} - p_{AC} - p_{AG}) \end{cases}$$

Parameter estimation

The transition probabilities are then estimated by:

$$\hat{p}_{AA} = \frac{n_{AA}}{n_{AA} + n_{AC} + n_{AG} + n_{AT}}$$

$$\hat{p}_{AC} = \frac{n_{AC}}{n_{AA} + n_{AC} + n_{AG} + n_{AT}}$$

$$\ldots$$

$$\hat{p}_{TT} = \frac{n_{TT}}{n_{TA} + n_{TC} + n_{TG} + n_{TT}}$$

Verify 2nd order part. derivatives of log-likehood are negative!

book Axelson-Fisk (2010): Example 2.1.

Parameter estimation

Thus, for the following sequence: **ATCGATCGCA,** tabulation yields

$$\begin{pmatrix} \#\mathtt{AA} & \#\mathtt{AC} & \#\mathtt{AG} & \#\mathtt{AT} \\ \#\mathtt{CA} & \#\mathtt{CC} & \#\mathtt{CG} & \#\mathtt{CT} \\ \#\mathtt{GA} & \#\mathtt{GC} & \#\mathtt{GG} & \#\mathtt{GT} \\ \#\mathtt{TA} & \#\mathtt{TC} & \#\mathtt{TG} & \#\mathtt{TT} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 2 \\ 1 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix}$$

The maximum likelihood estimates thus become:

$$\hat{p}_{AA} = 0, \hat{p}_{AC} = 0, \hat{p}_{AG} = 0, \hat{p}_{AT} = 1,$$
$$\hat{p}_{CA} = \frac{1}{3}, \hat{p}_{CC} = 0, \hat{p}_{CG} = \frac{2}{3}, \hat{p}_{CT} = 0$$
$$\ldots, \ldots, \ldots, \ldots$$

Parameter estimation

```
> DNAseq <- c("A", "T", "C", "G", "A",
                "T", "C", "G", "C" "A")

> table(DNAseq)
DNAseq
A C G T
3 3 2 2

> table(DNAseq[1:9], DNAseq[2:10])
    A C G T
  A 0 0 0 2
  C 1 0 2 0
  G 1 1 0 0
  T 0 2 0 0
```

Parameter estimation

If one may assume the observed data is a realization of a stationary Markov chain, the initial distribution is estimated by the *stationary distribution* (discussed next time).

If only one realization of a stationary Markov chain is available and stationarity cannot be assumed, the initial distribution is estimated by:

$$\hat{\pi}_i \quad = \quad I_{\{X_1 = E_i\}}$$

Testing the order of the Markov chain

Testing the order of a Markov chain

Often the order of the Markov chain is unknown and needs to be determined from the data. This is done using a *Chi-square test* for independence.

Idea : assess validity of a 0^{th} order Markov chain using data.

→ Under a 0^{th} order Markov chain, e.g.: $P(\mathbf{G}, \mathbf{T}) = P(\mathbf{G})\, P(\mathbf{T})$

→ All quantities, $P(\mathbf{G}, \mathbf{T})$, $P(\mathbf{G})$ and $P(\mathbf{T})$, can be estimated from the data by their frequencies.

→ Comparison of these frequencies measures the fit of a 0^{th} order Markov chain:

obs. # dinucl. (**G**, **T**) - exp # dinucl. (**G**, **T**) .

→ Large difference indicate poor fit of 0^{th} order Markov chain.

Testing the order of a Markov chain

Consider a DNA-sequence. To assess the order, one first tests whether the sequence consists of independent letters. Hereto count the nucleotides, di-nucleotides and tri-nucleotides in the sequence:

$$
\begin{aligned}
N(x_t) &= \{\#t \mid X_t = x_t\} \\
N(x_t, x_{t+1}) &= \{\#t \mid X_t = x_t, X_{t+1} = x_{t+1}\} \\
N(x_t, x_{t+1}, x_{t+2}) &= \{\#t \mid X_t = x_t, X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}\}
\end{aligned}
$$

E.g. for $\{x_1, x_2, x_3, x_4\} = \{\mathbf{T}, \mathbf{G}, \mathbf{A}, \mathbf{G}\}$:

$\rightarrow N(x_t = \mathbf{G}) = 2$, $N(x_t = \mathbf{A}) = 1$, ….

$\rightarrow N(x_t = \mathbf{G}, x_{t+1} = \mathbf{A}) = 1$, $N(x_t = \mathbf{T}, x_{t+1} = \mathbf{A}) = 0$, ….

$\rightarrow N(x_t = \mathbf{A}, x_{t+1} = \mathbf{A}, x_{t+2} = \mathbf{A}) = 0$, ….

Testing the order of a Markov chain

Assumping independence, we can calculate the expected frequency of a di-nucleotide for a sequence of length *n*:

$$\begin{aligned} E(\#AG) &= (n-1)\, p(A)\, p(G) \\ &= (n-1)\, \frac{N(A)}{n}\, \frac{N(G)}{n} \end{aligned}$$

To test the order of a Markov chain, we now compare the observed and expected of (say) di-nucleotides.

$$N(AG) - \frac{(n-1)^2}{n^2}\, \frac{N(A)\, N(G)}{n-1}$$

goes to one as n → ∞

Testing the order of a Markov chain

Compare data from a 1st order Markov chain with what is expected on the basis of a 0th order Markov chain.

