On some usual and some not-so-usual models of the evolution of nucleotidic sequences

Couplings with ambiguities and neighbour-dependent DNA dynamics

Didier Piau

Institut Fourier

&

Laboratoire d'Écologie Alpine

Grenoble

ALÉA 2011

[36/551 = 6,5%]

SETTING

Nucleotidic sequences: alphabet $\{A, C, G, T\}$, finite or infinite length, evolution by substitutions only (no insertion, no deletion).

In usual models, substitution rates at a site depend on the nucleotide at this site only. ((Or codon models.))

Hence:

– The dynamics of each site is independent on all the others.
– The evolution of a given site is ruled by a copy of a Markov process on $\{A, C, G, T\}$.
– Each site converges in distribution to the stationary distribution.
– The sequence converges in distribution to the product of the stationary distributions.

INDEPENDENT MODELS

At every site substitution rates: rate$(x \to y) =: q(x, y)$.

Dynamics of a given site : if $X(t) = x$, then

$$X(t + \mathrm{d}t) = \begin{cases} y \text{ with probability } q(x, y)\mathrm{d}t \text{ if } y \neq x, \\ x \text{ with probability } 1 + q(x, x)\mathrm{d}t \end{cases}$$

$$q(x, x) := -\sum_{y \neq x} q(x, y).$$

The **Q-matrix** characterizes the dynamics:

$$Q = (q(x, y))_{x,y}$$

1) When to jump If $X(t) = x$, $X(u) = x$ for every u in the time interval $[t, t+s[$ with probability

$$(1 + q(x,x)\mathrm{d}t)^{s/\mathrm{d}t} \approx \mathrm{e}^{q(x,x)s}.$$

That is, one leaves x after an exponential time of parameter $|q(x,x)|$ and mean $1/|q(x,x)|$.

+ Bus paradox: mean sojourn in state x is $1/|q(x,x)|$.

2) Where to jump If $X(t)$ leaves x at time $t+s$, $X(t+s) = y$ with probability

$$q(x,y)/|q(x,x)|$$

independently of the time of the transition.

The result $(X(t))_t$ is a Markov process with transitions

$$\mathbb{P}(X(t+s) = y | X(t) = x) = (\mathrm{e}^{sQ})(x, y)$$

with

$$\mathrm{e}^{sQ} = \sum_{n=0}^{+\infty} \frac{s^n}{n!} Q^n$$

The term Q^n corresponds to n transitions from one state to another state during the time interval $[t, t+s[$.

If $\mathbb{P}(X(0) = x) = \mu(x)$ for every x, $X(t)(\mathbb{P}) = \mu \cdot \mathrm{e}^{tQ}$ in the sense that

$$\mathbb{P}(X(t) = x) = (\mu \cdot \mathrm{e}^{tQ})(x) = \sum_y \mu(y) \cdot (\mathrm{e}^{tQ})(y, x).$$

The continuous time process $(X(t))_t$ contains several discrete time Markov chains...

Embedded chain The sequence $(\xi_n)_n$ of the different states successively occupied by $(X(t))_t$ is a Markov chain in the usual (discrete) sense

$$\mathbb{P}(\xi_{n+1} = y | \xi_n = x) = q(x,y)/|q(x,x)| \quad (y \neq x).$$

Transition matrix of $(\xi_n)_n$: $P = (p(x,y))_{x,y}$ with

$$p(x,x) := 0, \quad p(x,y) := q(x,y)/|q(x,x)| \quad (y \neq x).$$

P nonnegative matrix whose every line sums to 1.

To recover $(X(t))_t$ from $(\xi_n)_n$, use (conditionally) independent random times $(\tau_n)_n$

$$X(t) = \xi_n \text{ while } \tau_0 + \cdots + \tau_{n-1} \leqslant t < \tau_0 + \cdots + \tau_n$$

with τ_n exponential of parameter $|q(\xi_n, \xi_n)|$.

Discretized Markov chain For every scale h, $\zeta_n^{(h)} = X(nh)$ yields (another) Markov chain

$$\mathbb{P}(\zeta_{n+1}^{(h)} = y | \zeta_n^{(h)} = x) = (\mathrm{e}^{hQ})(x, y)$$

Conversely,

$$X(t) \approx \zeta_{t/h}^{(h)} \quad (h \to 0)$$

Caveat: Not every Markov chain is a ζ chain (embeddability problem).

