Modelling DNA sequence evolution with interacting particle systems

Mikael Falconnet

Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne – CNRS

Chaire Modélisation Mathématique et Biodiversité,
Museum National d'Histoire Naturelle, the 16th of November 2012

Stochastic nucleotidic substitution models

Common assumptions of the usual models

- A DNA sequence is an element of $\{A, T, C, G\}^N$, $N \in \mathbb{N}^*$.
- **Independent** evolution of the sites according to a **Markovian kernel**.

Example: Jukes and Cantor model (1969)

- Rate matrix ($\lambda > 0$)

	A	T	C	G
A	·	λ	λ	λ
T	λ	·	λ	λ
C	λ	λ	·	λ
G	λ	λ	λ	·

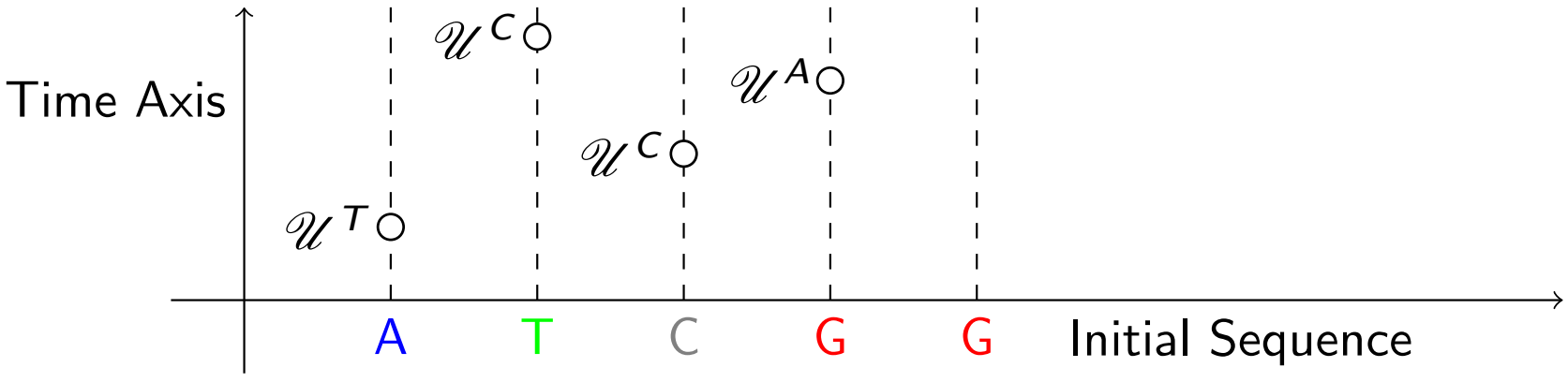
- **Diagonal** entry $-q_{aa}$ is the substitution rate of nucleotide a, here $q_{aa} = -3\lambda$.
- **Non-diagonal** entry q_{ab} is the substitution rate of nucleotide a by b, here $q_{ab} = \lambda$.

Modelisation

- At any site x, we run a **Poisson point process** with parameter 3λ.
- At any **point**, the nucleotide $\eta(x)$ is substituted by $a \in \{A, T, C, G\} \setminus \{\eta(x)\}$ with probability $1/3$.

Alternative, equivalent but faster modelisation

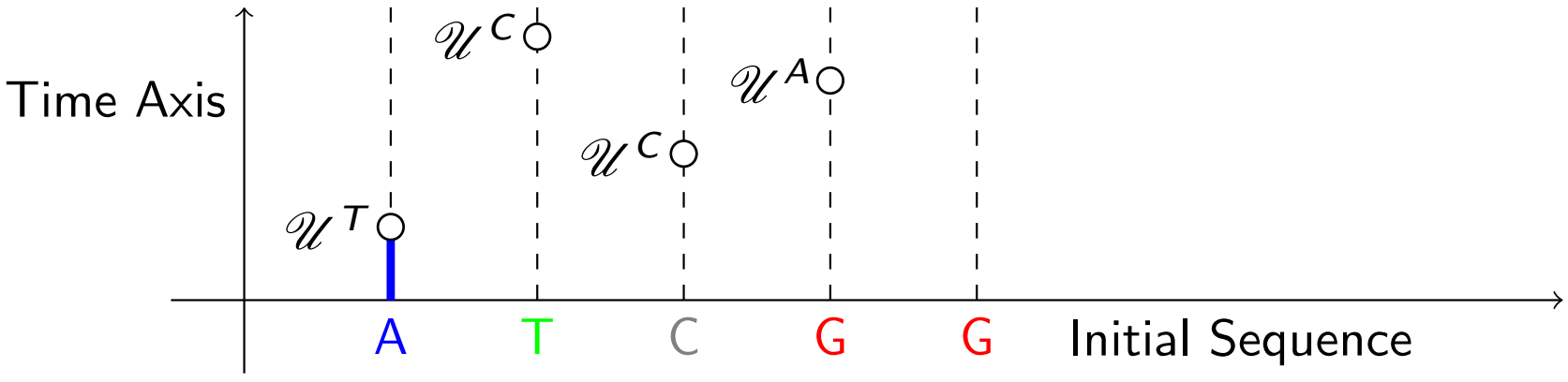
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

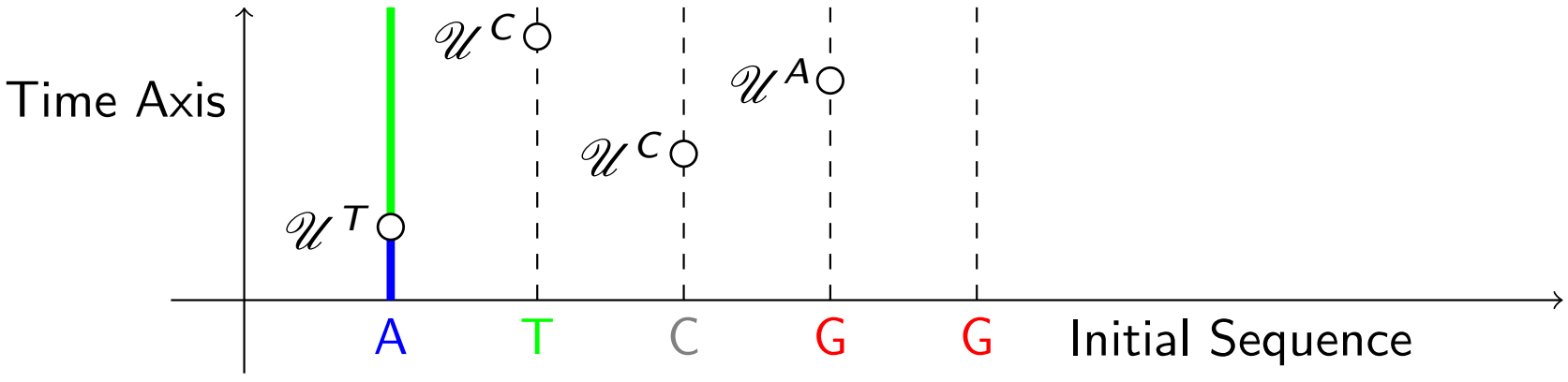
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

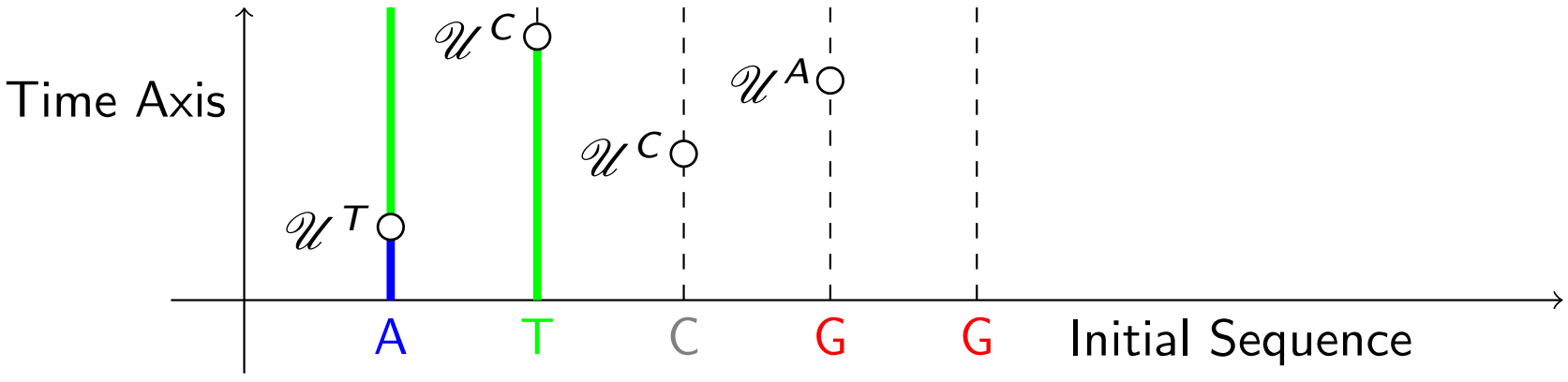
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

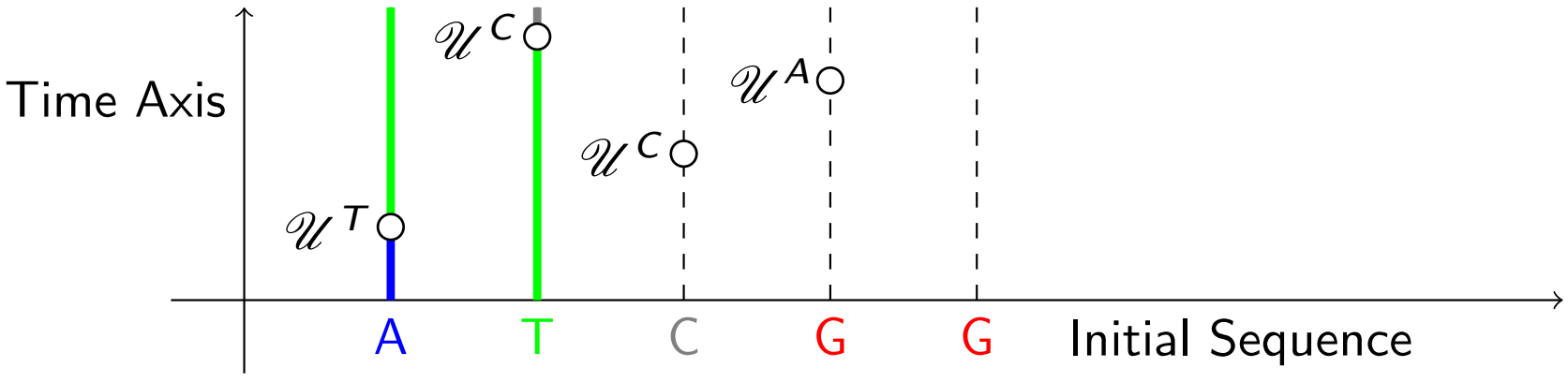
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

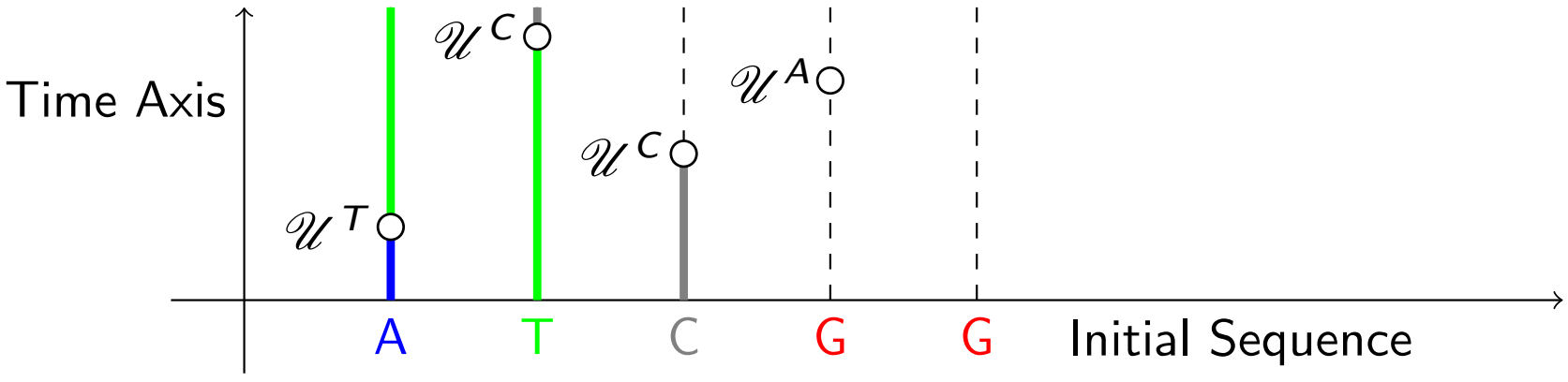
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

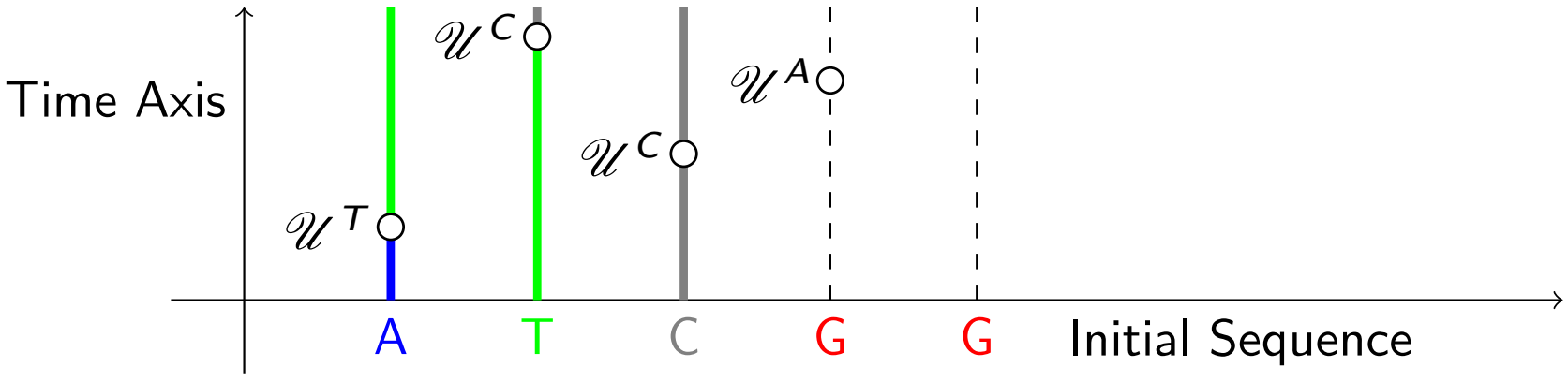
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

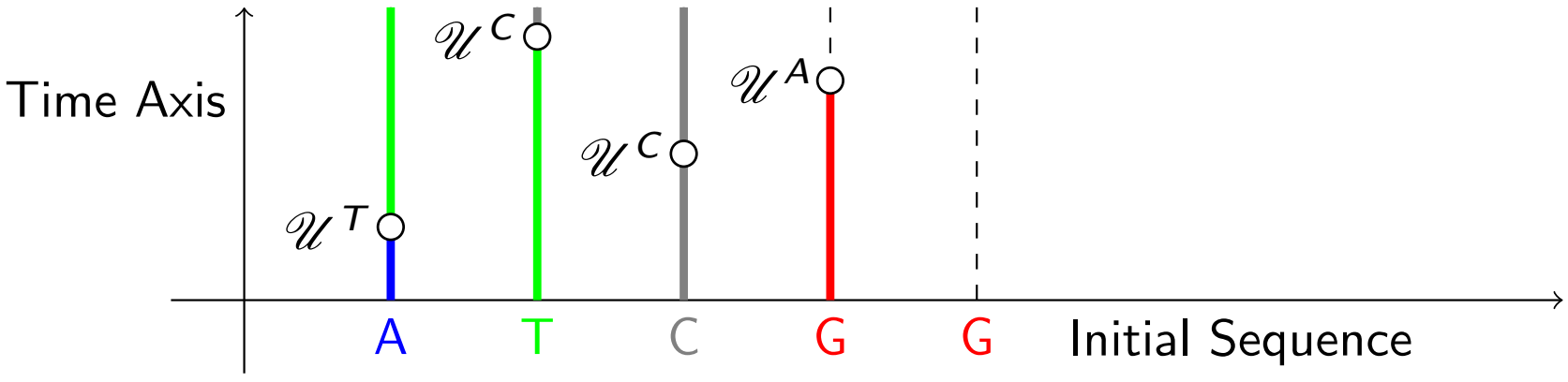
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

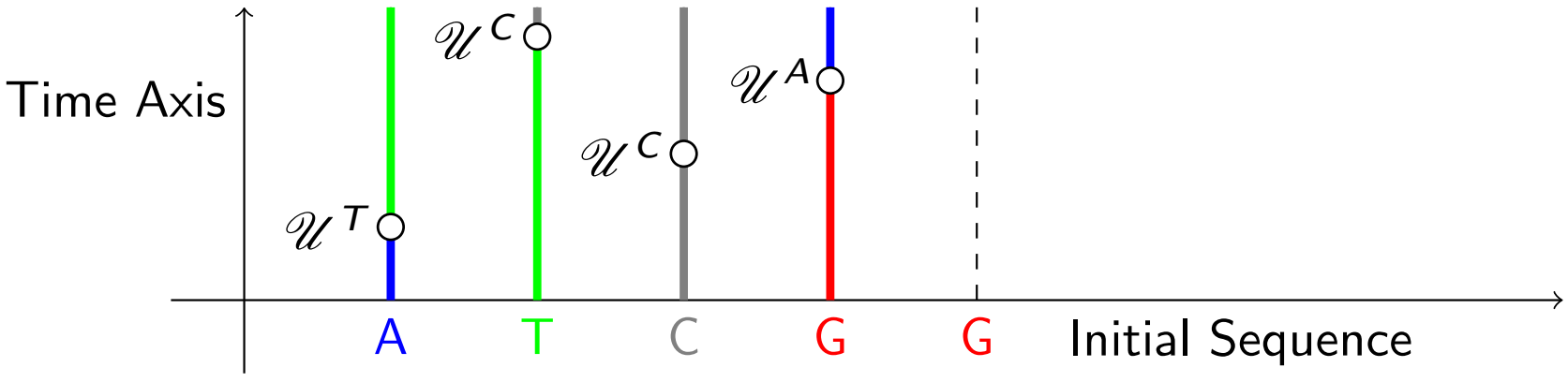
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

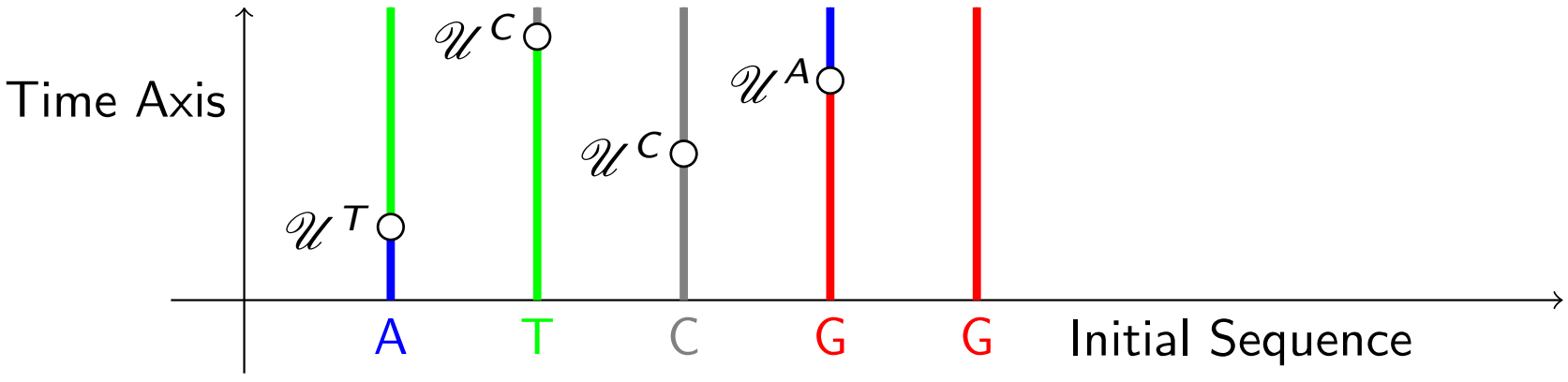
	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Alternative, equivalent but faster modelisation

	A	T	C	G
A	$\cdot$	λ	λ	λ
T	λ	$\cdot$	λ	λ
C	λ	λ	$\cdot$	λ
G	λ	λ	λ	$\cdot$



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	λ	... unconditionally.