Observed dimer frequency

	A	C	G	T
A	98	103	109	65
C	95	216	37	185
G	81	128	62	13
T	100	87	76	44

Expected dimer frequency

	A	C	G	T
A	93.8	133.6	71.0	76.8
C	133.6	190.2	101.2	109.4
G	71.0	101.2	53.8	58.2
T	76.8	109.4	58.2	62.9

According the 0th order Markov model the CG-dimer ought to be observed ≈3x more than is done.

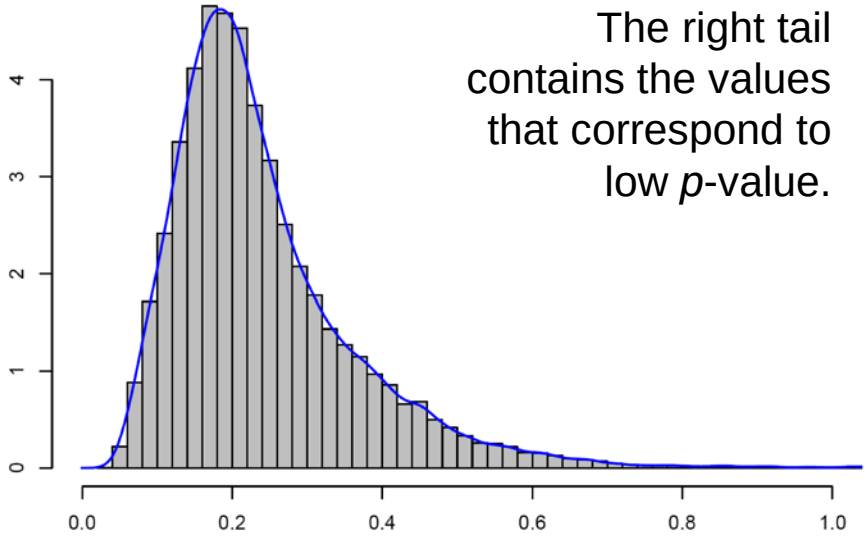
Testing the order of a Markov chain

The null hypothesis of independence between the letters of the DNA sequence evaluated by the χ^2 statistic:

$$\chi^2 = \sum_{x_t, x_{t+1} \in \{A,C,G,T\}} \frac{\left(N(x_t, x_{t+1}) - N(x_t)\, N(x_{t+1})/(n-1)\right)^2}{N(x_t)\, N(x_{t+1})/(n-1)}$$

which has (4-1) x (4-1) = 9 degrees of freedom.

If the null hypothesis cannot be rejected, one would fit a Markov model of order m=0. However, if H_0 can be rejected one would test for higher order dependence.



The right tail contains the values that correspond to low *p*-value.

Testing the order of a Markov chain

To test for 1st order dependence, consider to implications of the 1st order Markov chain model. E.g.:

$$P(X_t=\mathbf{G}, X_{t+1}=\mathbf{T}, X_{t+2}=\mathbf{A}) =$$

$$P(X_t=\mathbf{G}) \times P(X_{t+1}=\mathbf{T} \mid X_t=\mathbf{G}) \times P(X_{t+2}=\mathbf{A} \mid X_{t+1}=\mathbf{T})$$

Using the definition of conditional probability:

$$P(X_t=\mathbf{G}, X_{t+1}=\mathbf{T}, X_{t+2}=\mathbf{A}) =$$

$$P(X_{t+1}=\mathbf{T}, X_t=\mathbf{G}) \times P(X_{t+2}=\mathbf{A}, X_{t+1}=\mathbf{T}) / P(X_t=\mathbf{T}).$$

All these probabilities can be estimated directly from the data.

Testing the order of a Markov chain

The 1st order dependence hypothesis is evaluated by:

$$\sum_{x_t, x_{t+1}, x_{t+2} \in \{A,C,G,T\}} \frac{\left[N(x_t, x_{t+1}, x_{t+2}) - E\{N(x_t, x_{t+1}, x_{t+2})\}\right]^2}{E(\{N(x_t, x_{t+1}, x_{t+2})\}}$$

where

$$E\{N(x_t, x_{t+1}, x_{t+2})\} = \frac{N(x_t, x_{t+1})\, N(x_{t+1}, x_{t+2})}{N(x_{t+1})}$$

This is χ^2 distributed with (16-1) x (4-1) = 45 d.o.f..

Testing the order of a Markov chain

A 1st order Markov chain provides a reasonable description of the sequence of the hlyE gene of the E.coli bacteria.

Independence test:
Chi-sq stat: 22.45717, p-value: 0.00754

1st order dependence test:
Chi-sq stat: 55.27470, p-value: 0.14025

The sequence of the prrA gene of the E.coli bacteria requires a higher order Markov chain model.

Independence test:
Chi-sq stat: 33.51356, p-value: 1.266532e-04

1st order dependence test:
Chi-sq stat: 114.56290, p-value: 5.506452e-08

Testing the order of a Markov chain

The independence case, assuming the **`DNAseq`**-object is a **`character`**-object containing the sequence:

```
> # calculate nucleotide and dimer frequencies
> nuclFreq <- matrix(table(DNAseq), ncol=1)
> dimerFreq <- table(DNAseq[1:(length(DNAseq)-1)],
DNAseq[2:length(DNAseq)])

> # calculate expected dimer frequencies
> dimerExp <- nuclFreq %*% t(nuclFreq)/(length(DNAseq)-1)

> # calculate test statistic and p-value
> teststat <- sum((dimerFreq - dimerExp)^2 / dimerExp)
> pval <- exp(pchisq(teststat, 9, lower.tail=FALSE,
      log.p=TRUE))
```

Exercise: modify the code above for the 1st order test.