Summary (Continuous time) Markov process = (discrete time) Markov chain, cooled down and sped up.

Main question: long time behaviour

Key-object: **stationary distribution**

(Def) Probability measure π such that if $X(0) \rightsquigarrow \pi$ then $X(t) \rightsquigarrow \pi$

$$\pi Q = 0, \quad \sum_y \pi(y) q(y,x) = 0$$

Equilibrium in/out: fraction $|q(x,x)|$ of $\pi(x)$ leaves x (out) while fractions $q(y,x)$ of $\pi(y)$ arrive on x coming from each y (in).

1) Stationary distribution = ergodic limit

$$\int_0^t [X(t) = x] \, \mathrm{d}t = \pi(x)t + o(t) \text{ almost surely}$$

Hence, π is accessible from the observation of a single sample path.

2) Stationary distribution = long time asymptotics

For every initial distribution of $X(0)$, the distribution of $X(t)$ converges to π in the sense that $\mathbb{P}(X(t) = x) \to \pi(x)$ for every x,

$$\|X(t)(\mathbb{P}) - \pi\| \to 0$$

Idea: $e^{tQ} \to \Pi$ where $\Pi(x, y) = \pi(y)$ and $\mu\Pi = \pi$ for every probability measure μ, hence $\mu e^{tQ} \to \pi$

Spectral decomposition of Q: $\mu_\lambda Q = \lambda\mu_\lambda$ with $\mu_0 = \pi$, $\mu_\lambda e^{tQ} = e^{t\lambda}\mu_\lambda \to 0$ for every $\lambda \neq 0$, $\mathrm{Re}(\lambda) < 0$, hence

$$(*) \qquad \|X_t(\mathbb{P}) - \pi\| \leqslant Ke^{-tG(Q)},$$

with K constant and $G(Q) :=$ spectral gap of Q.

Measuring the distance between probability measures: the total variation distance

$$\|\mu - \nu\| = \max_B \mu(B) - \nu(B)$$

Realization by random variables (=coupling):

$$\|\mu - \nu\| = \min\{\mathbb{P}(U \neq V)\,;\, U \rightsquigarrow \mu,\, V \rightsquigarrow \nu\}$$

Consequence: $\mu = \mu_1 \otimes \cdots \otimes \mu_N, \nu = \nu_1 \otimes \cdots \otimes \nu_N$

$$(**) \qquad \|\mu - \nu\| \leqslant \|\mu_1 - \nu_1\| + \cdots + \|\mu_N - \nu_N\|$$

Back to the evolution of DNA sequences

Box of size N: convergence of the distribution after time t_N with

$$N\mathrm{e}^{-t_N G(Q)} \approx 1 \text{ hence } t_N = \Theta(\log N)$$

Simplest model of DNA evolution: Jukes-Cantor

$$\text{rate}(x \to y) = u, \quad x \neq y \text{ in } \{A, C, G, T\}$$

Q-matrix:

$$\begin{pmatrix} - & u & u & u \\ u & - & u & u \\ u & u & - & u \\ u & u & u & - \end{pmatrix}$$

Construction of the dynamics: on each site, place a (time) Poisson process of intensity $4u$, when one meets a Poisson event, replace x by a uniform letter.

If at least one Poisson event is already met, the distribution is π, hence the distribution of $X_{1:N}$ is $\pi^{\otimes N}$ unless at least one site did not meet a Poisson event yet:

$$\|X_{1:N}(t)(\mathbb{P}) - \pi^{\otimes N}\| \leqslant 1 - (1 - \mathbb{P}(\text{Poisson}(4ut) = \emptyset))^N$$

hence

$$\|X_{1:N}(t)(\mathbb{P}) - \pi^{\otimes N}\| \leqslant N\mathrm{e}^{-4ut}$$

(And $Q^2 = -4uQ$ hence $G(Q) = 4u$.)

Convergence of a box of N sites: distance at most e^{-s} at time

$$t(s) = (4u)^{-1}(\log(N) + s)$$

Same kind of result for every matrix of rates with $4u$ replaced by the spectral gap $G(Q)$.