Stochastic nucleotidic substitution models

Consequences

- Convergence in distribution at any site

- Convergenge in distribution of the whole sequence to the **product measure**.

Problems

- $(a_1 \dots a_\ell)_{\text{obs}} \neq (a_1)_{\text{obs}} \dots (a_\ell)_{\text{obs}}$.

- The substitution rate $\eta(x) \to a$ may **dépend** de $\eta(x-1)$, $\eta(x)$ and $\eta(x+1)$.

Famous example : CpG dinucleotides

- Rate C $\to$ T up to ten times larger when C is involved in a CpG (in fact C*pG).

JC+CpG model

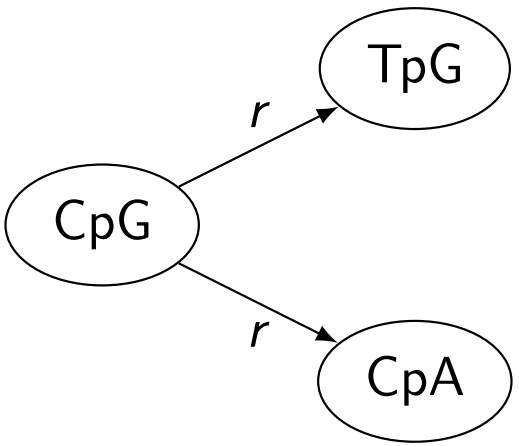
Bérard, Gouéré et Piau, *Mathematical Biosciences* (2008)

- A DNA sequence is now **doubly infinite**, that is, an element of $\{A, T, C, G\}^{\mathbb{Z}}$.

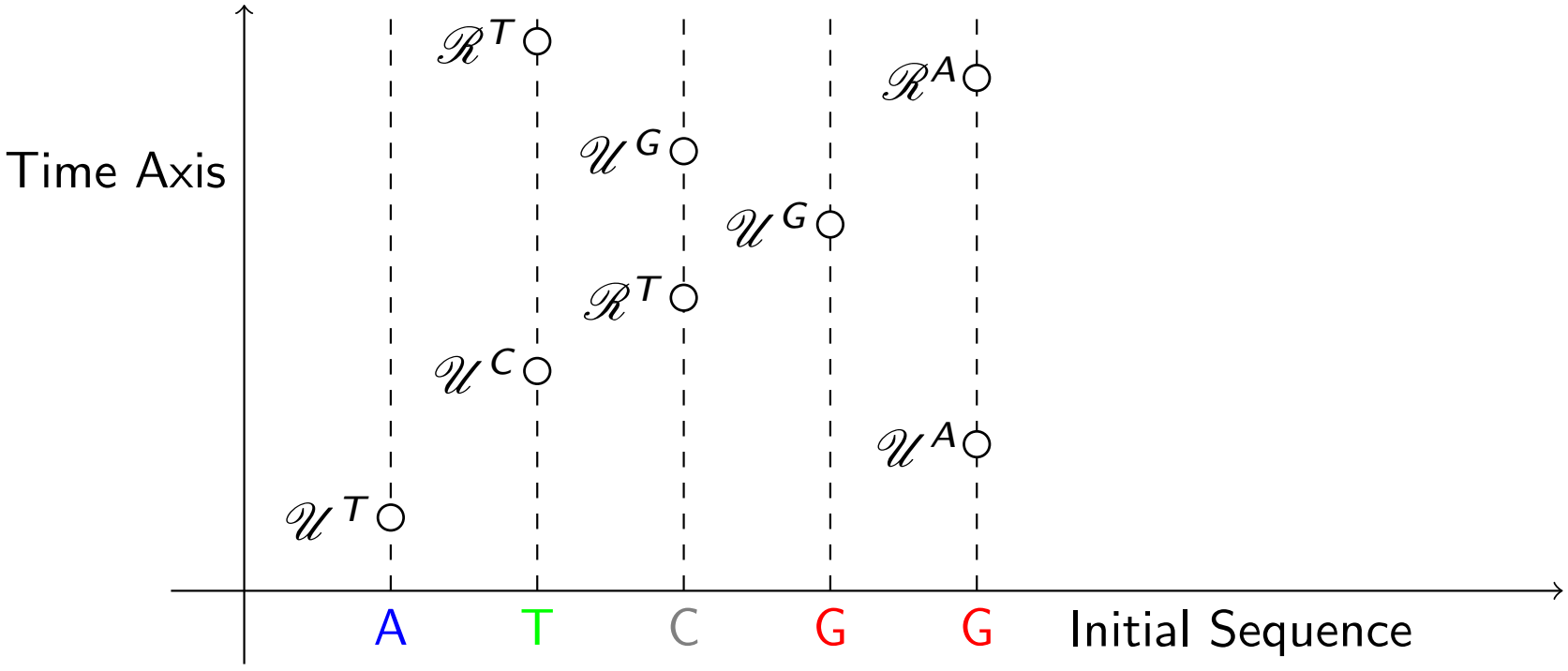
- **Keep** Jukes and Cantor model

	A	T	C	G
A	·	1	1	1
T	1	·	1	1
C	1	1	·	1
G	1	1	1	·

- **Superimpose** "double" substitution mechanism

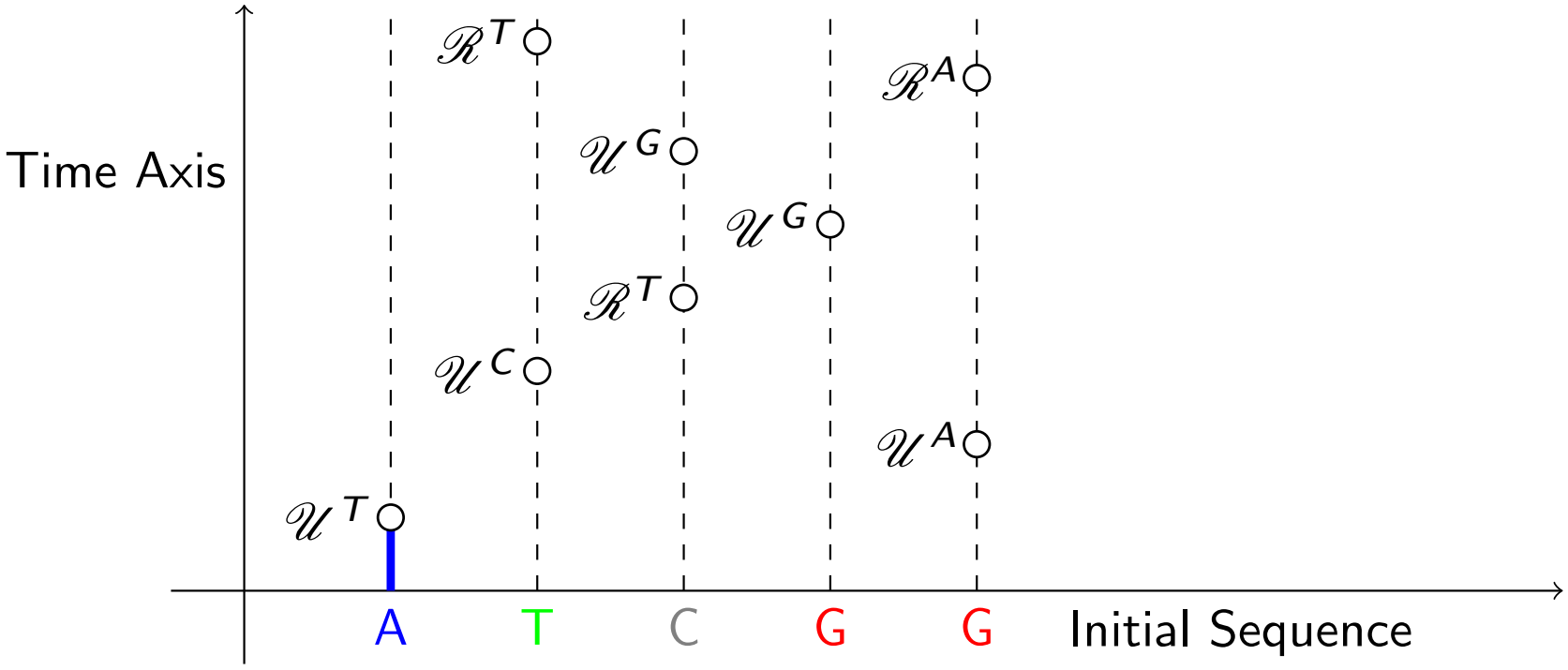


Construction with marked Poisson point processes



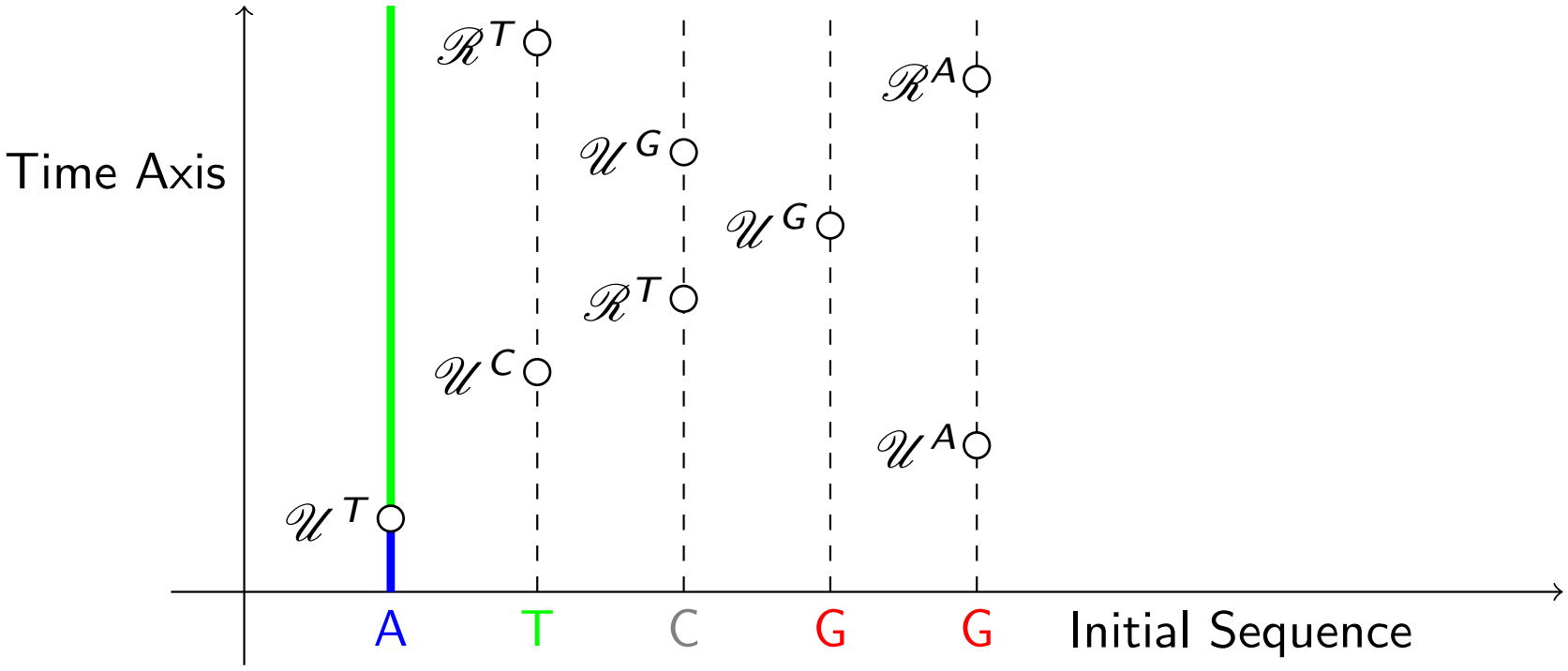
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



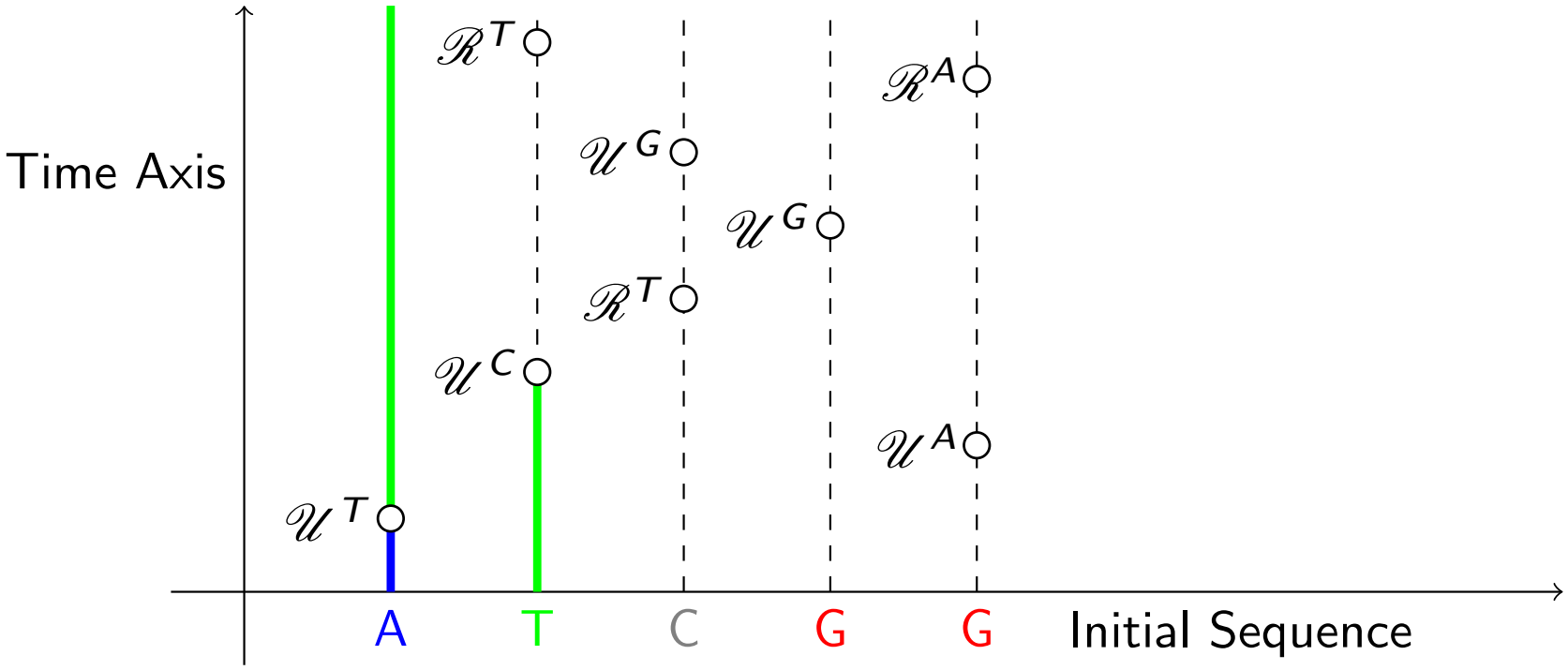
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