Example
---
sequence discrimination

Example: sequence discrimination

We have fitted 1st order Markov chain models to a representative coding and noncoding sequence of the E.coli bacteria.

$\mathbf{P}_{coding}$

	A	C	G	T
A	0.321	0.257	0.211	0.211
C	0.319	0.207	0.266	0.208
G	0.259	0.284	0.237	0.219
T	0.223	0.243	0.309	0.225

$\mathbf{P}_{noncoding}$

	A	C	G	T
A	0.320	0.278	0.231	0.172
C	0.295	0.205	0.286	0.214
G	0.241	0.261	0.233	0.265
T	0.283	0.238	0.256	0.223

These models can be used to discriminate between coding and noncoding sequences.

Example: sequence discrimination

For a new sequence with unknown function calculate the likelihood under the coding and noncoding model, e.g.:

$$\begin{aligned} L_{coding}(\mathbf{X}) &= P_{coding}(X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T) \\ &= \hat{\pi}_{coding}(x_1)\,(\hat{\mathbf{P}}_{coding})_{AA}^{n_{AA}} \cdot \ldots \cdot (\hat{\mathbf{P}}_{coding})_{TT}^{n_{TT}} \end{aligned}$$

These likelihoods are compared by means of their log ratio:

$$LR(\mathbf{X}) = \log[L_{coding}(\mathbf{X})] - \log[L_{noncoding}(\mathbf{X})]$$

the *log-likelihood ratio*.

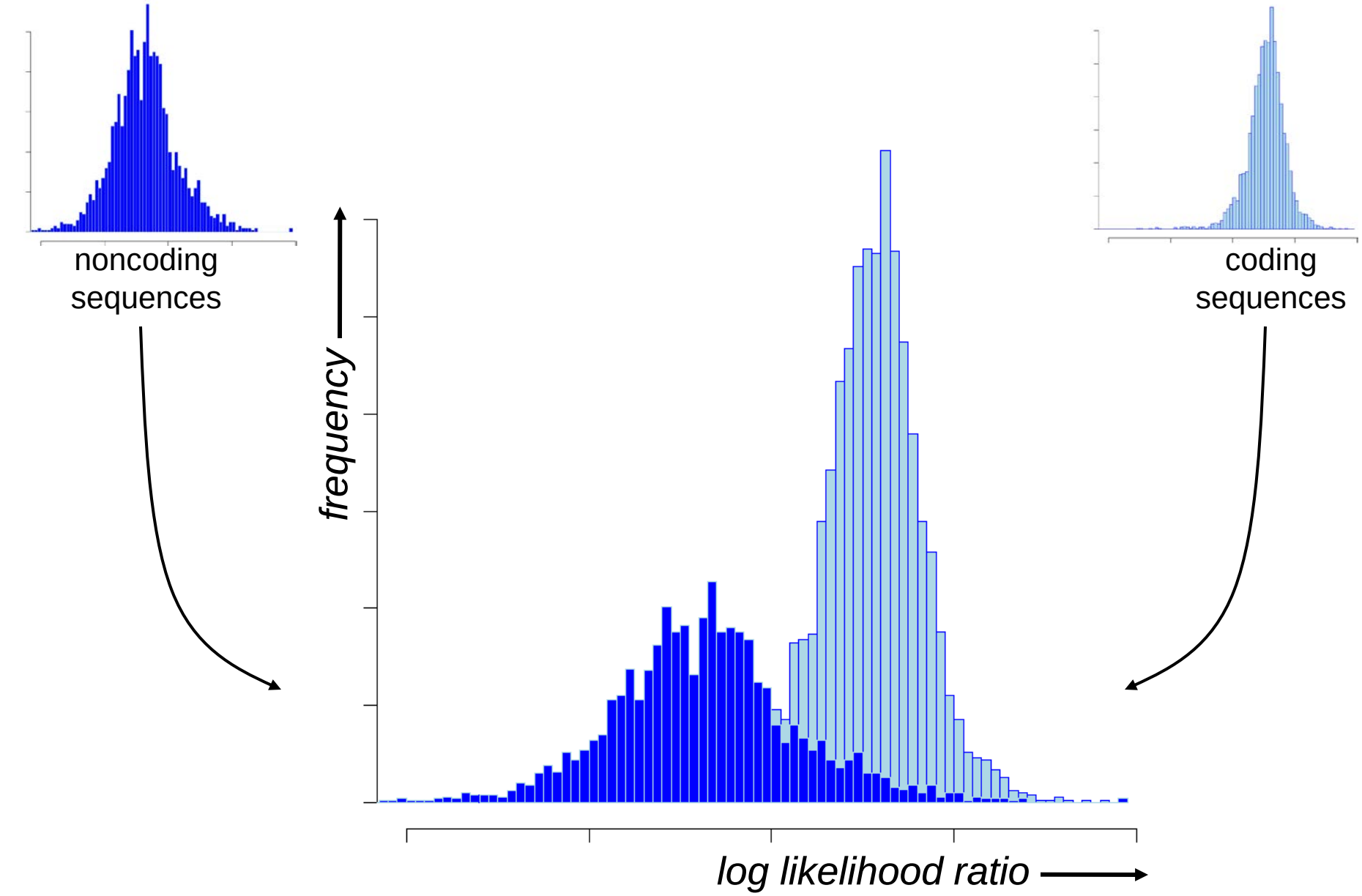
Example: sequence discrimination

If the log likelihood $LR(\mathbf{X})$ of a new sequence exceeds a certain threshold, the sequence is classified as coding and non-coding otherwise.