Second simplest example of DNA evolution: Kimura Different rates of transitions and of transversions

$$Y = \{C, T\} \text{ (pyrimidines)}, \quad R = \{A, G\} \text{ (purines)}$$

$$\text{rate}(x \to y) = \begin{cases} u \text{ if } x \neq y \text{ belong to the same class,} \\ v \text{ otherwise.} \end{cases}$$

Q-matrix:

$$\begin{pmatrix} - & u & v & v \\ u & - & v & v \\ v & v & - & u \\ v & v & u & - \end{pmatrix}$$

Phylogenies

Problem: given n present sequences, build the tree of life with these sequences as leaves.

Step zero: how to compute the (evolutionary) distance (=time elapsed) between two sequences?

Assume X' and X'' evolved during time t from a MRCA X. Comparing X' and X or X'' and X yields the time t elapsed between the ancestral sequence X and the present ones. For example,

$$\text{Proportion}\{k; X'_k = X_k\} \approx p_0(t)$$

with

$$p_0(t) = \sum_x \pi(x)(\mathrm{e}^{tQ})(x,x)$$

Then solve equation in t...

For JC, the (normalised) elapsed time is

$$t_{X\to X'}(D) = -\frac{3}{4}\log\left(1-\frac{4}{3}D\right)$$

with $D =$ proportion of non coinciding sites in X and X'.

But X is not known hence one must compare X' and X''. Let

$$p(t) = \mathbb{P}(X'_k = X''_k)$$

Assuming X is at stationarity

$$p(t) = \sum_x \pi(x) \sum_y (\mathrm{e}^{tQ})(x,y)(\mathrm{e}^{tQ})(x,y)$$

Hypothesis: reversibility $\pi(x)q(x,y) = \pi(y)q(y,x)$

Then $\pi(x)(\mathrm{e}^{tQ})(x,y) = \pi(y)(\mathrm{e}^{tQ})(y,x)$ hence

$$p(t) = \sum_x \pi(x)(\mathrm{e}^{2tQ})(x,x)$$

In a nutshell: The problem is equivalent to assuming that X' produced X'' in time $2t$.

For JC,

$$t_{X'-X''}(D) = -\frac{3}{8}\log\left(1 - \frac{4}{3}D\right)$$

with $D =$ proportion of non coinciding sites in X' and X''.

Estimators statistically founded (MLE)

TOWARDS MODELS WITH DEPENDENCE

Biologists know that:
(a) the observed frequencies are not product ones (not nearly),
(b) the substitution rates at a site do depend on the neighbours of the site.

One massive and well known example: CpG islands

> G $\to$ A [up to] 10fold when G is in CG (in fact CG*),
> (hence) C $\to$ T [up to] 10fold when C is in CG.

General model: Substitution rate $x \to x'$ at each site

$$\text{rate}(x \to x' \text{ if } x \text{ in } yxz).$$

Dependency cone: (Stationary) frequencies (x) depend on frequencies (yxz), which depend on frequencies $(uyxzv)$, etc.

Hence one is stuck.

Approximate solutions of models with "double" substitutions related to CpG effects

Duret & Galtier, *Molecular Biology and Evolution* (2000)
Tamura model with 2 parameters + CpG $\rightarrow$ CpA and TpG at the same rate.

Here, each (x) depends on some (xy), each (xy) depends on some (xyz), etc.

Idea: $$(xyz) \approx (xy)\,(yz)/(y) \qquad (*)$$

Note:
– In other contexts, this is called Bethe Ansatz, Kikuchi approximation, cluster approximation, etc.
– Formula $(*)$ would be exact for a (spatial) Markov chain.

Assuming $(*)$, the 16 frequencies (xy) are solutions of an autonomous nonlinear system.
Which can be solved, at least numerically. The solutions $(xy)^*$ are close to a true distribution (all positive and summing to almost 1).

Duret & Galtier: TpA frequency also modified, no need of an auxiliary mechanism (call this a *mathematical artefact*). Typically

$$\mathrm{CpGo/e} \ll \mathrm{TpAo/e} \ll 1.$$

((Mention here : Arndt, Burge, Hwa, Jensen, Pedersen, Lunter, Hein, others.))