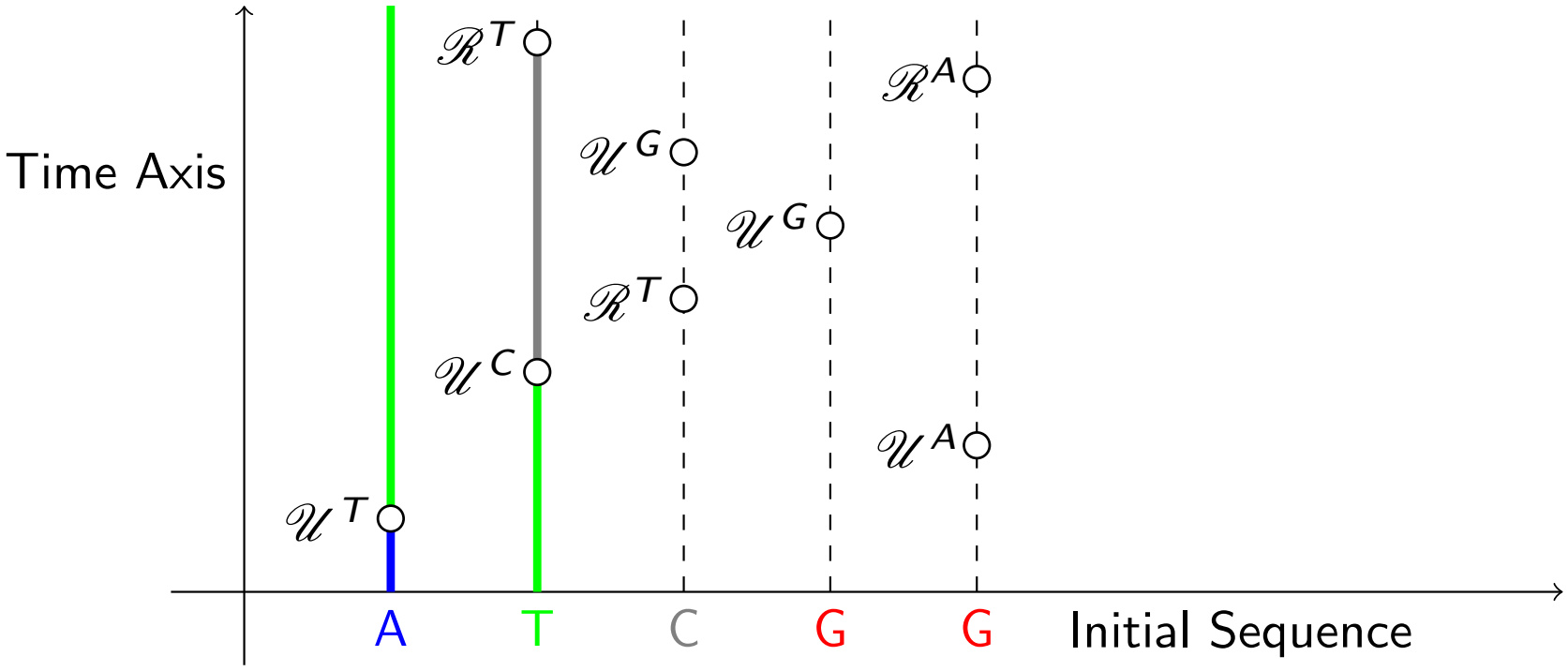
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



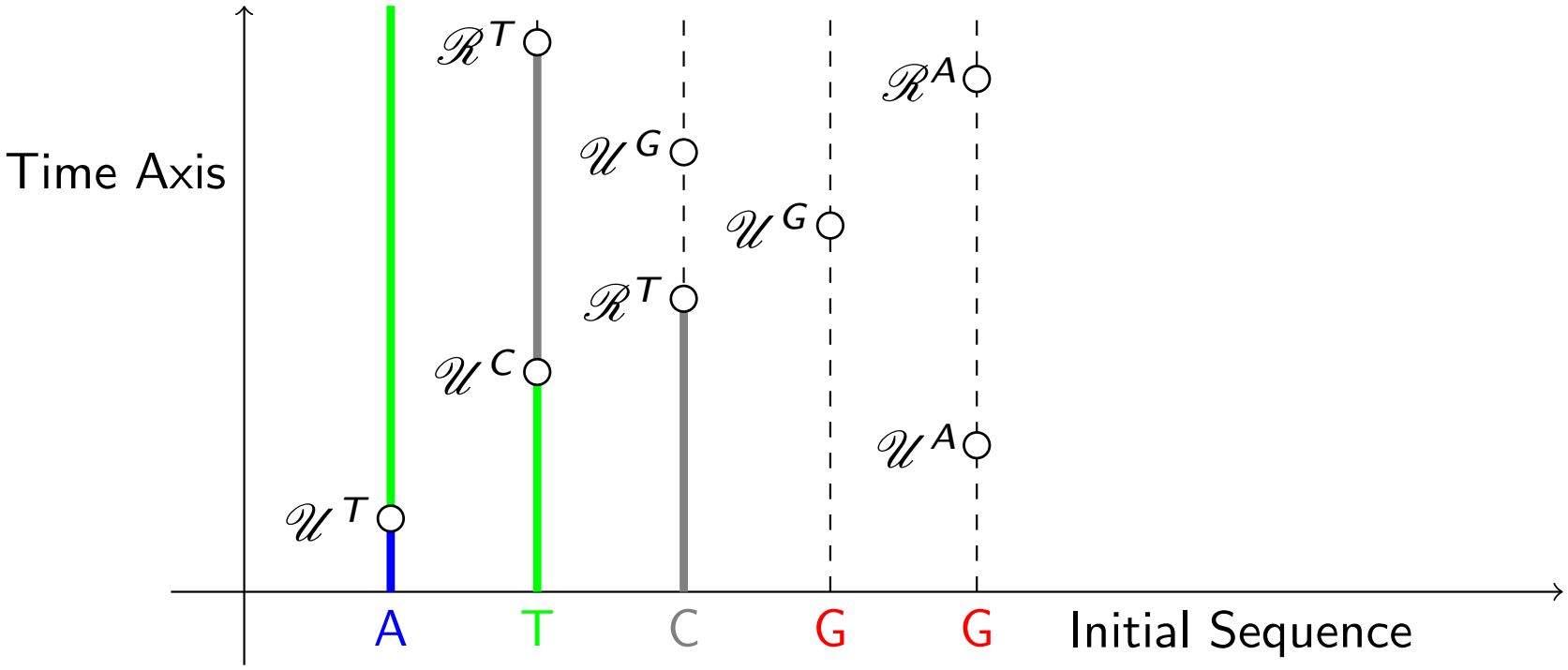
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



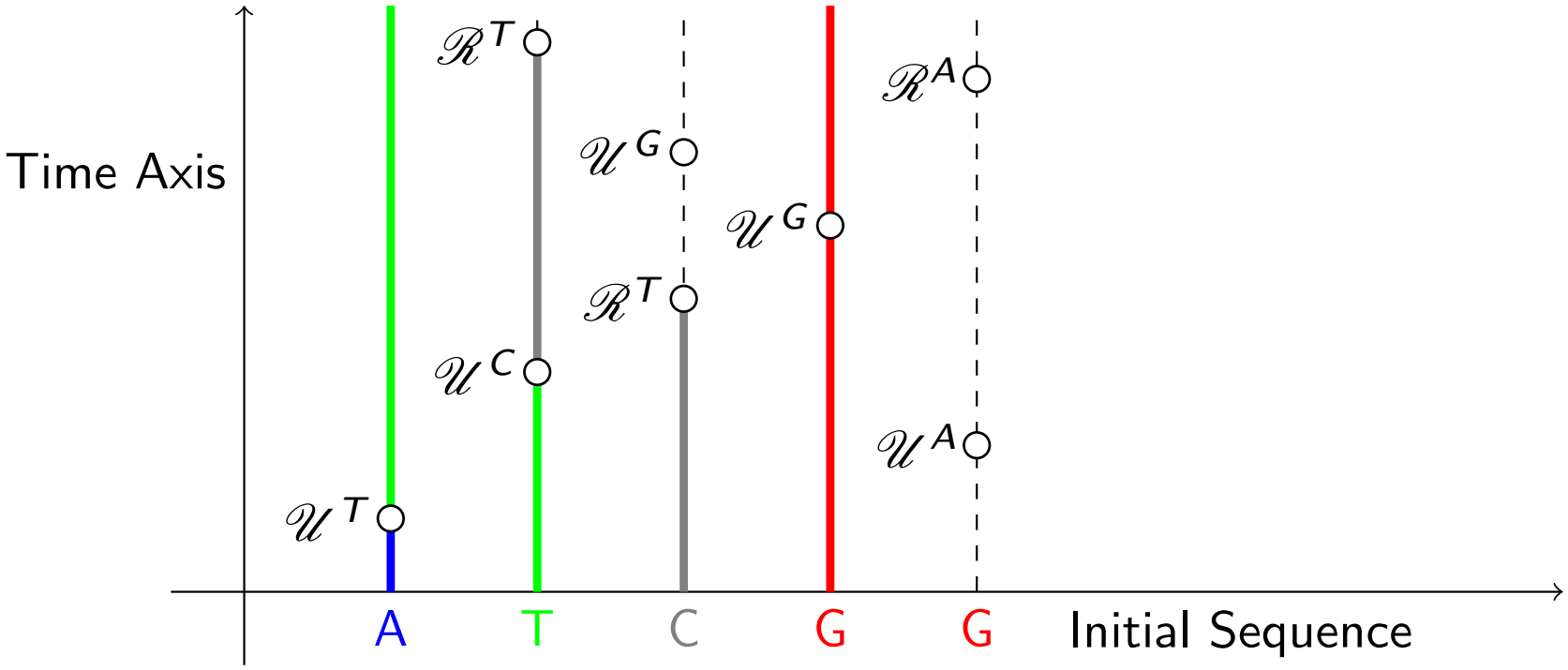
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



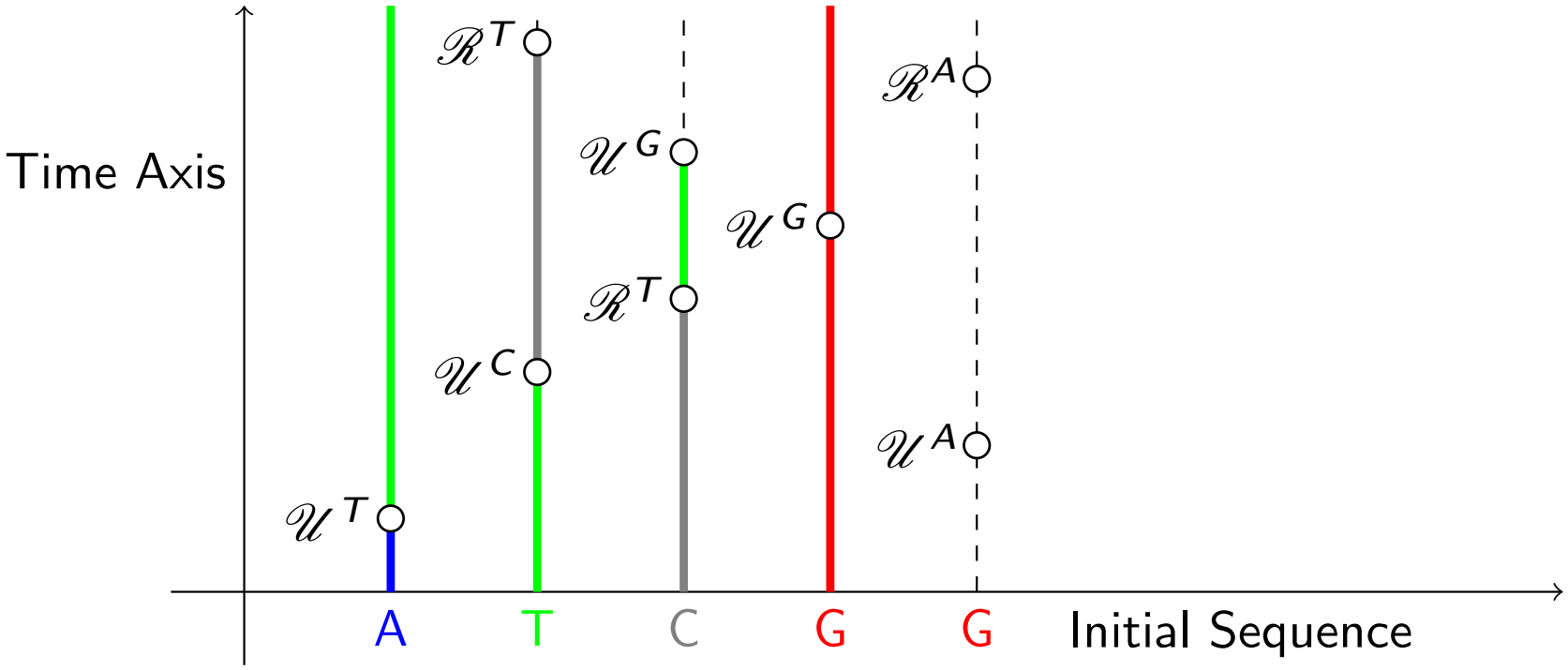
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



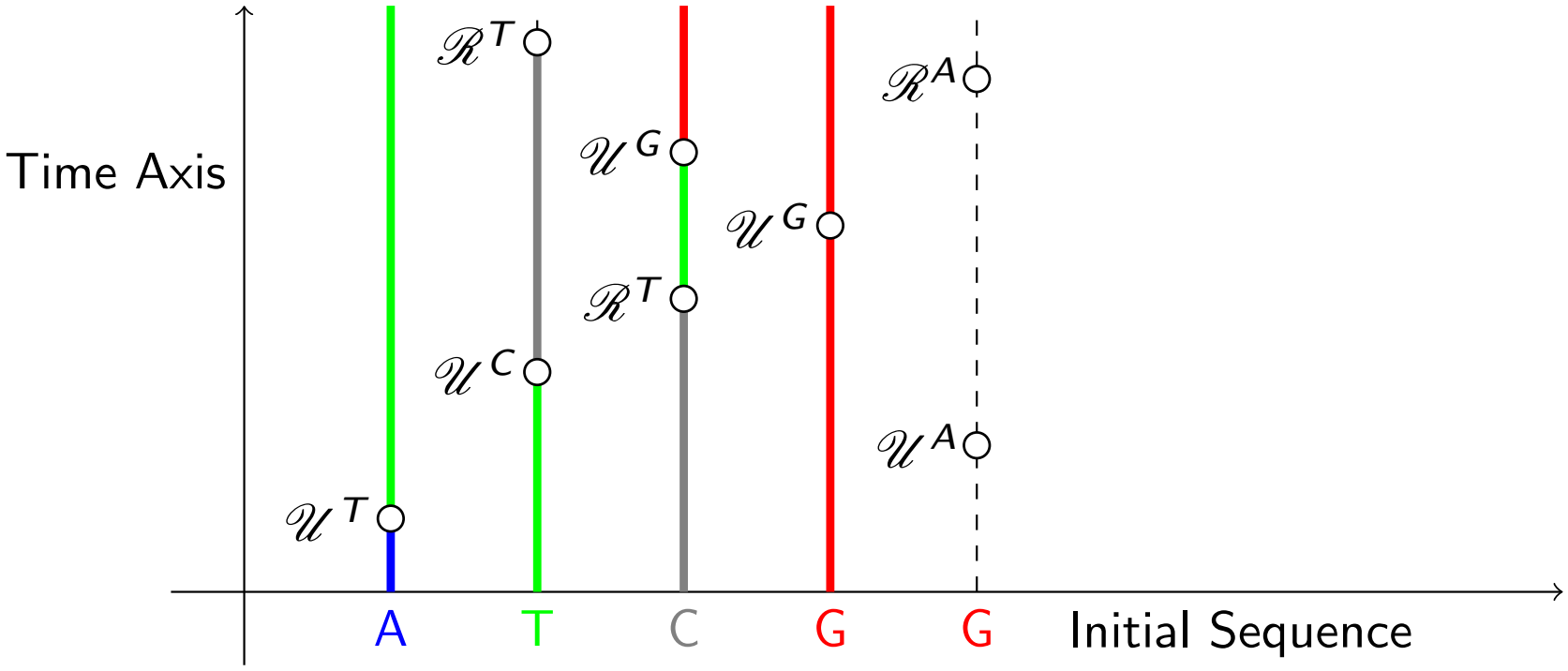
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



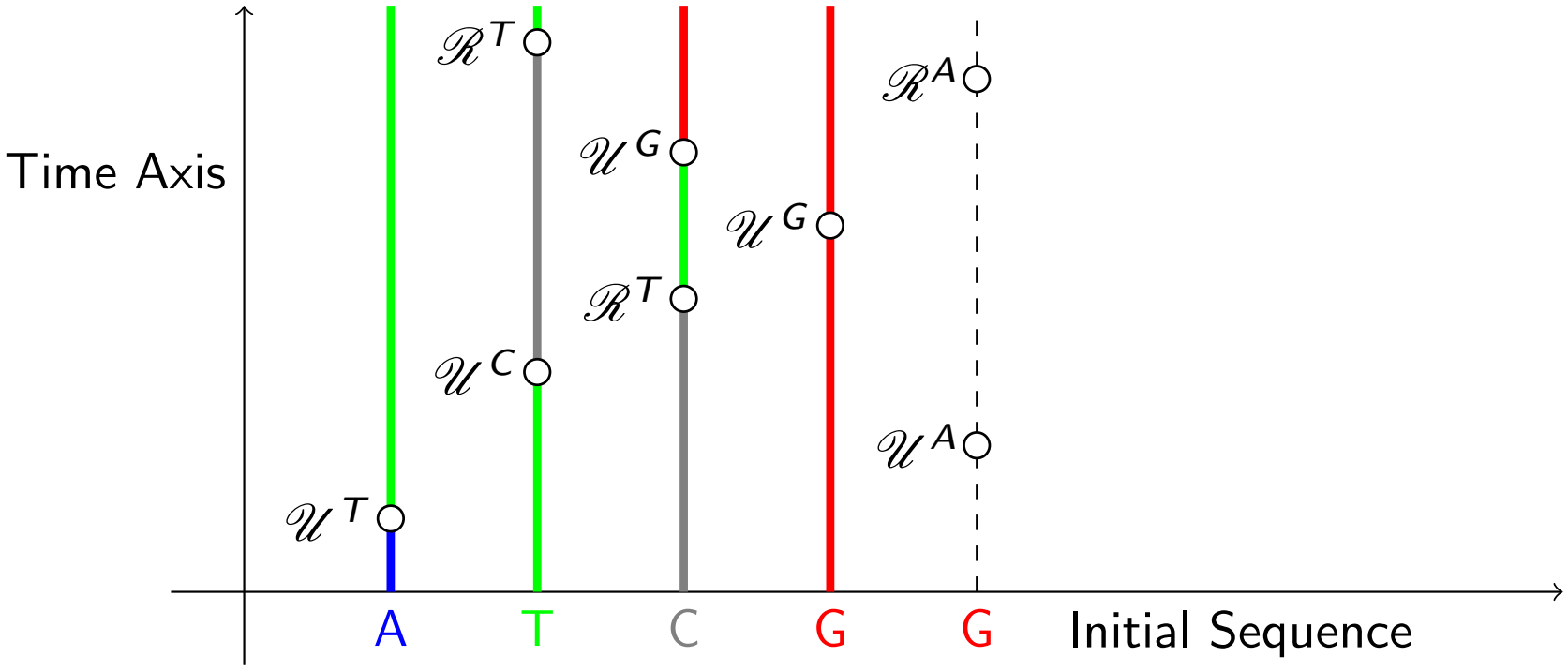
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