Back to E.coli

To illustrate the potential of the discrimination approach, we calculate the log likelihood ratio for large set of known coding and noncoding sequences. The distribution of the two sets of $LR(X)$'s are compared.

Example: sequence discrimination



Example: sequence discrimination

Conclusion E.coli example

Comparison of the distributions indicate that one could discriminate reasonably between coding and noncoding sequences on the basis of simple 1st order Markov.

Improvements:

- Higher order Markov model.
- Incorporate more structure information of the DNA.

Example

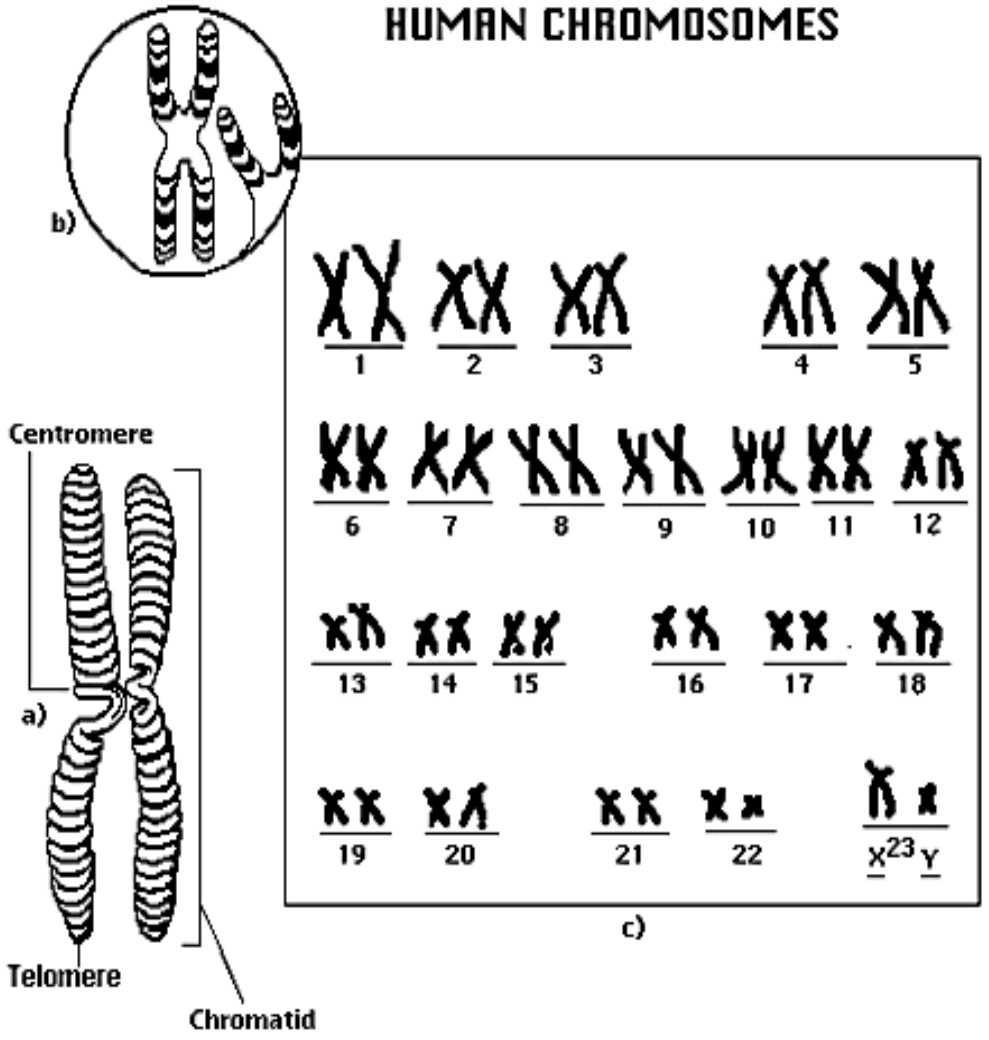
---

evolution of aberrations

Example

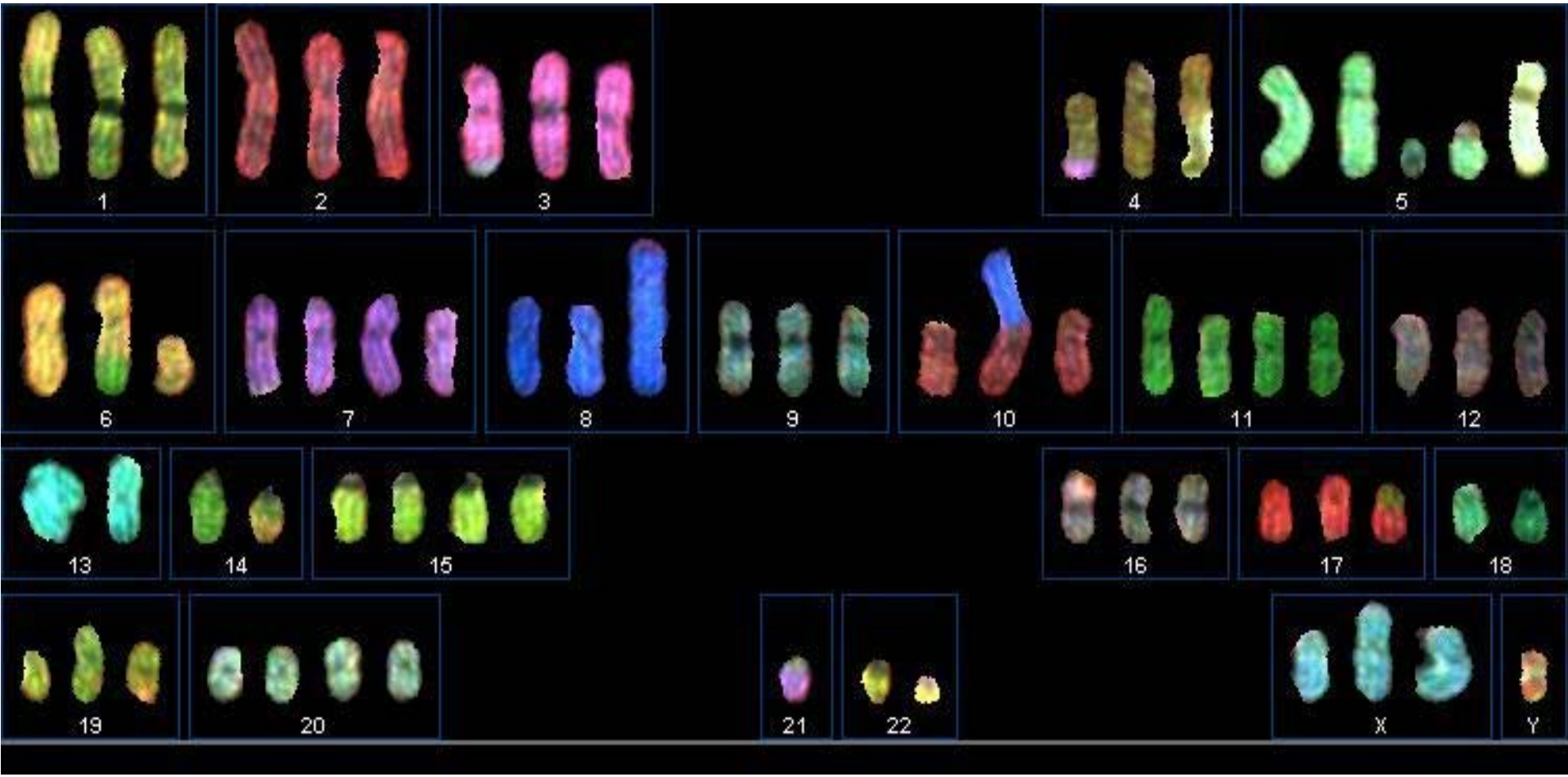
DNA copy number of a genomic segment is simply the number of copies of that segment present in the cell under study.

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

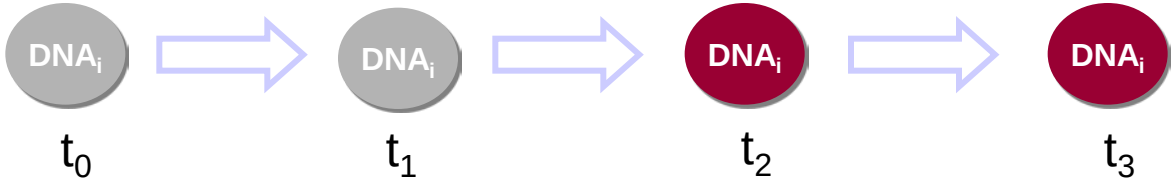
In cancer:

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

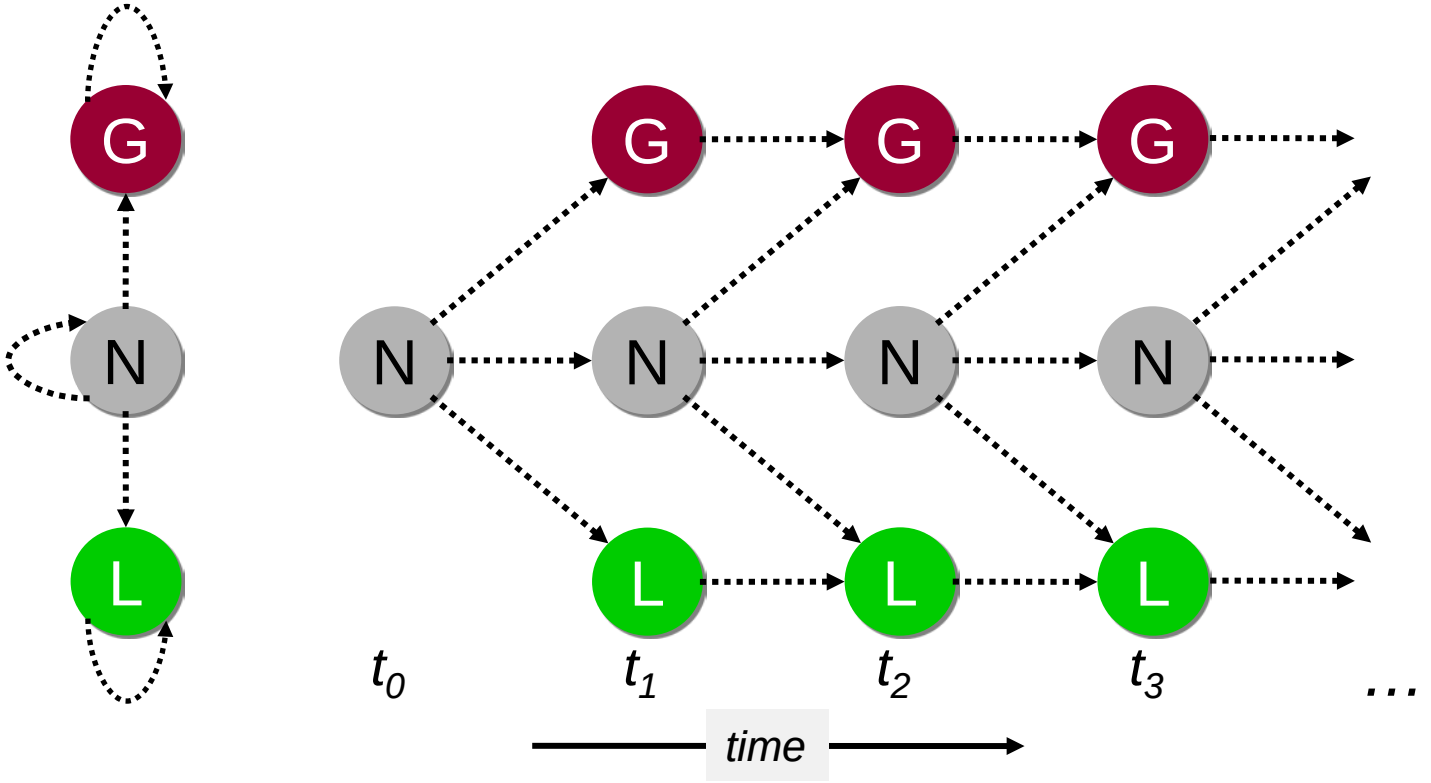
Let us model the accumulation process of DNA copy number aberrations.

Example

So far, we only considered one locus. Hence:



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:

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

Using:

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

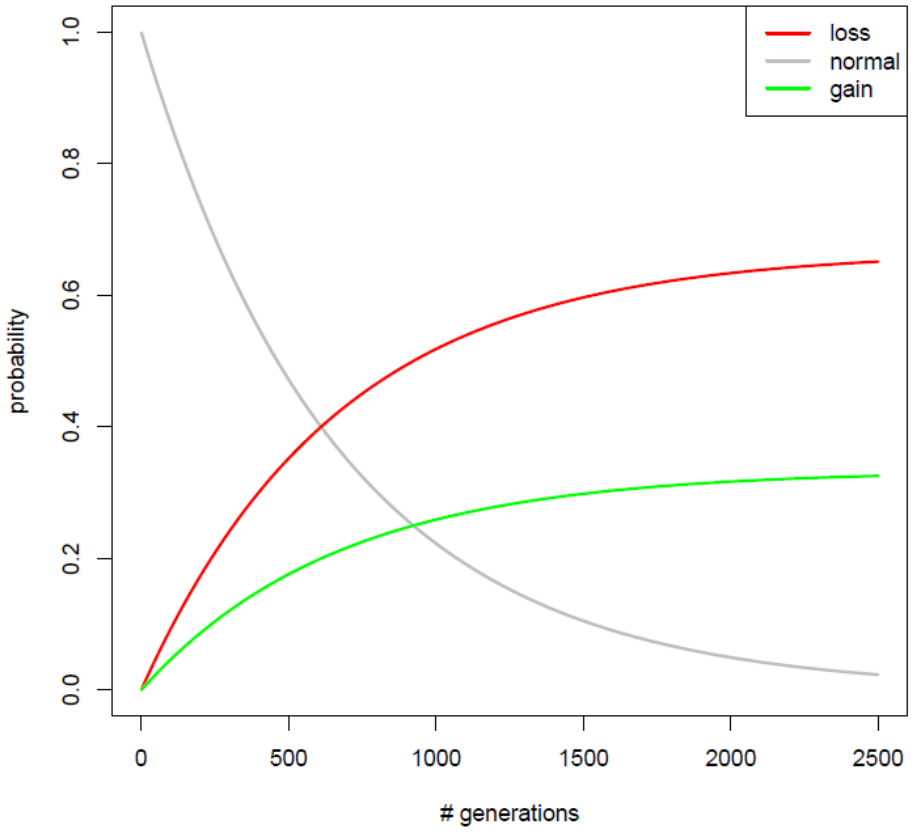
These probabilities simplify to, e.g.:

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

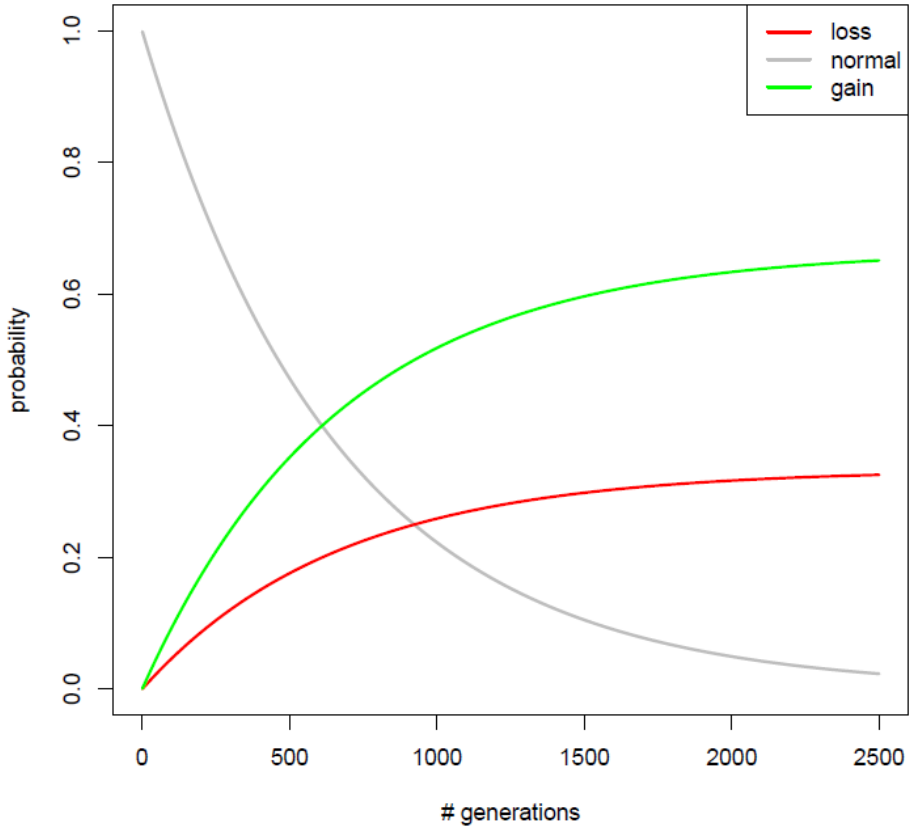
Example

Parameters α and β determine which aberration type is most prevalent.

α > β



α < β



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:

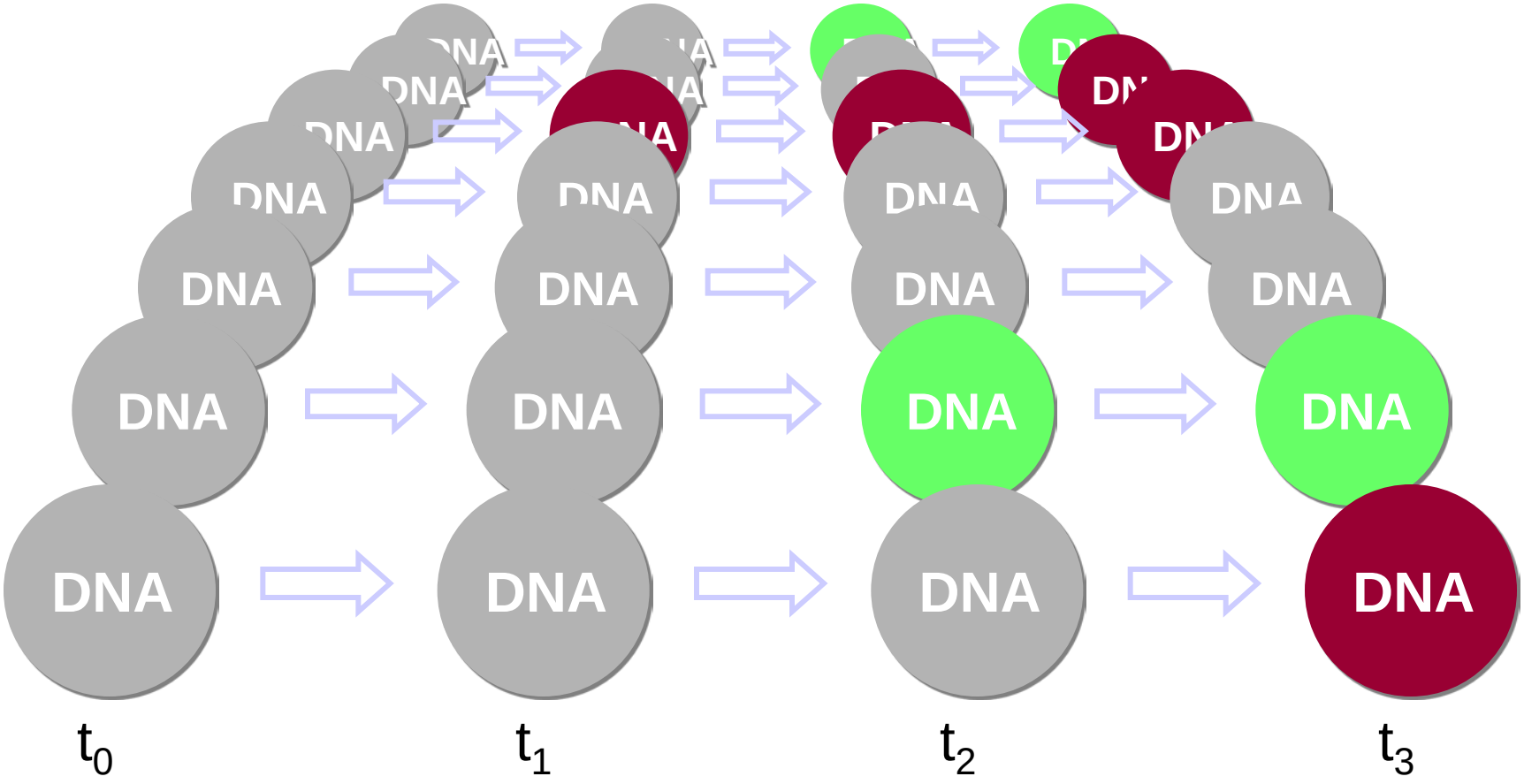
$$P(Y = p) = \lambda^p \, \exp(-\lambda)/p!$$