How to check the effect of approximation $(*)$?

Simulations (?).
Linear finite box, or discrete finite circle: close to the behaviour of the system on the line (or not)? For what size of the box?

((Add here: voter model.)) ((Add here: Gács & Gray.))

Summary of the results

For RN+YpR models,

(a) the system converges to a unique stationary measure, invariant by (spatial) translations,
(b) one can quantify the rate of convergence,
(c) one can compute exactly the marginals at equilibrium (polynucleotidic frequencies),
(d) the equilibrium measure has some strong (unforeseen) independence properties.

Results (cont'd)

For RN+YpR models,

(e) one can (begin to) build phylogenies.

For models in a "neighborhood" of RN+YpR,

(a) convergence and (b) rate of convergence still valid,
(c) exact marginals replaced by asymptotic expansions,
(d) independence replaced by explicit bounds of the decrease of the correlations.

(A FAIRY TALE BASED ON) POISSON DYNAMICS

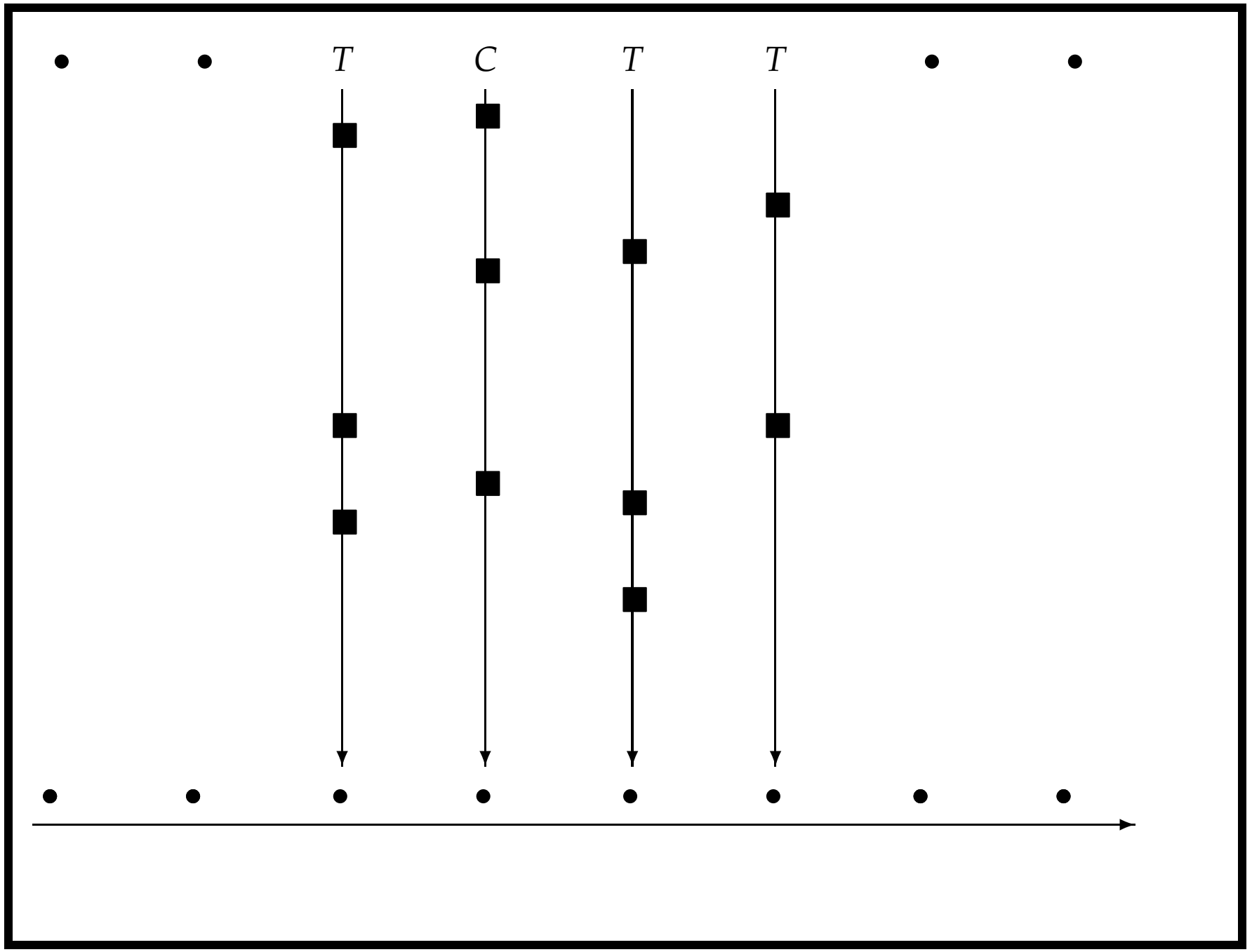
■ = "simple" substitutions.