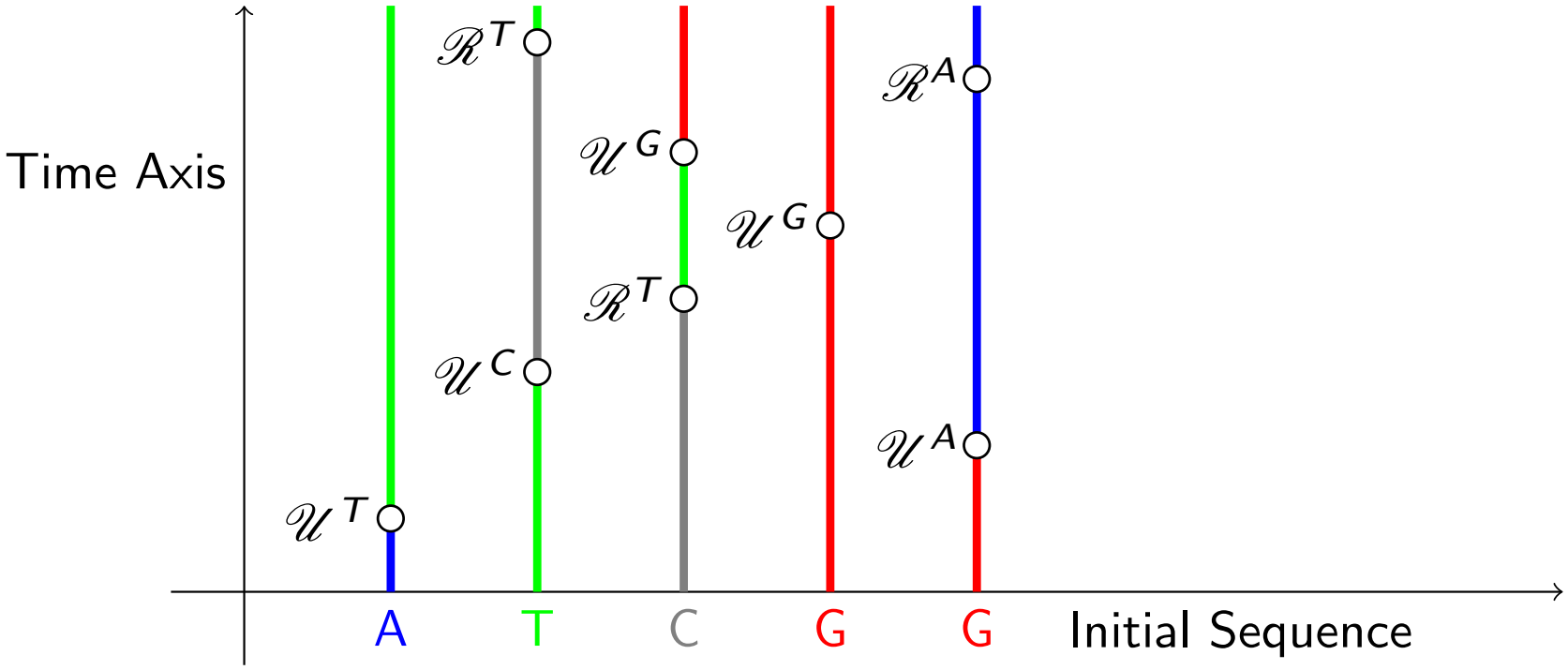
Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

Construction with marked Poisson point processes



Mark at x	**Rate**	**Action**: $\eta(x)$ moves to a ...
$\mathscr{U}^a$	1	... unconditionally.
$\mathscr{R}^a$	r	... if $\eta(x, x+1) =$CG and $a =$T or if $\eta(x-1, x) =$CG and $a =$A.

To know more about interacting particle systems

Bible: **Liggett, *Interacting particle systems*, Springer (1985)**

Durrett, *Ten lectures on interacting particle systems*, Springer (1993)

Properties

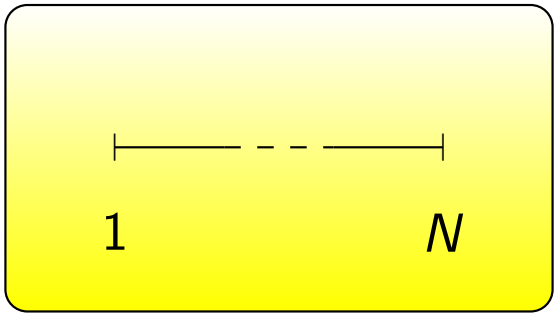
Bérard, Gouéré et Piau, *Mathematical Biosciences* (2008)

- **There exists a unique Markov process** on $\mathscr{A}^{\mathbb{Z}}$ with the transition rates defined before.

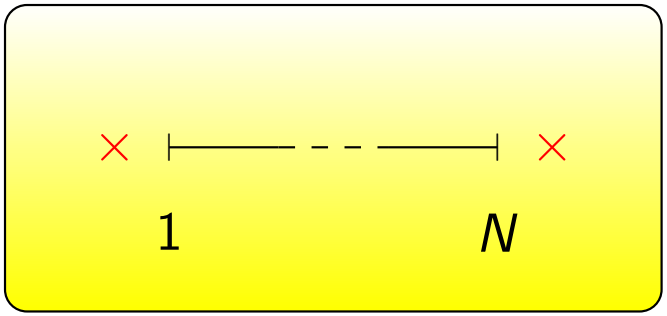
- The process is **ergodic**, its unique invariant probability measure π on $\mathscr{A}^{\mathbb{Z}}$ is **translation invariant** and **ergodic** with respect to the translations on $\mathbb{Z}$.

- Starting from equilibrium, any collections $(\eta_x)_{x\in I}$ and $(\eta_y)_{y\in J}$ are **indépendent** as soon as $\mathrm{dist}(I, J) \geqslant 3$.

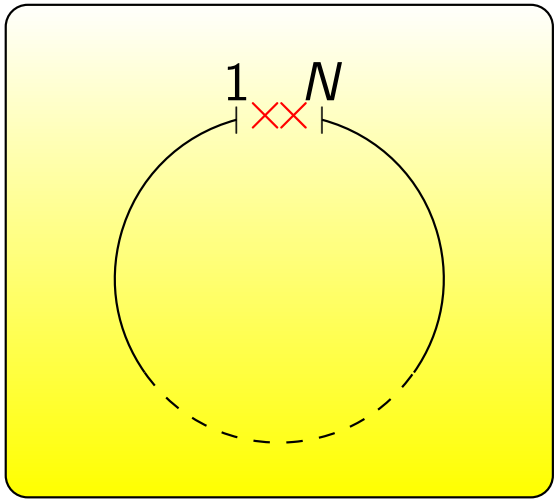
Simulate the evolution of a finite DNA sequence



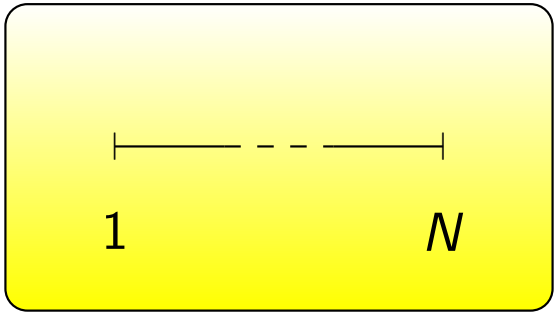
Add two artificial sites



Transform the linear sequence into a circular one

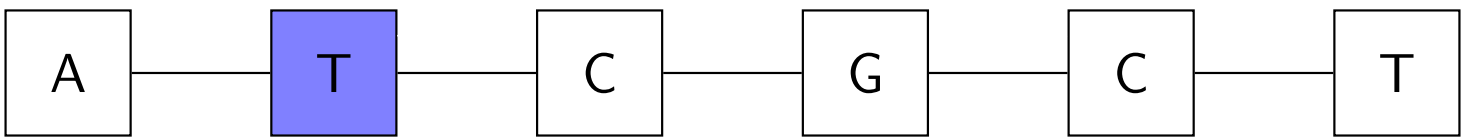


Run the evolution process with marked Ppp's and remove the two artificial sites

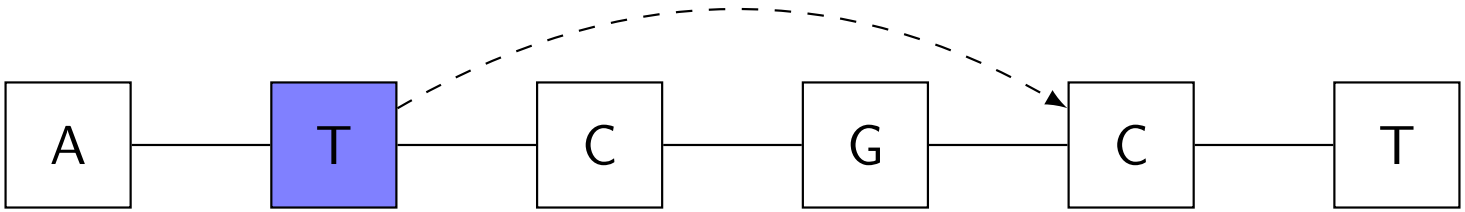


Example of translocation

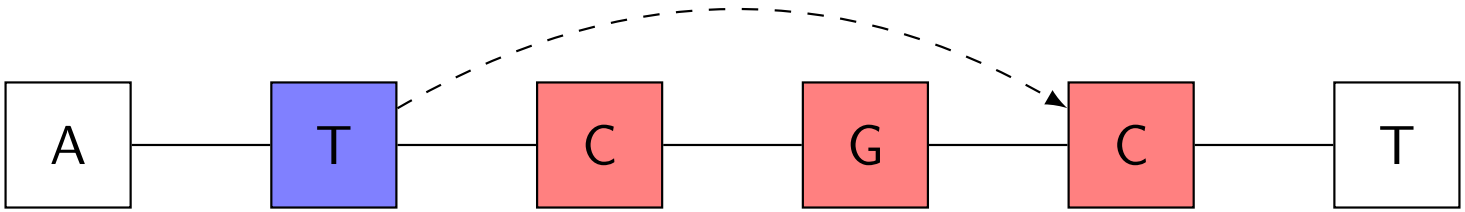
Example of translocation



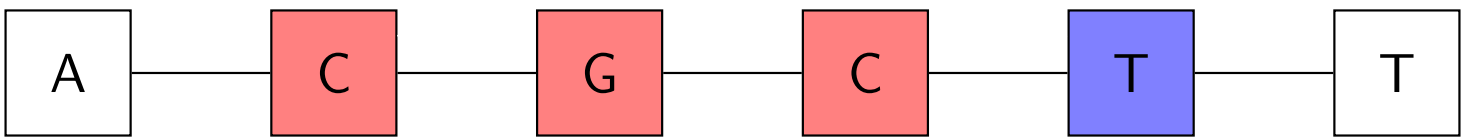
Example of translocation



Example of translocation



Example of translocation



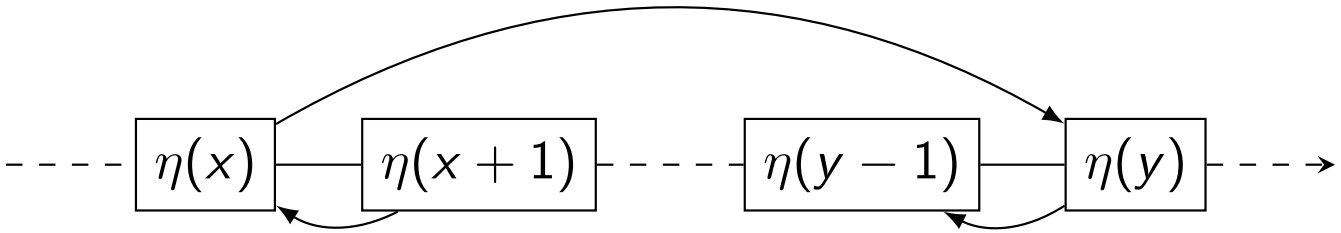
Definition with Markov generator

Translocation process

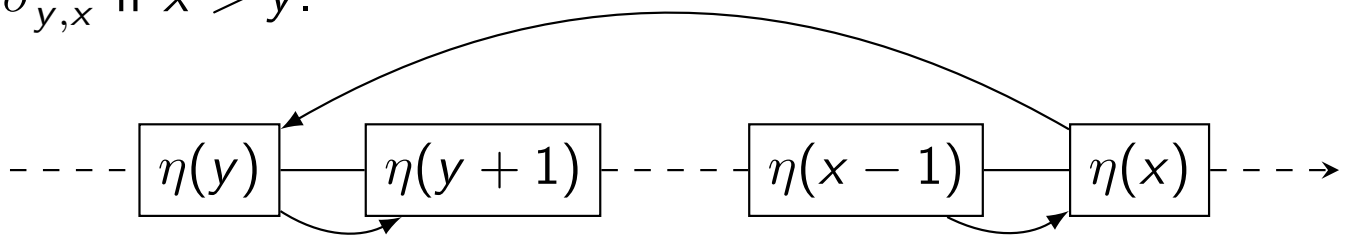
$$\mathscr{L}_2 f(\eta) = \sum_{x,y\in\mathbb{Z}} p(x,y)[f(\eta\circ\sigma_{x,y}) - f(\eta)], \tag{1}$$

with $\sigma_{x,y}$ defined for any $x < y$ by

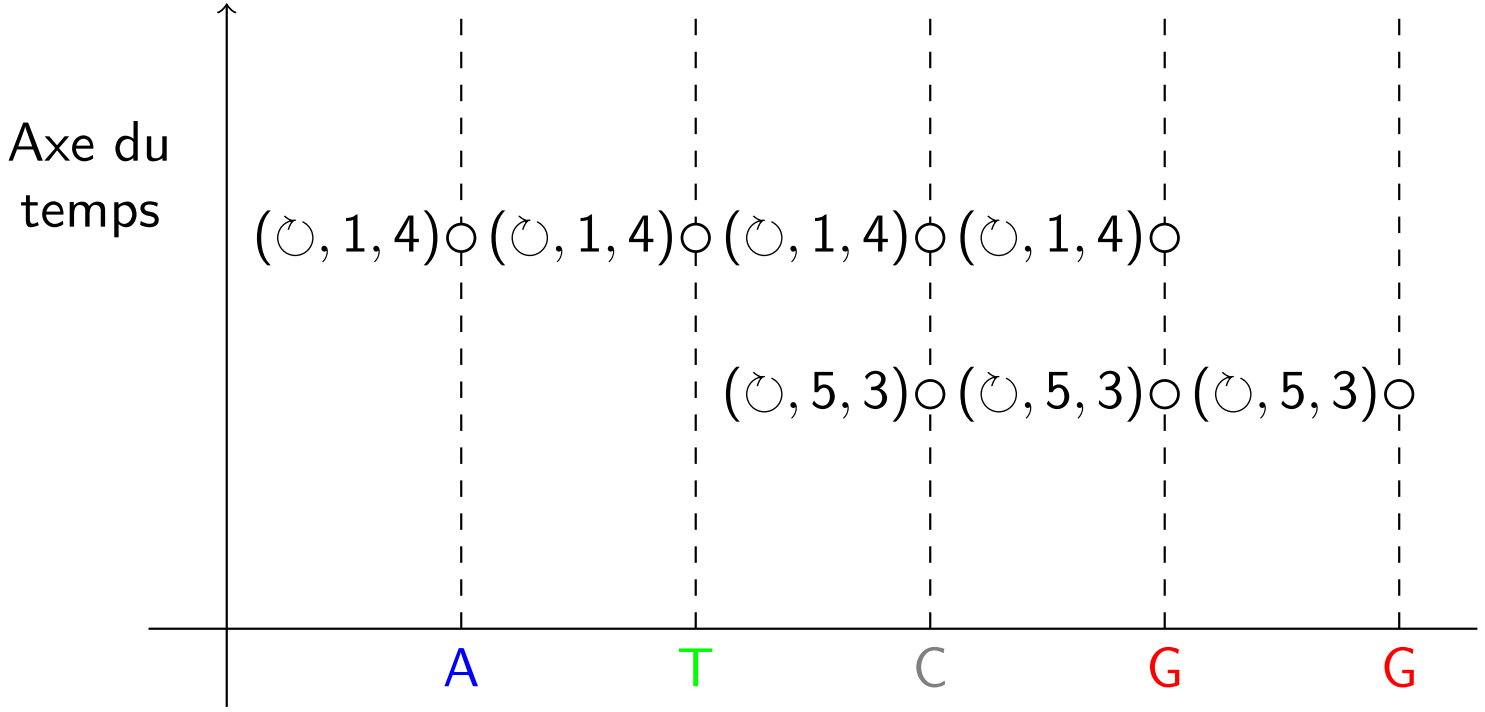
$$\sigma_{x,y}(z) = \begin{cases} z & \text{if} \quad z \notin \{x, x+1, \ldots, y\}, \\ y & \text{if} \quad z = x, \\ z-1 & \text{if} \quad x < z \leqslant y, \end{cases}$$



and by $\sigma_{y,x}^{-1}$ if $x > y$.