This yields, e.g.:

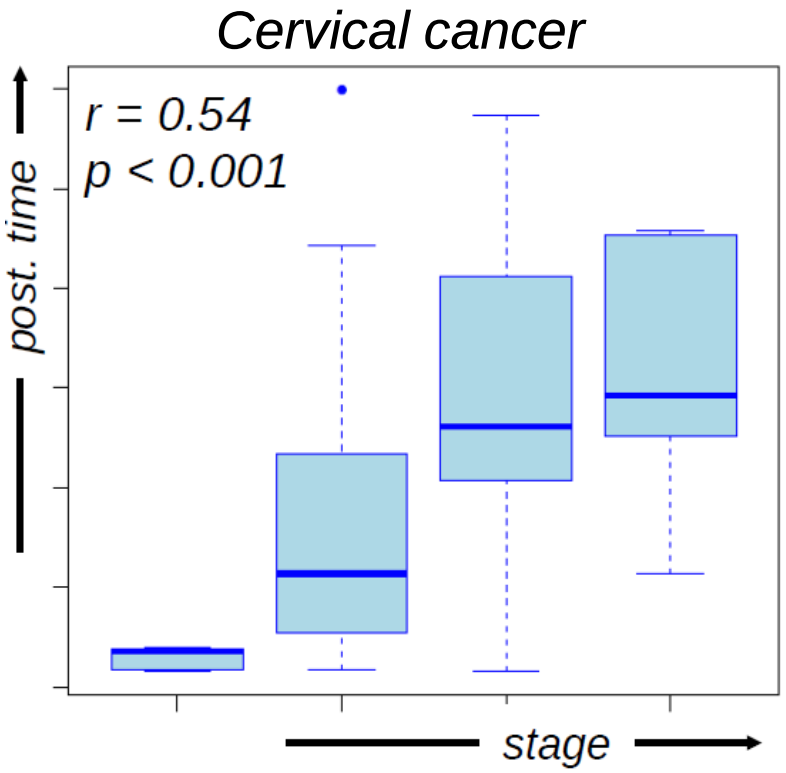
$$\begin{aligned} P(X = N) &= \sum_{p=0}^{\infty} P(X = N \,|\, Y = p) \, P(Y = p) \\ &= \sum_{p=0}^{\infty} (1 - \alpha - \beta)^p \, \lambda^p \, \exp(-\lambda)/p! \end{aligned}$$

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.

Supplementary material:
Proof of
Kolmogorov-Chapman
equations

SM: Markov processes

Proof of the Kolmogorov-Chapman equations:

$$
\begin{aligned}
(\mathbf{P}^{n+m})_{ij} &= P(X_{n+m} = j \mid X_0 = i) \\
&= \sum_{k=1}^{S} P(X_{n+m} = j, X_n = k \mid X_0 = i) \\
&= \sum_{k=1}^{S} P(X_{n+m} = j, X_n = k, X_0 = i)/P(X_0 = i) \\
&= \sum_{k=1}^{S} P(X_{n+m} = j \mid X_n = k, X_0 = i)\, P(X_n = k, X_0 = i)/P(X_0 = i) \\
&= \sum_{k=1}^{S} (\mathbf{P}^m)_{kj}\, P(X_n = k, X_0 = i)/P(X_0 = i) \\
&= \sum_{k=1}^{S} (\mathbf{P}^n)_{ik} (\mathbf{P}^m)_{kj}
\end{aligned}
$$

SM: Markov processes

Induction proof of $\mathbf{P}^{(n+1)} = \mathbf{P}^{n+1}$

Assume $\mathbf{P}^{(n)} = \mathbf{P}^{n}$

Then:

$$
\begin{aligned}
&(\mathbf{P}^{(n+1)})_{ij} \\
&\quad = P(X_{t+n+1} = E_j \,|\, X_t = E_i) \\
&\quad = \sum_{k=1}^{S} P(X_{t+n+1} = E_j, X_{t+n} = E_k \,|\, X_t = E_i) \\
&\quad = \sum_{k=1}^{S} P(X_{t+n+1} = E_j \,|\, X_{t+n} = E_k)\, P(X_{t+n} = E_k \,|\, X_t = E_i) \\
&\quad = \sum_{k=1}^{S} (\mathbf{P})_{kj} (\mathbf{P}^{(n)})_{ik} = \sum_{k=1}^{S} (\mathbf{P})_{kj} (\mathbf{P}^{n})_{ik} = (\mathbf{P}^{n+1})_{ij}
\end{aligned}
$$

References & further reading

References and further reading

Axelson-Fisk, M. (2010), *Comparative Gene Finding: Models, Algorithms and Implementation*, Springer.

Basharin, G.P., Langville, A.N., Naumov, V.A (2004), “The life and work of A.A. Markov”, *Linear Algebra and its Applications*, **386**, 3-26.

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.

Van Wieringen, W.N., Ros, B.P., Wilting, S.M. (2013), “Modeling the DNA copy number aberration patterns in observational high-throughput cancer data”, *Statistical Applications in Genetics and Molecular Biology, 12(2),*143-174.