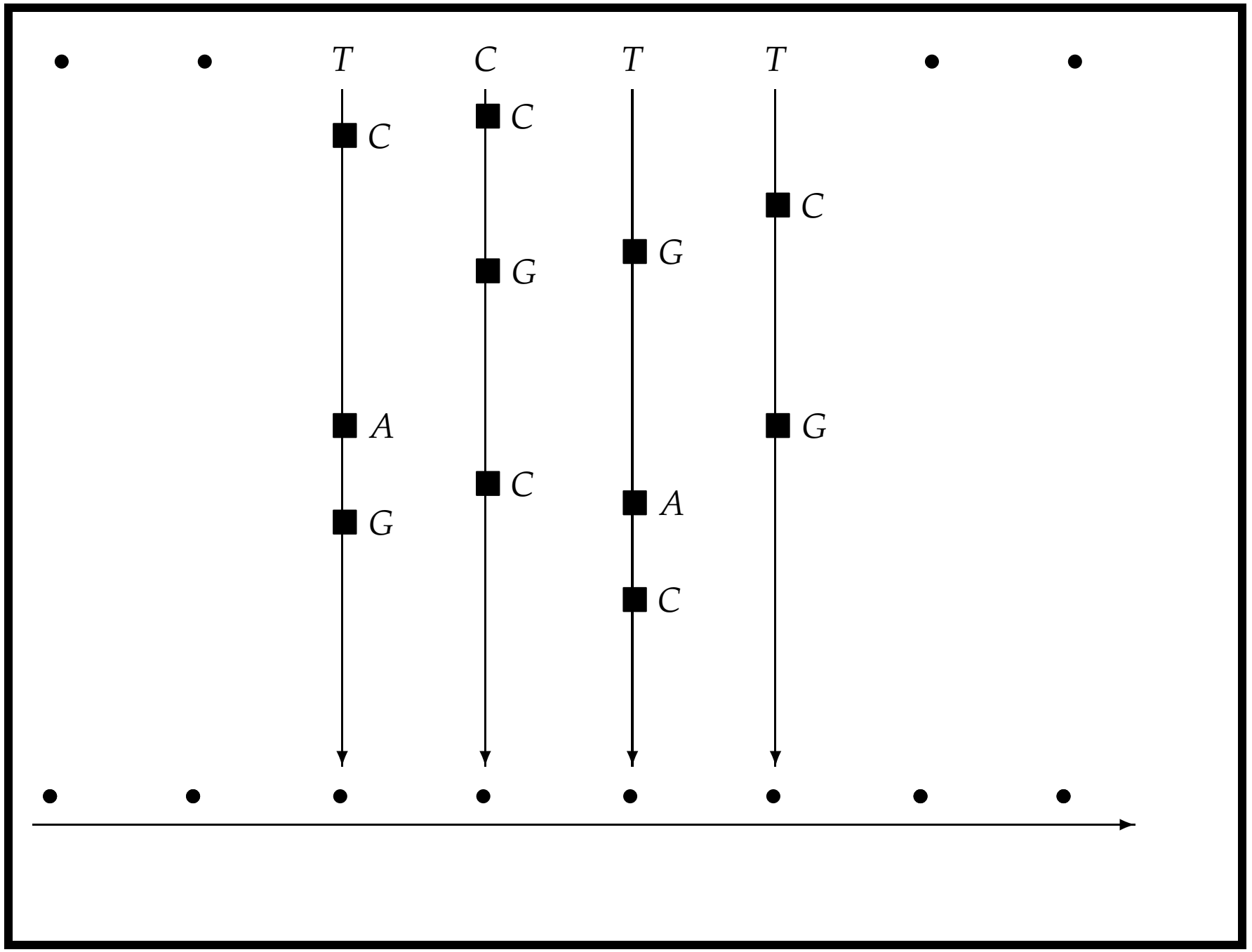
$\tau(x, y)$ = rate of substitution of x by y.