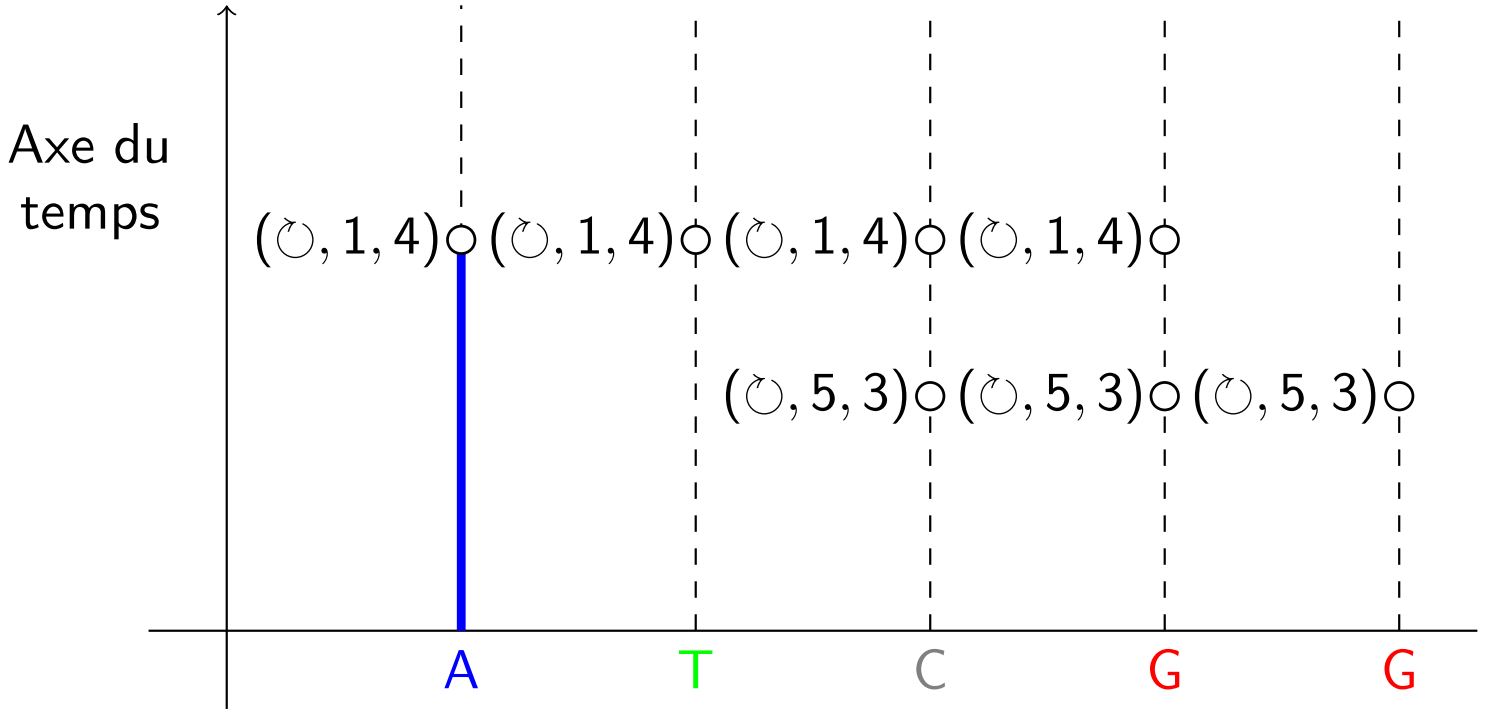


Construction with marked Poisson point processes



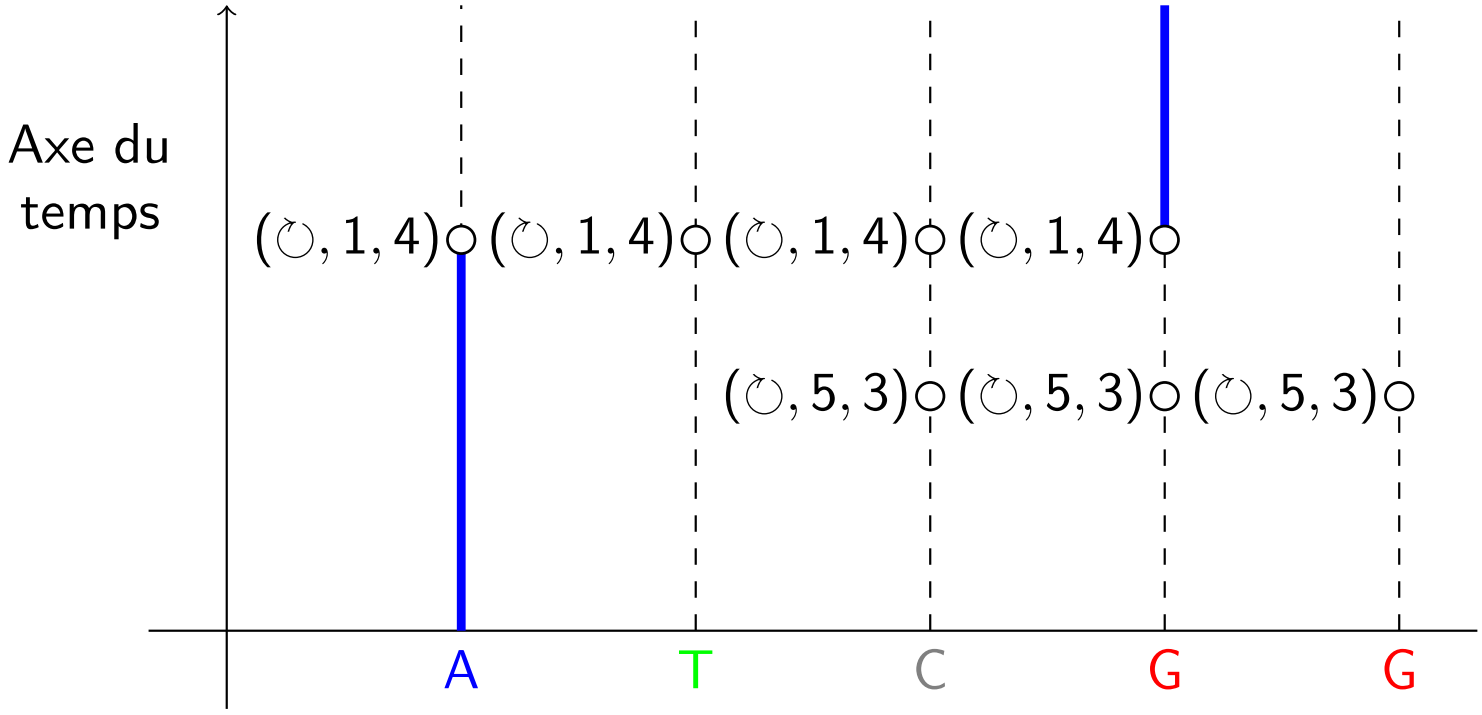
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



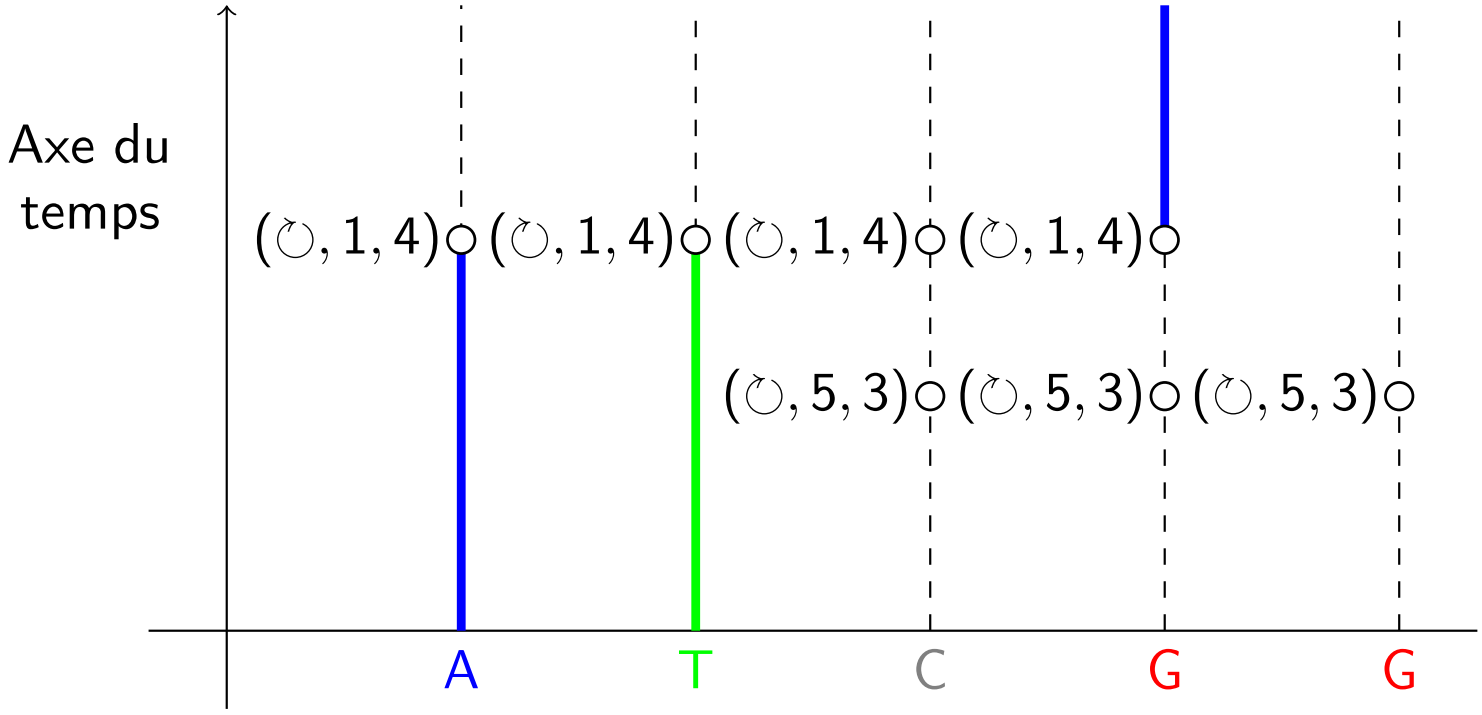
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



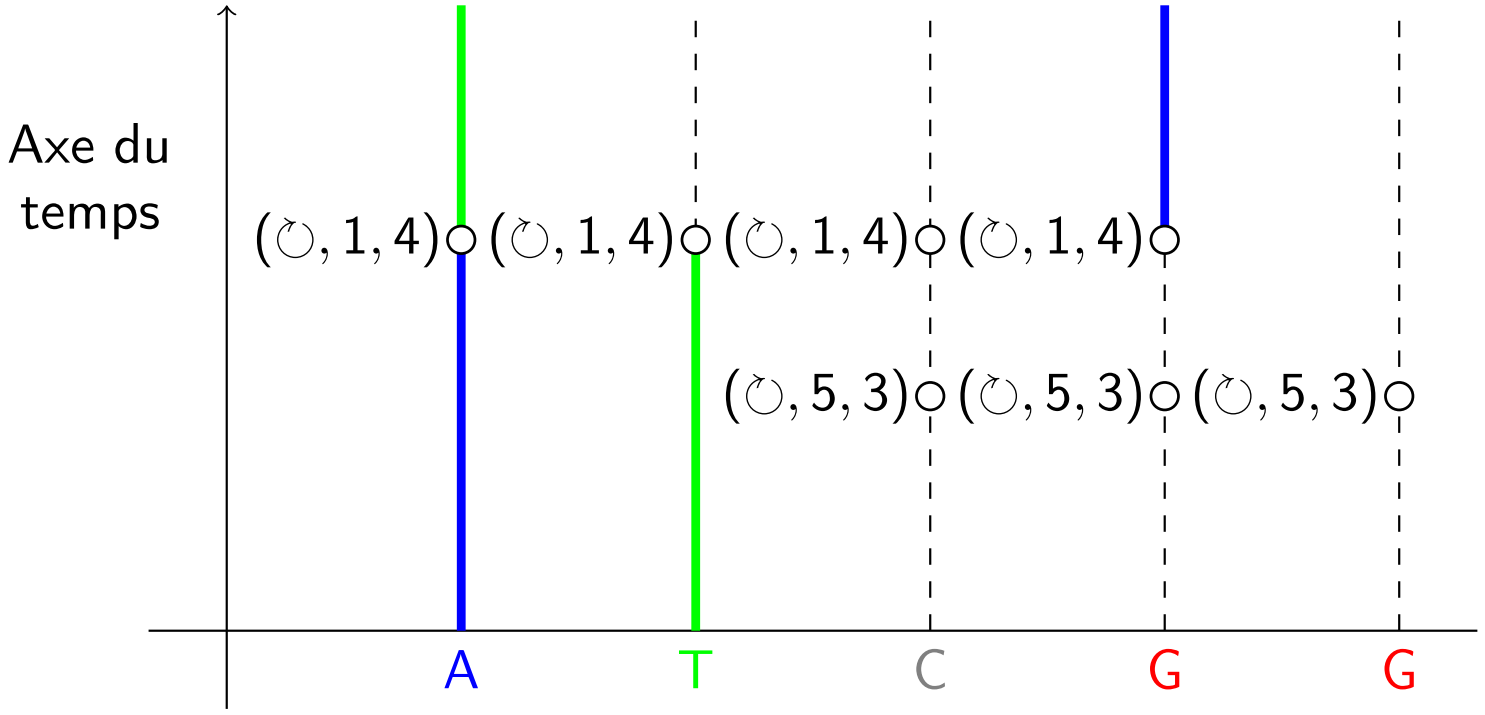
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



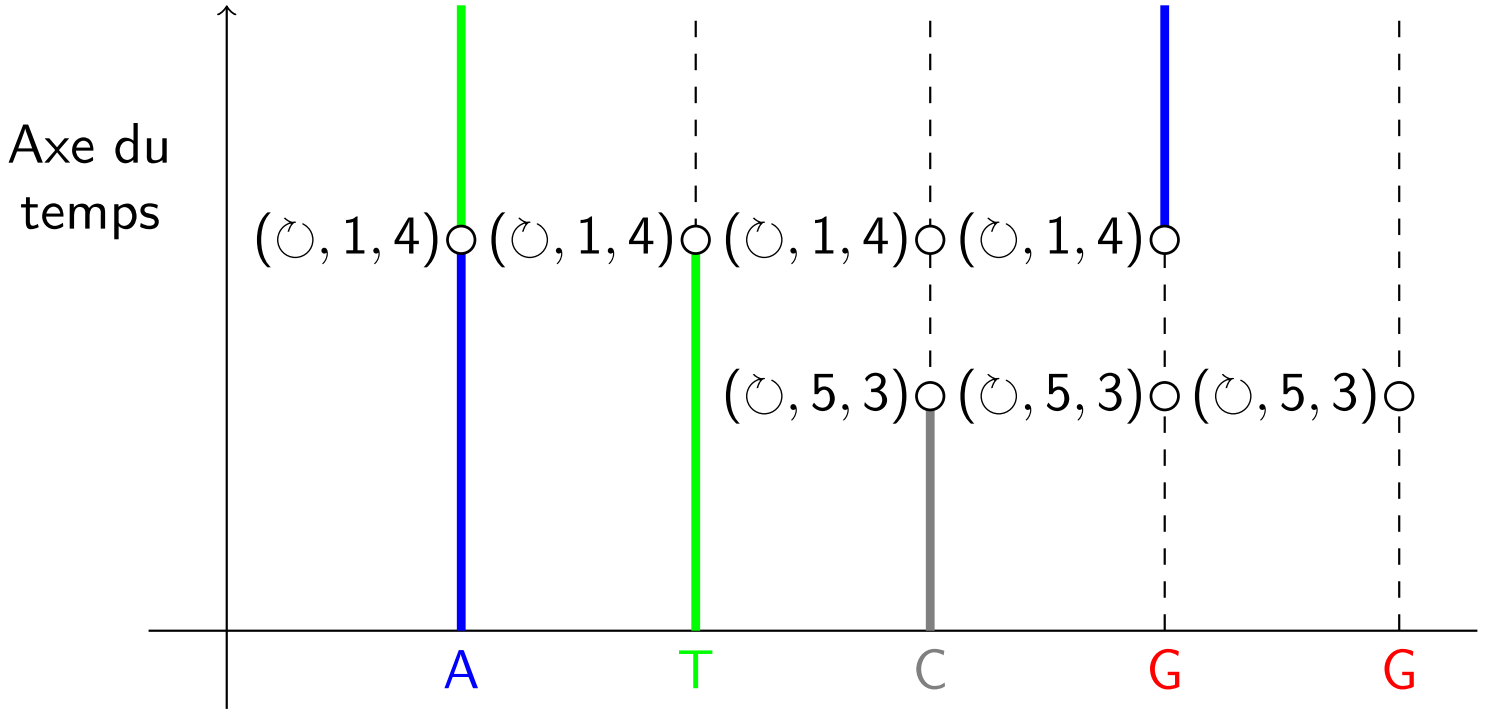
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



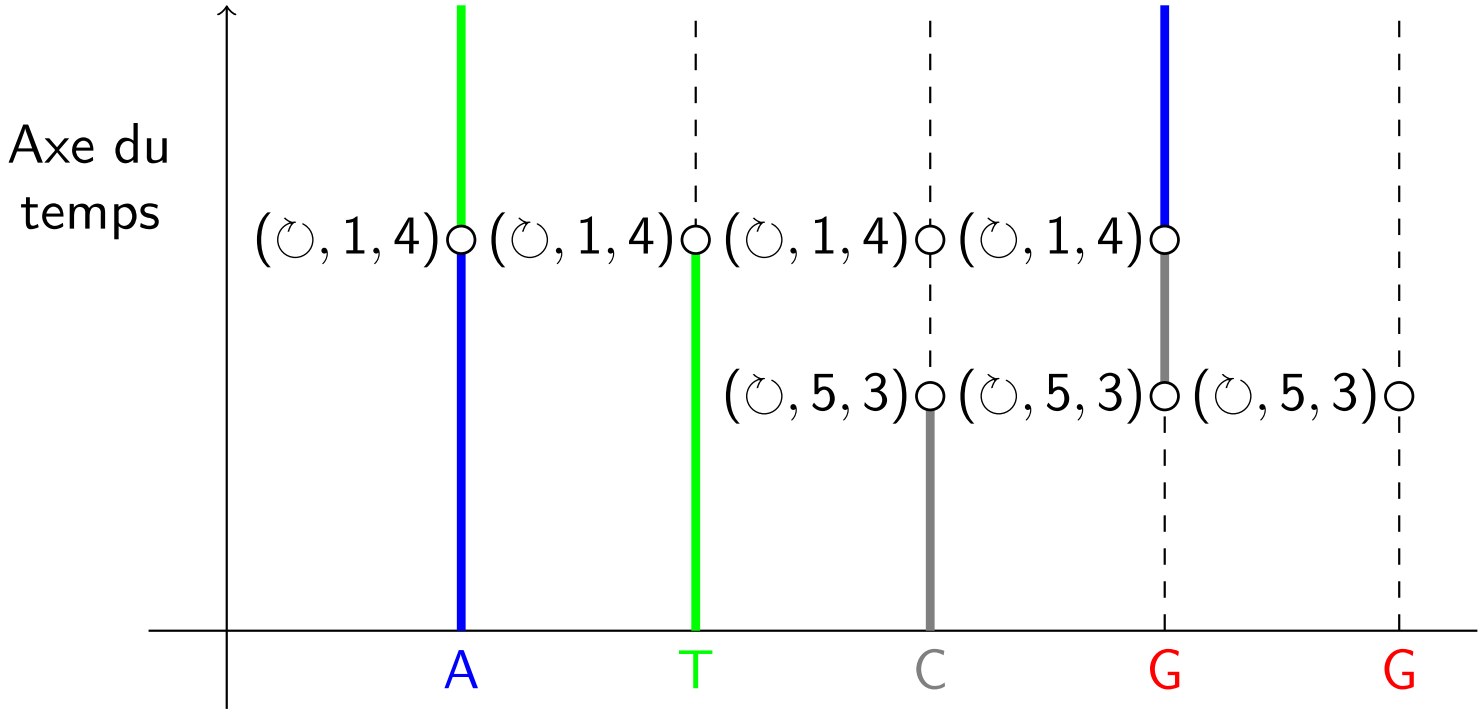
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