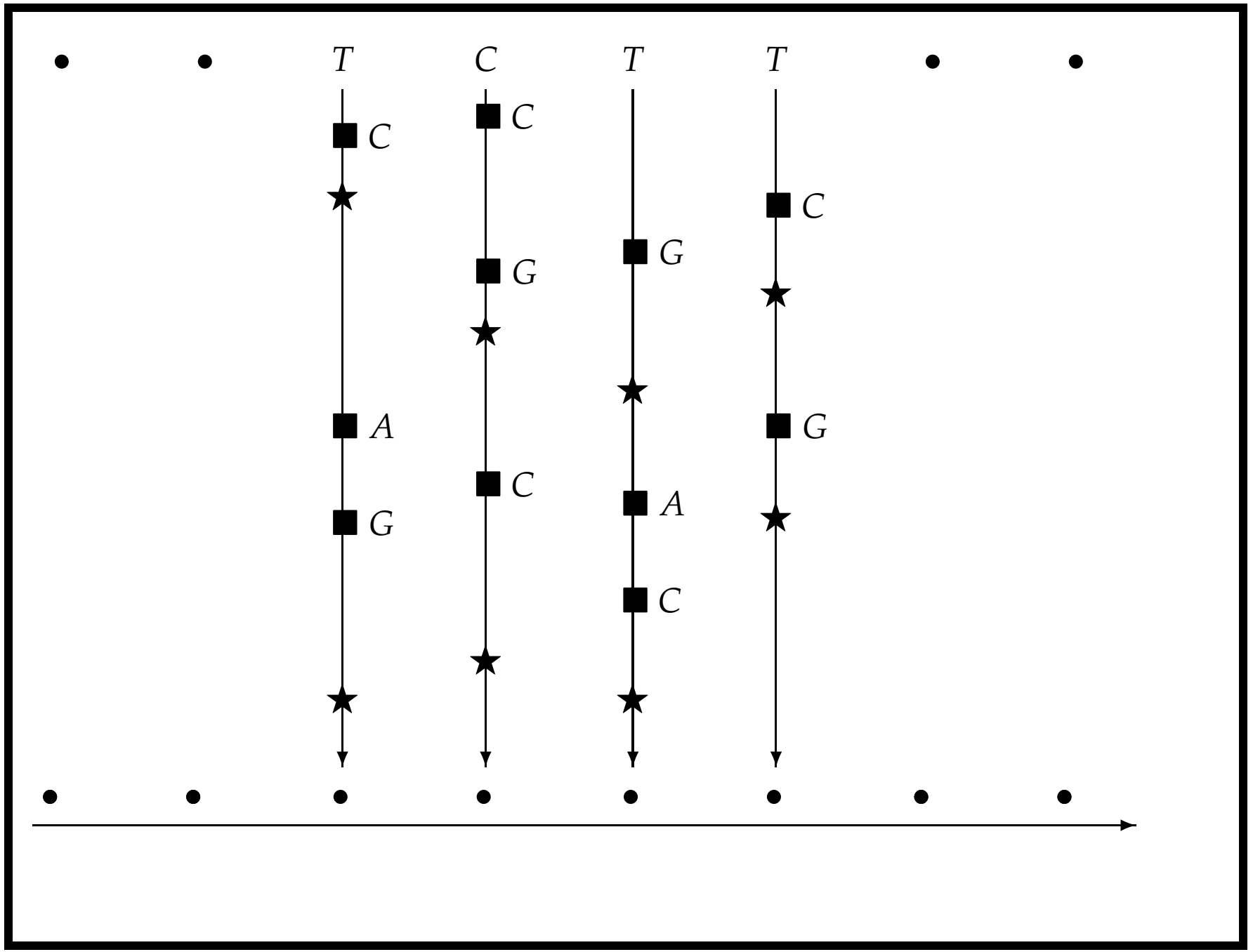
Caution: one authorizes simple (virtual) substitutions of x by x, hence every $\tau(x, x)$ is a free parameter.