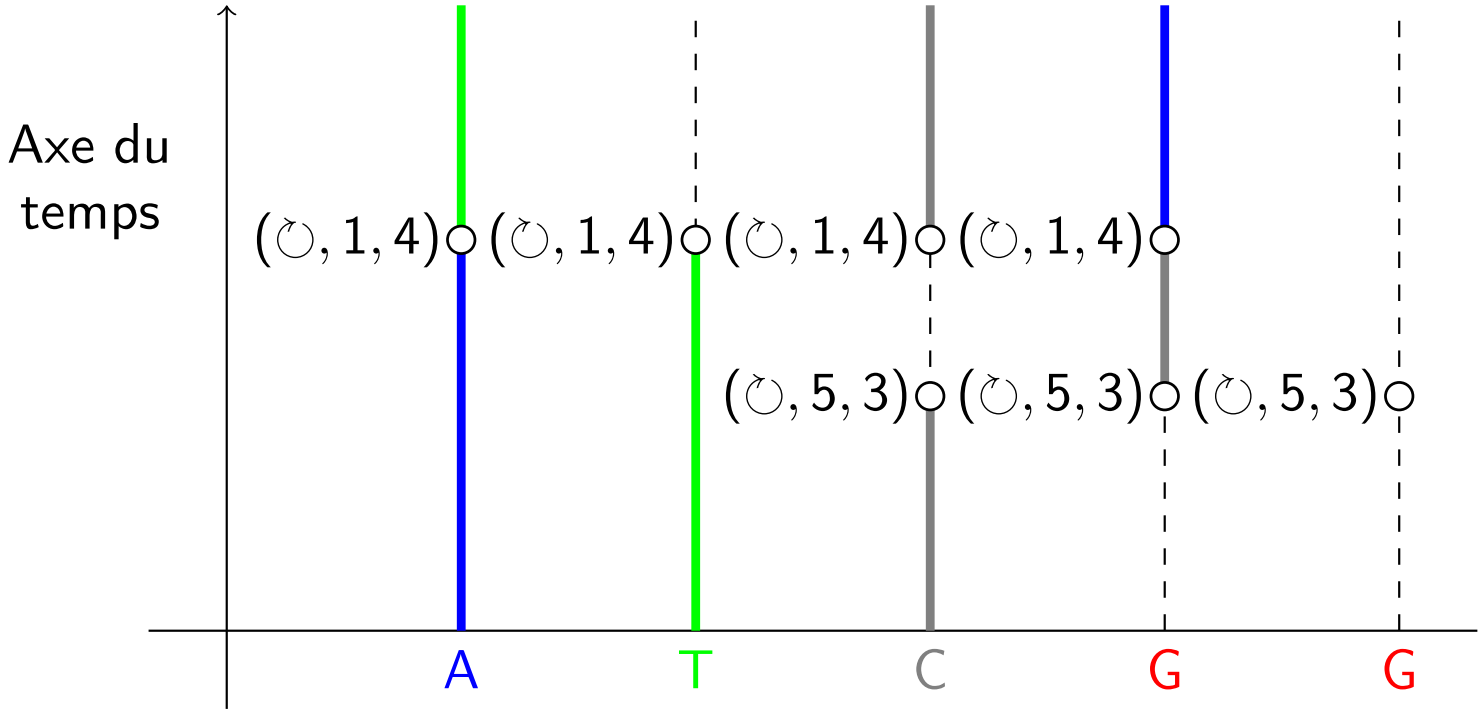
- Mark $(\circlearrowright, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



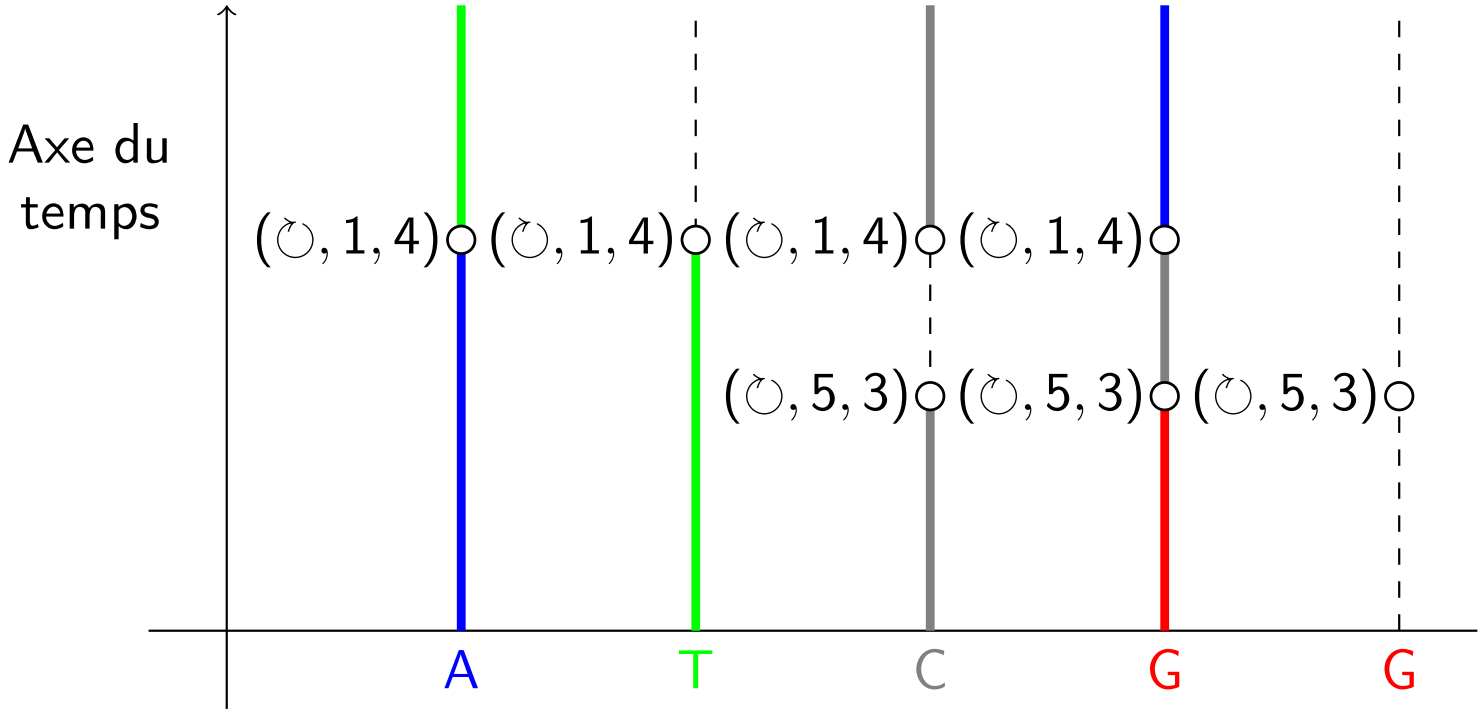
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



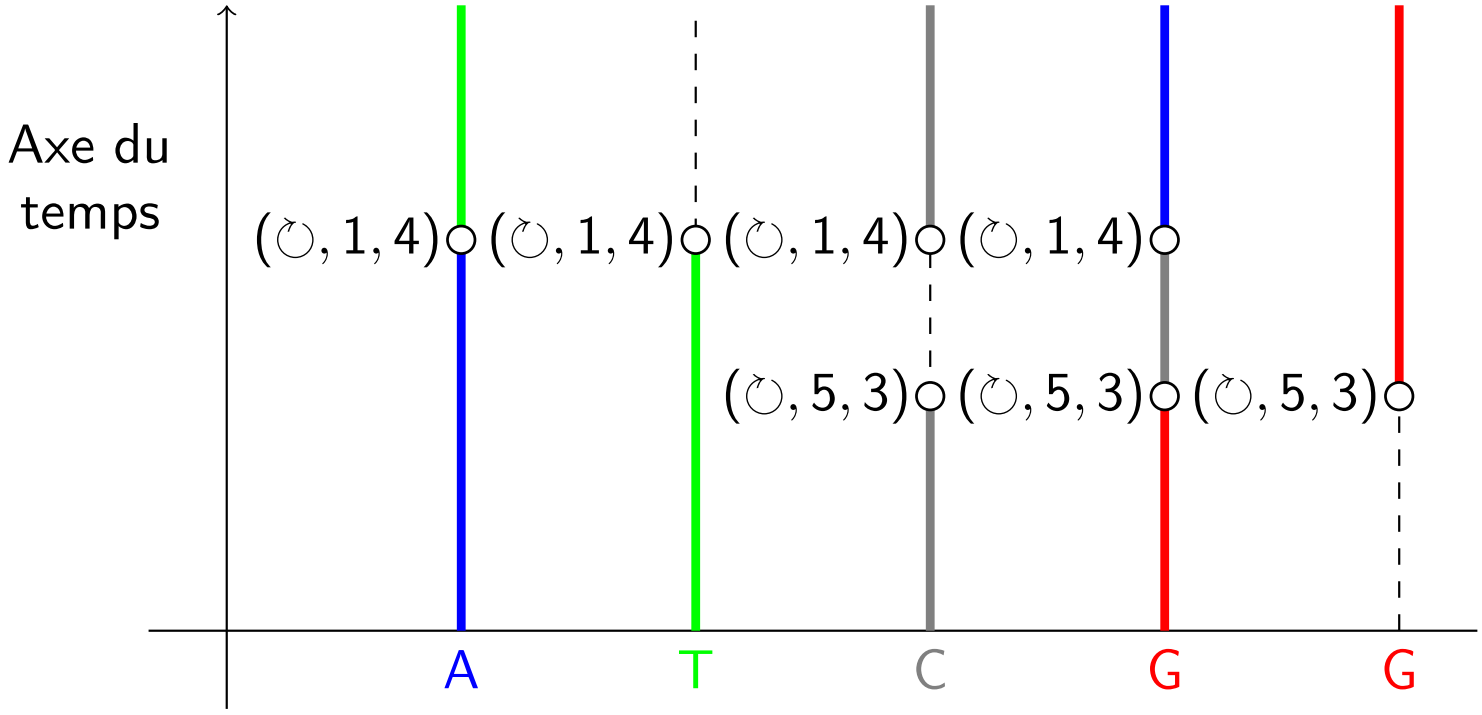
- Mark $(\circlearrowright, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



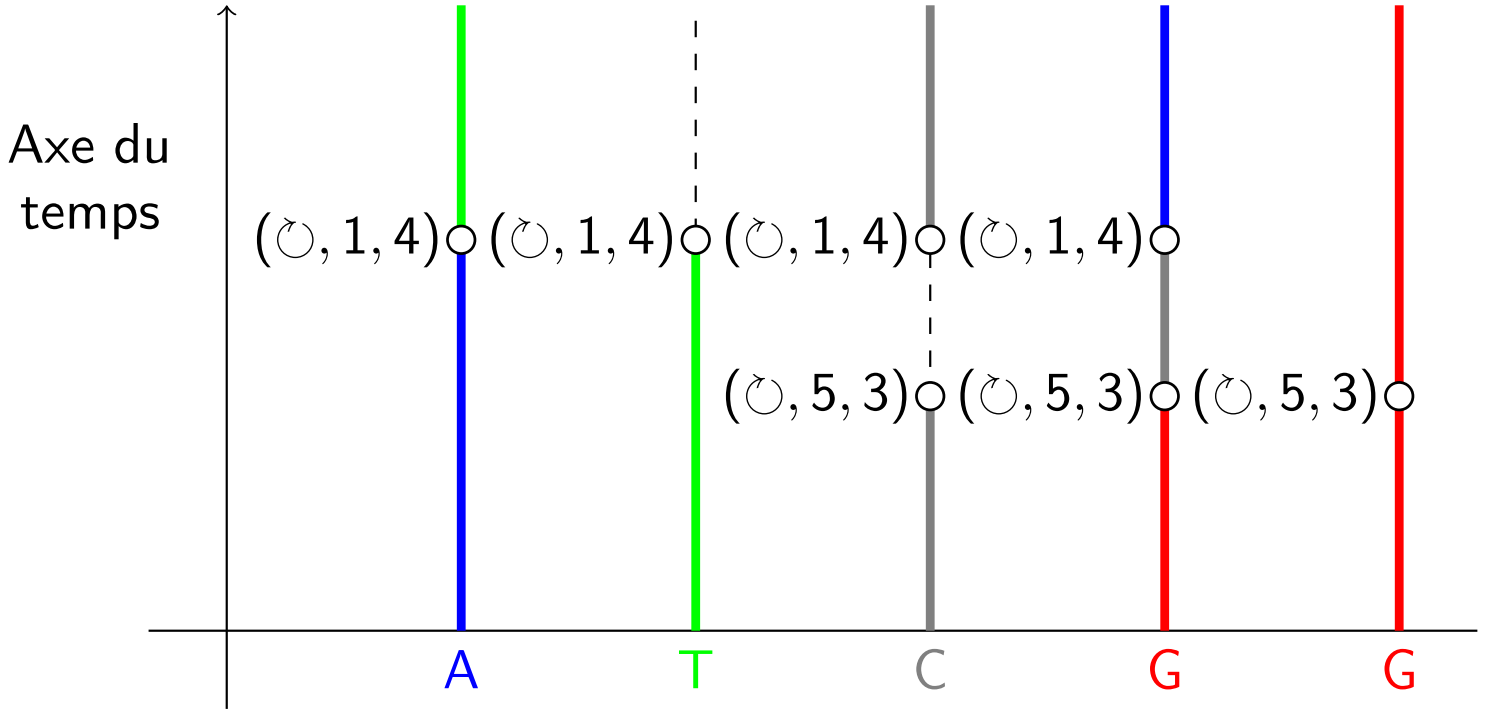
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



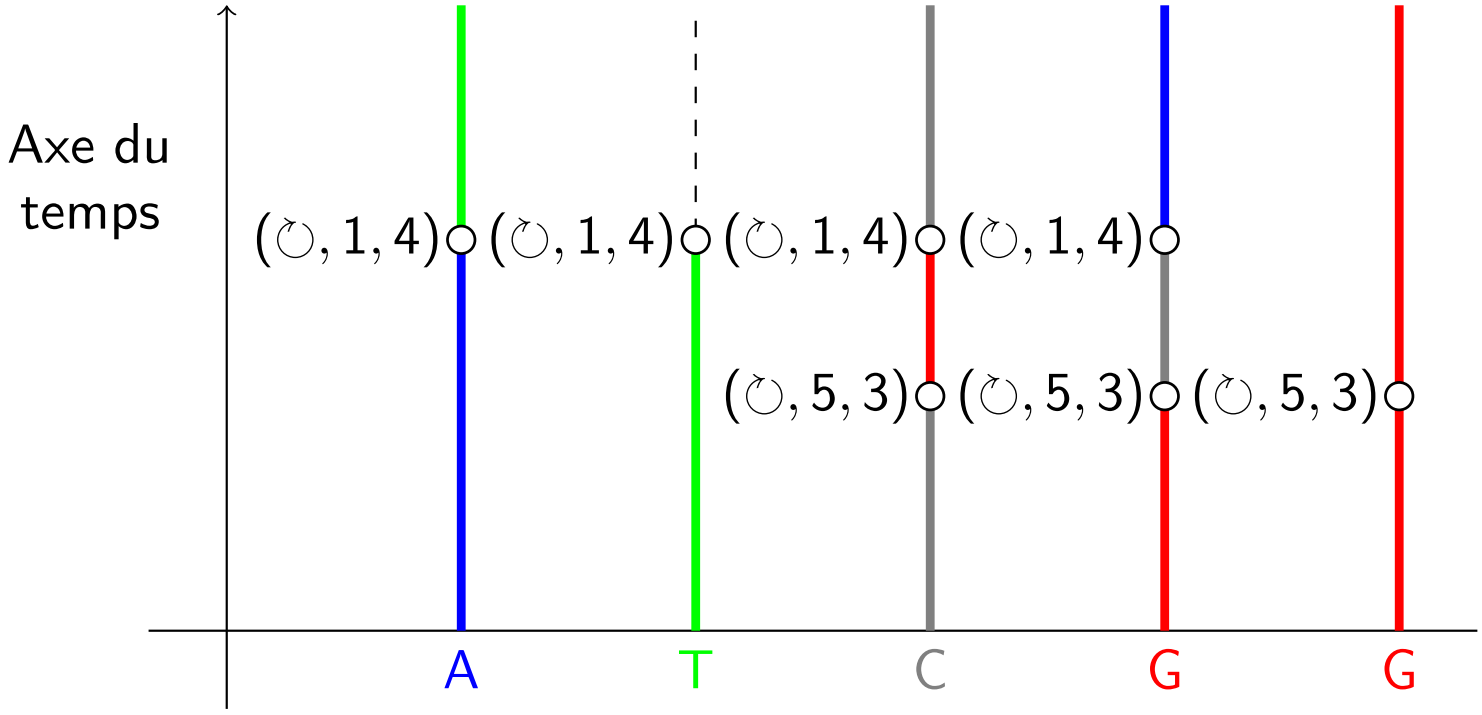
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



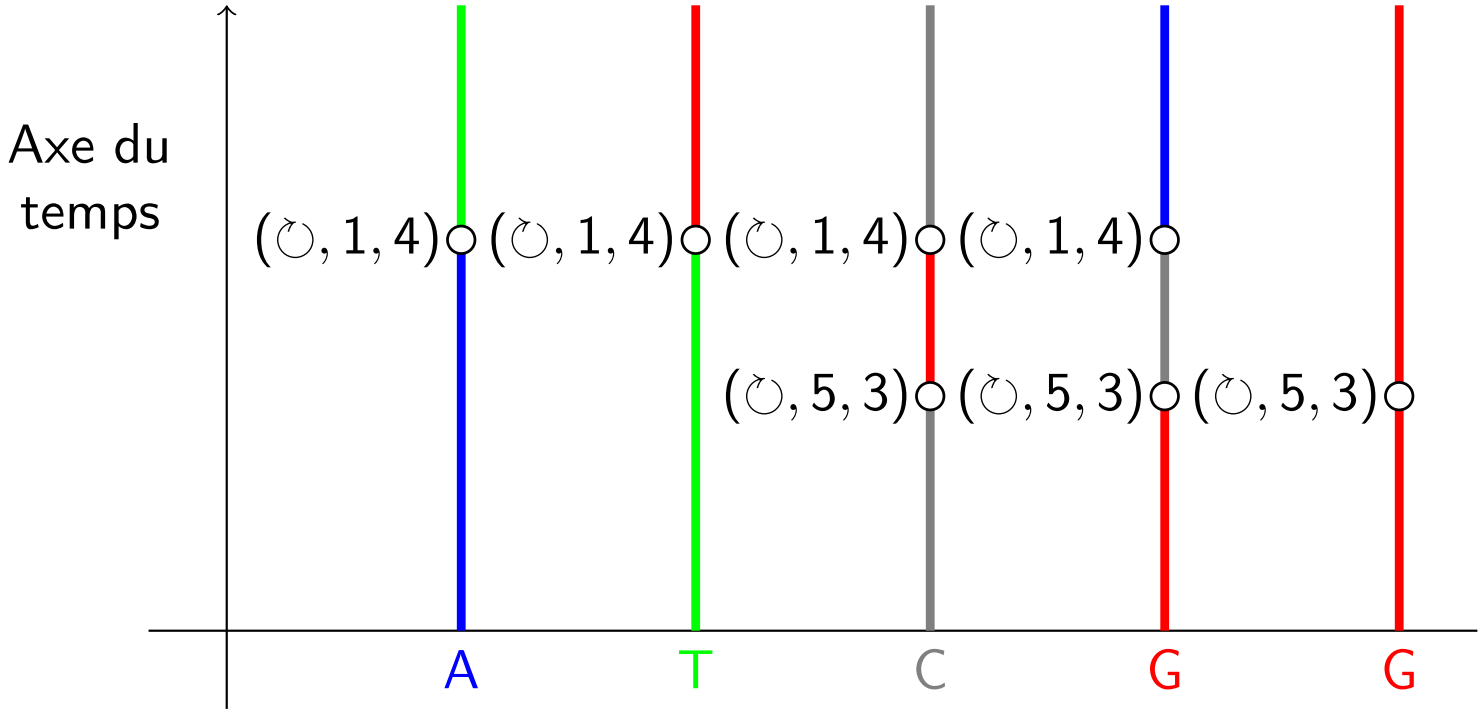
- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

Construction with marked Poisson point processes



- Mark $(\circlearrowleft, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

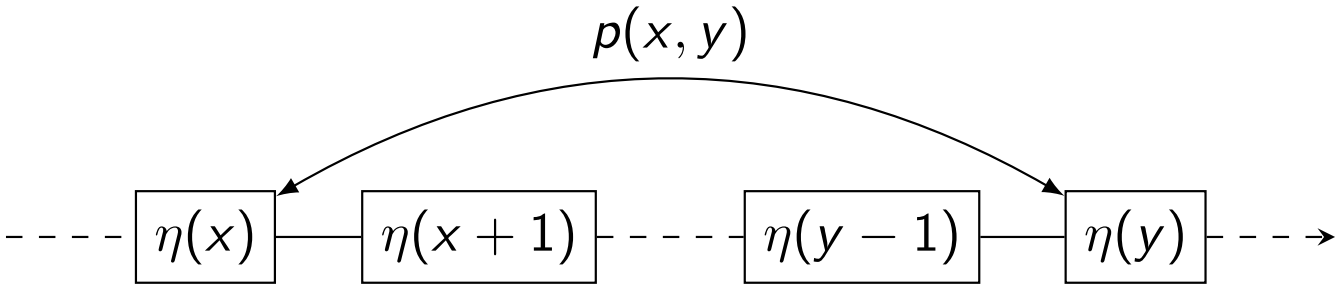
Construction with marked Poisson point processes



- Mark $(\circlearrowright, x, y)$ distributed at rate $\rho p(x, y)$. If $x < y$, the contents of sites x, $x + 1$, ..., y are right circularly permuted. If $x < y$, the contents of sites y, $y + 1$, ..., x are left circularly permuted.

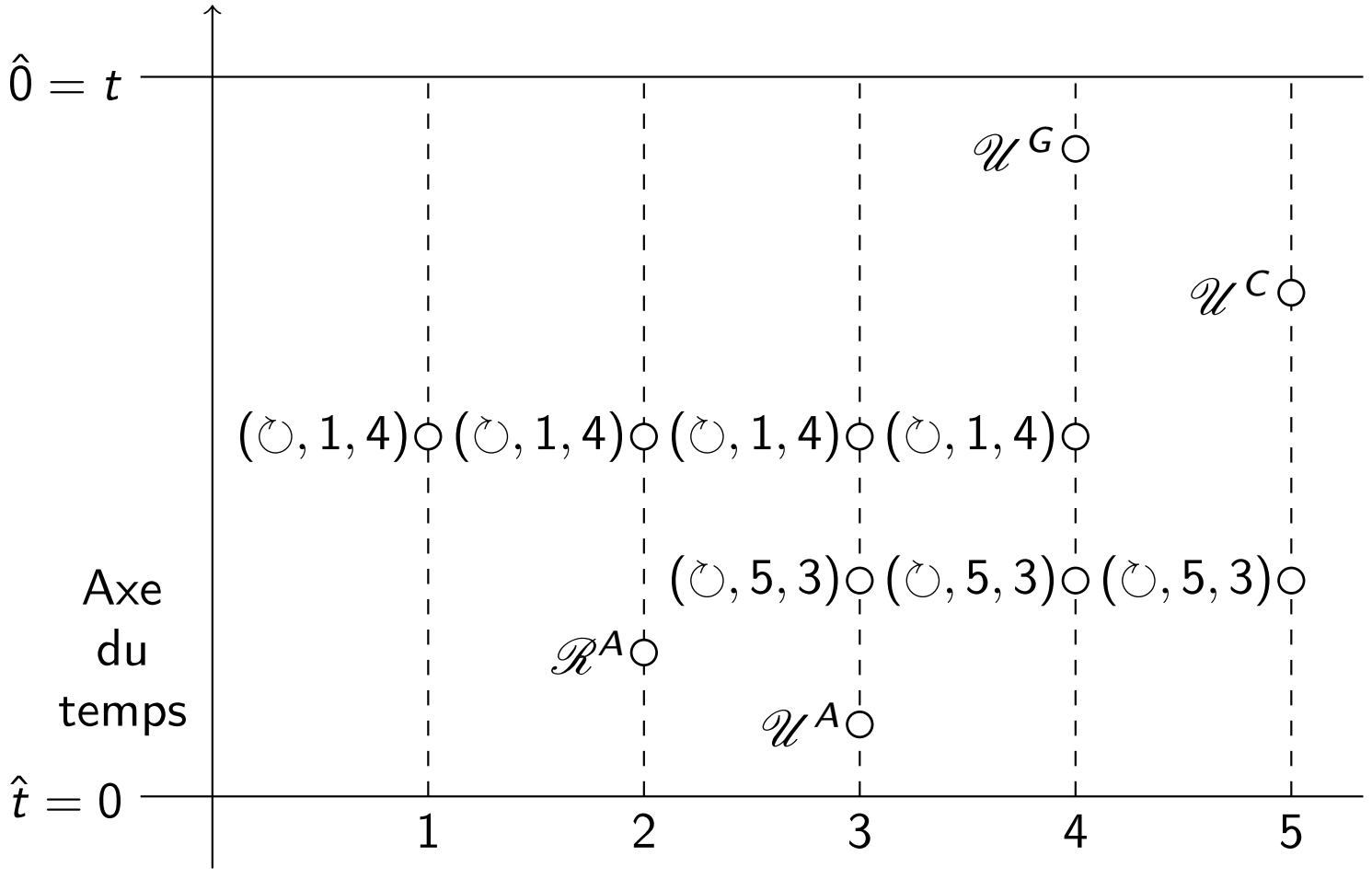
Spin + stirring

Ferrari, *Annals of Probability* (1990)

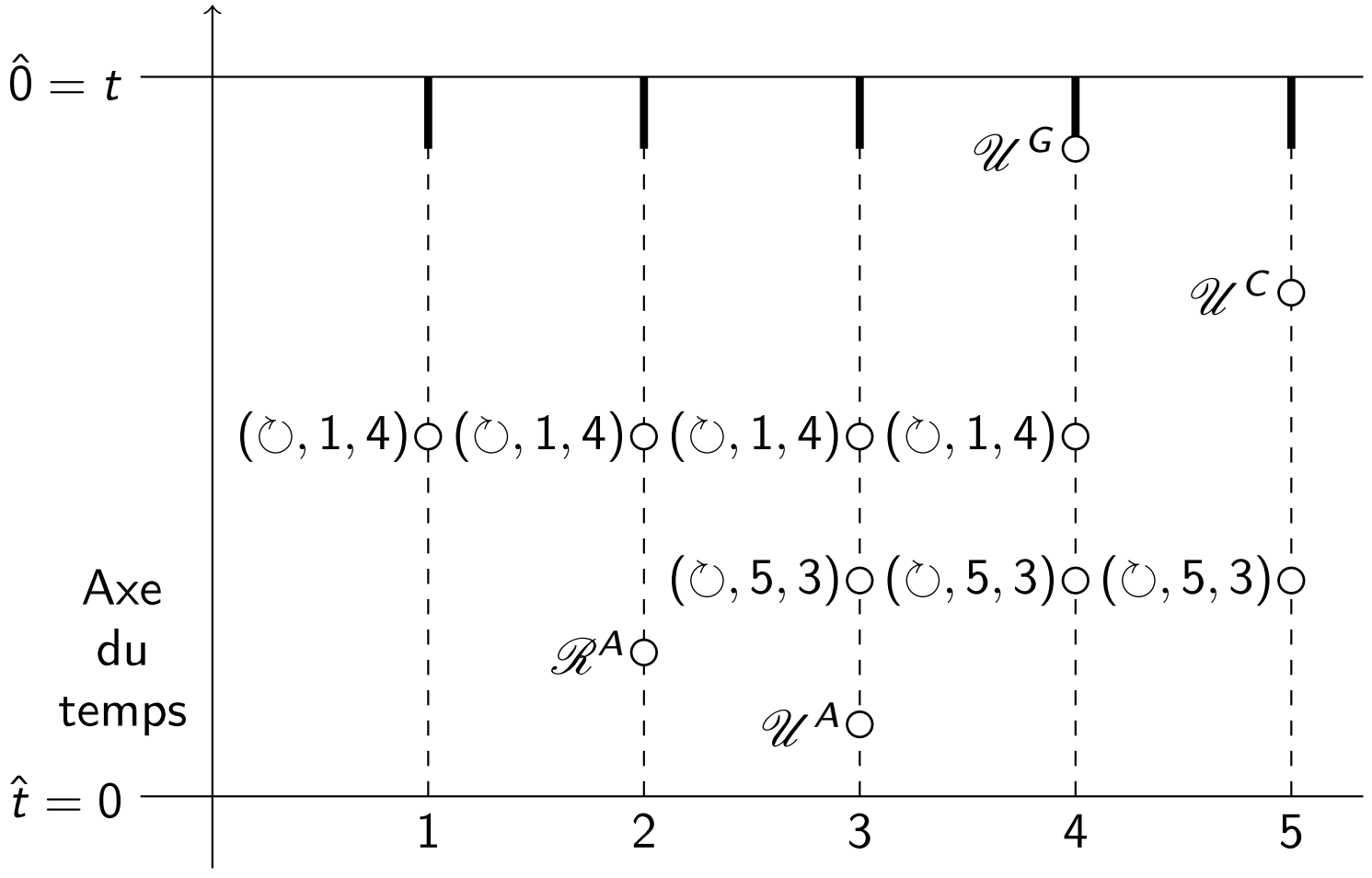


To prove ergodicity, Ferrari introduces the construction of a **dual process**

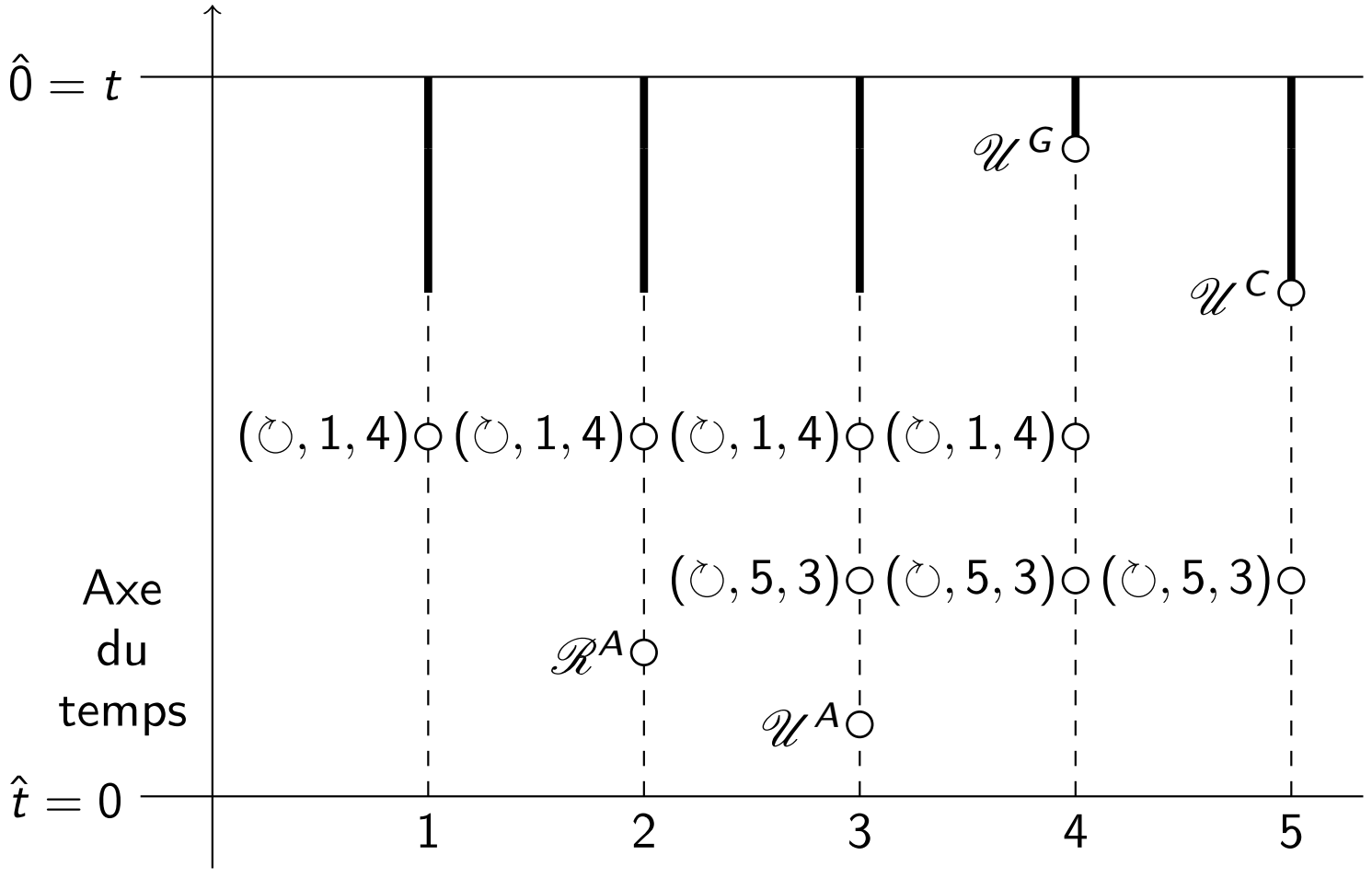
Introduction of a branching process



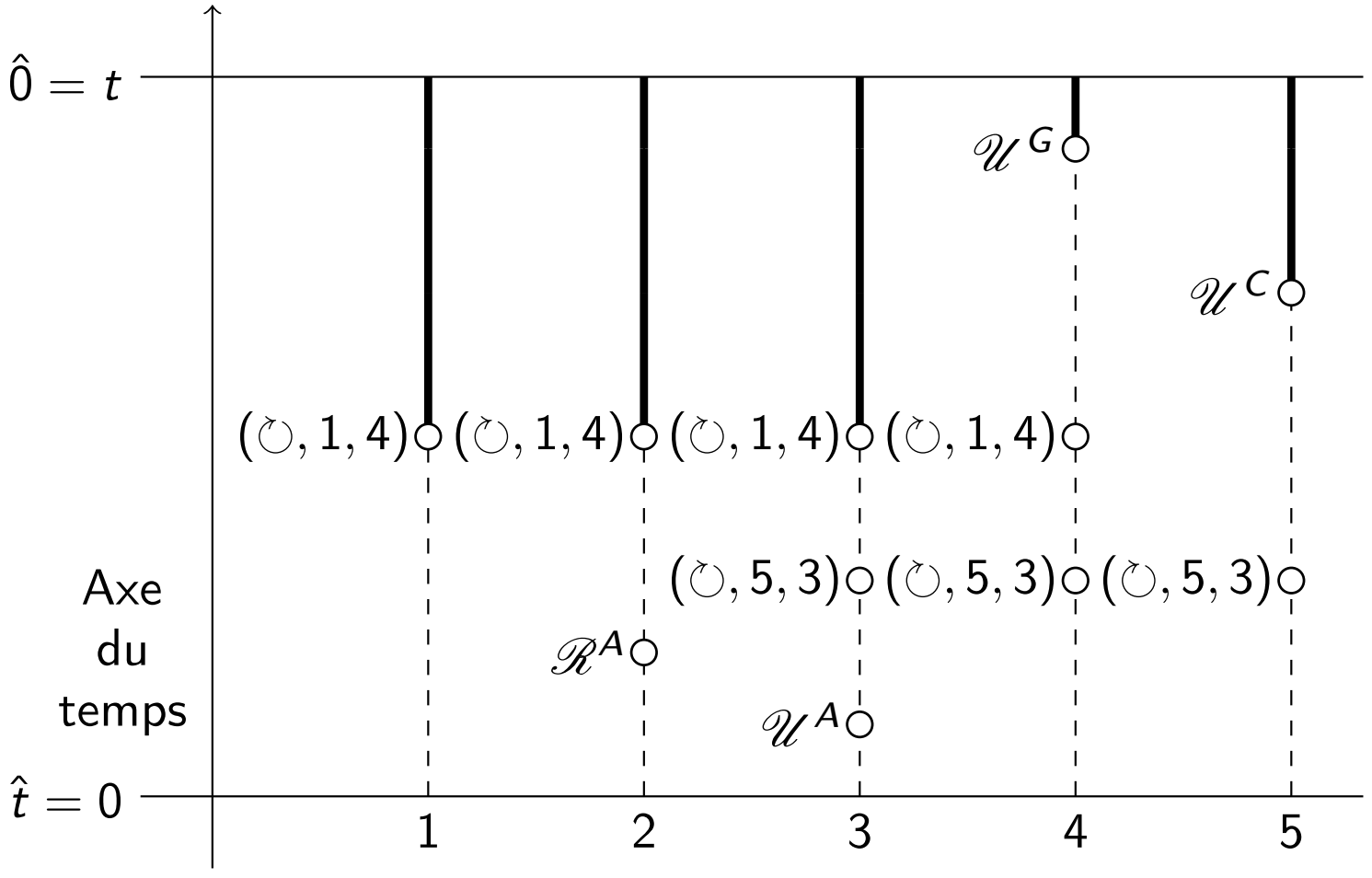
Introduction of a branching process



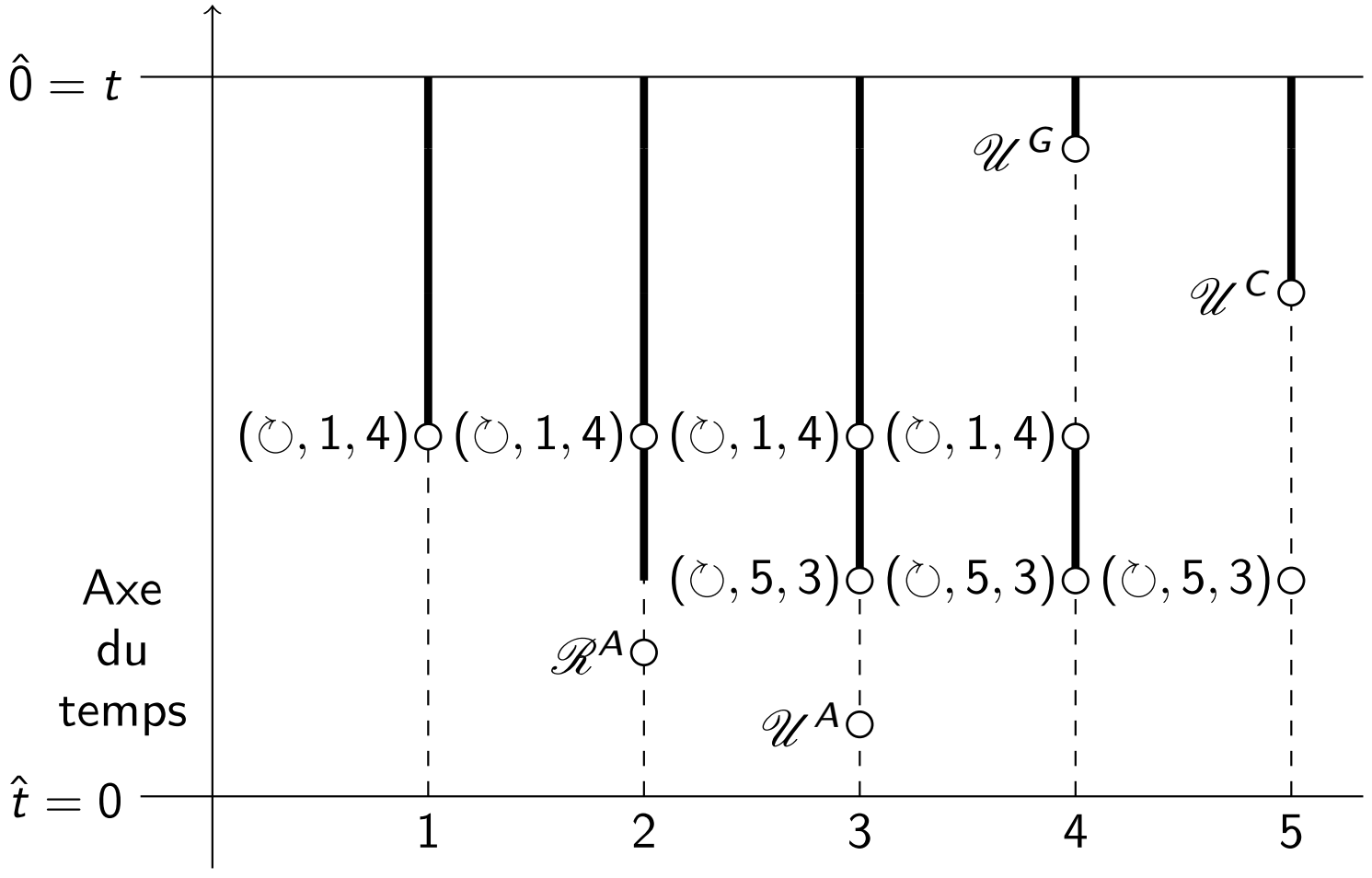
Introduction of a branching process



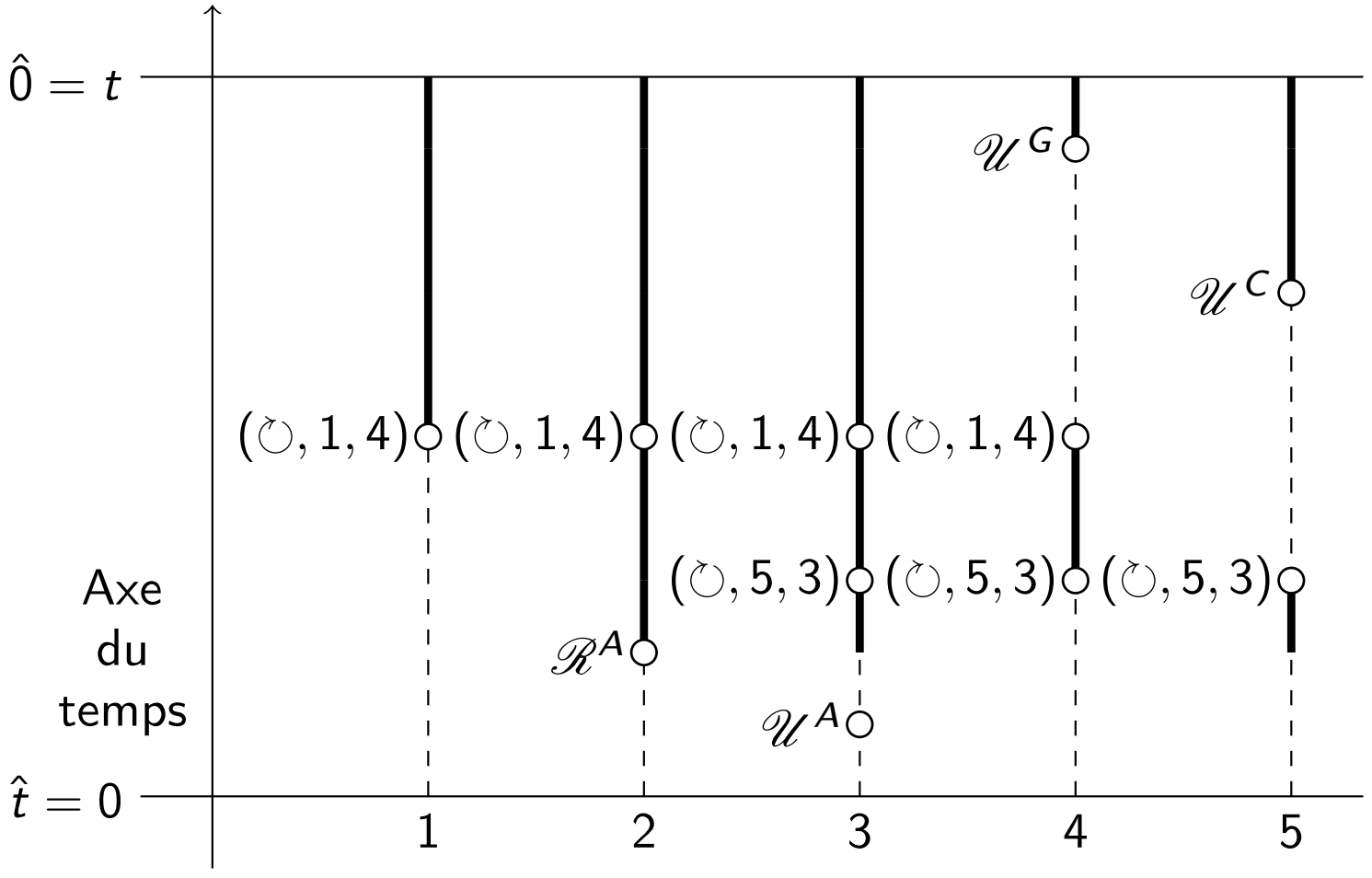
Introduction of a branching process



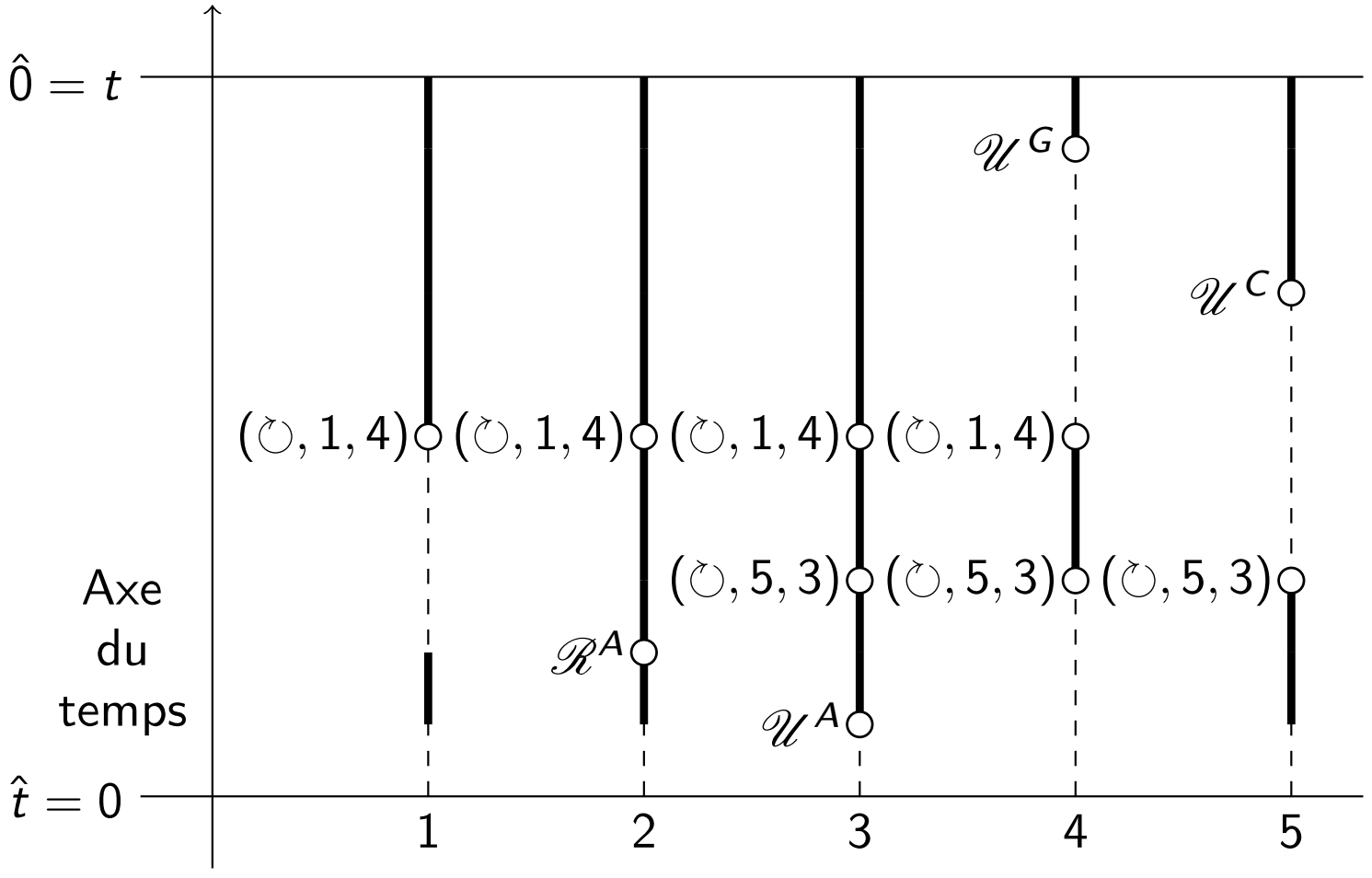
Introduction of a branching process



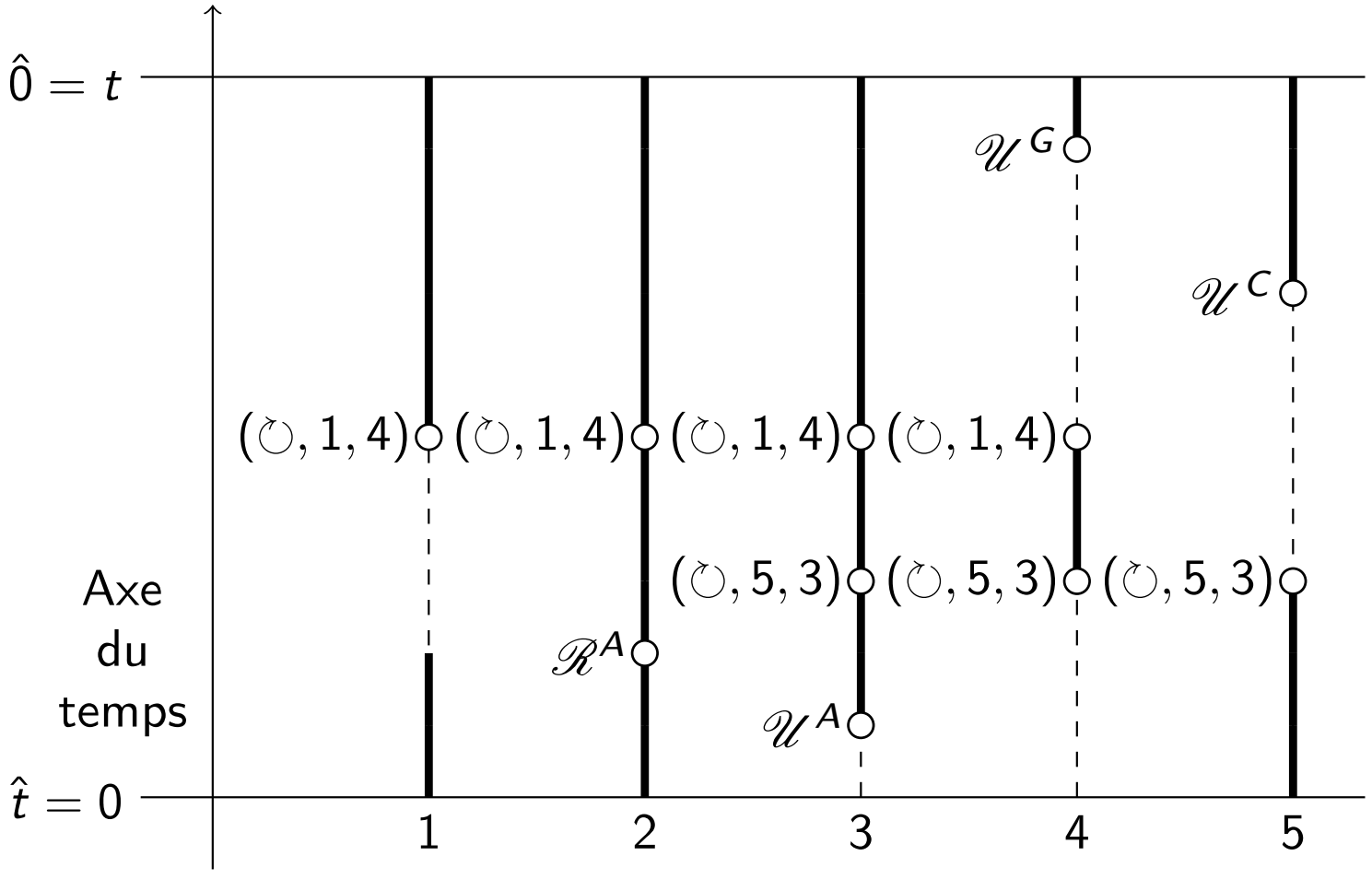
Introduction of a branching process



Introduction of a branching process



Introduction of a branching process



Results for independent evolution models

Falconnet, Gantert and Saada, (2012)

Assume that the substitution rates are **independent** and are Markovian. The the process is **ergodic** and the invariant measure is the product measure on $\mathbb{Z}$.

- Especially, for any usual substitution model (JC69, K80, T92, etc.) and any translocation mechanism invariant by translation, the dynamic of the process is ergodic.

Results for spin + stirring models

Neuhauser, *Annals of Probability* (1990)

Consider the Ising model. If the rate of stirring is small enough, then the process remains **ergodic**.

Ferrari, *Annals of Probability* (1990)

Define

$$\begin{aligned} m &= \inf\{c(x,\eta) \,:\, x \in \mathbb{Z}, \eta \in X\}, \\ K &= \sup\{c(x,\eta) \,:\, x \in \mathbb{Z}, \eta \in X\}, \\ h &= \max_{x\in\mathbb{Z}} |R_x|. \end{aligned} \tag{2}$$

Then if $m > 0$ and if

$$(h-1)(K-m) < 2m,$$

the process is **exponentially ergodic**.

Results for substitution process with translocation mechanism

Falconnet, Gantert et Saada, (2012)

Ferrari's result can be transposed. Define

$$\begin{aligned} m &= \inf\{c(x,a,\eta) \,:\, x \in \mathbb{Z}, a \in \mathscr{A}, \eta \in X\}, \\ K &= \sup\{c(x,a,\eta) \,:\, x \in \mathbb{Z}, a \in \mathscr{A}, \eta \in X\}, \\ h &= \max_{x\in\mathbb{Z}, a\in\mathscr{A}} |R_x^a|. \end{aligned} \tag{3}$$

Then if $m > 0$ and if

$$(h-1)(K-m) < |\mathscr{A}|m,$$

the process is **exponentially ergodic**.

Especially, JC+CpG+Translocation model is ergodic as soon as

$$r < 4\lambda.$$

Open questions

Contact process

The contact process is such that

$$c(x,\eta) = \begin{cases} \lambda[\eta(x-1)+\eta(x+1)] & \text{if} \quad \eta(x)=0, \\ 1 & \text{if} \quad \eta(x)=1, \end{cases}$$

where $\lambda \geqslant 0$. One can see that

$$m = \inf\{c(x,\eta) \,:\, x \in \mathbb{Z}, \eta \in X\} = 0,$$

hence the theorem cannot be used there.

One can show that there exists a critical value $\lambda_c(\rho)$ depending on ρ, but we do not know its behavior. At the moment, we only know that

$$\forall \rho \geqslant 0, \quad \lambda_c(\rho) \geqslant \frac{1}{2}, \quad \text{and}$$

$$\lambda_c(0) - \rho \leqslant \lambda_c(0) \leqslant (1+2\rho)\lambda_c(0).$$

Questions ouvertes

Modèle d'Ising

The one dimensional Ising model is defined as

$$c(x,\eta) = \begin{cases} \mathrm{e}^{-2\beta} & \text{if} \quad \eta(x-1) = \eta(x) = \eta(x+1), \\ \mathrm{e}^{2\beta} & \text{if} \quad \eta(x) \neq \eta(x-1) = \eta(x+1), \\ 1 & \text{else.} \end{cases}$$

This process is ergodic for any $\beta \geqslant 0$. We think that translocation mechanism should not change this fact.

Statistics

Would it be possible to use this model to improve DNA sequences alignment ?