★ = "double" substitutions of CpG by CpA or TpG.

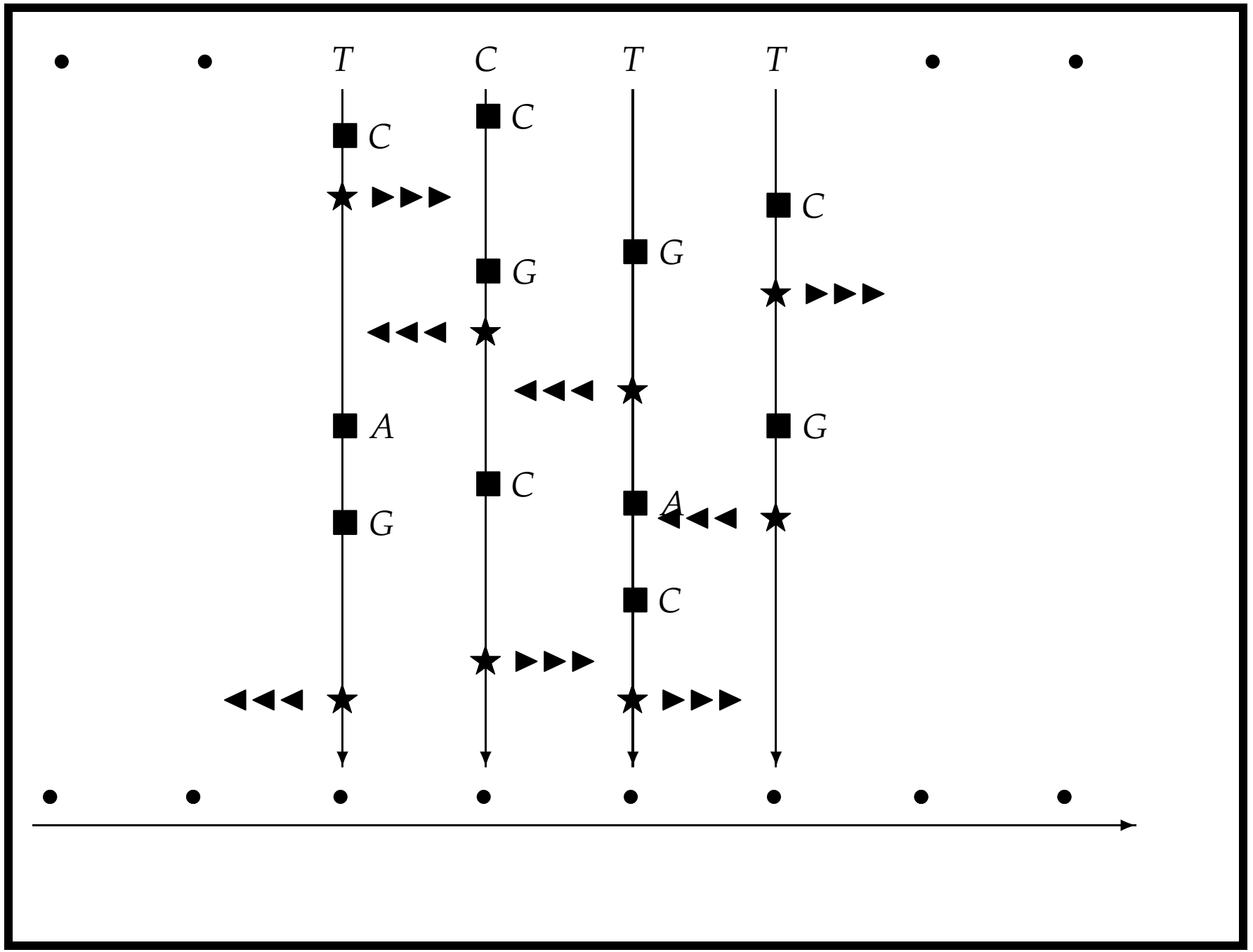
One wants to represent the **evolution of a finite collection of sites**.


T
C
T
T

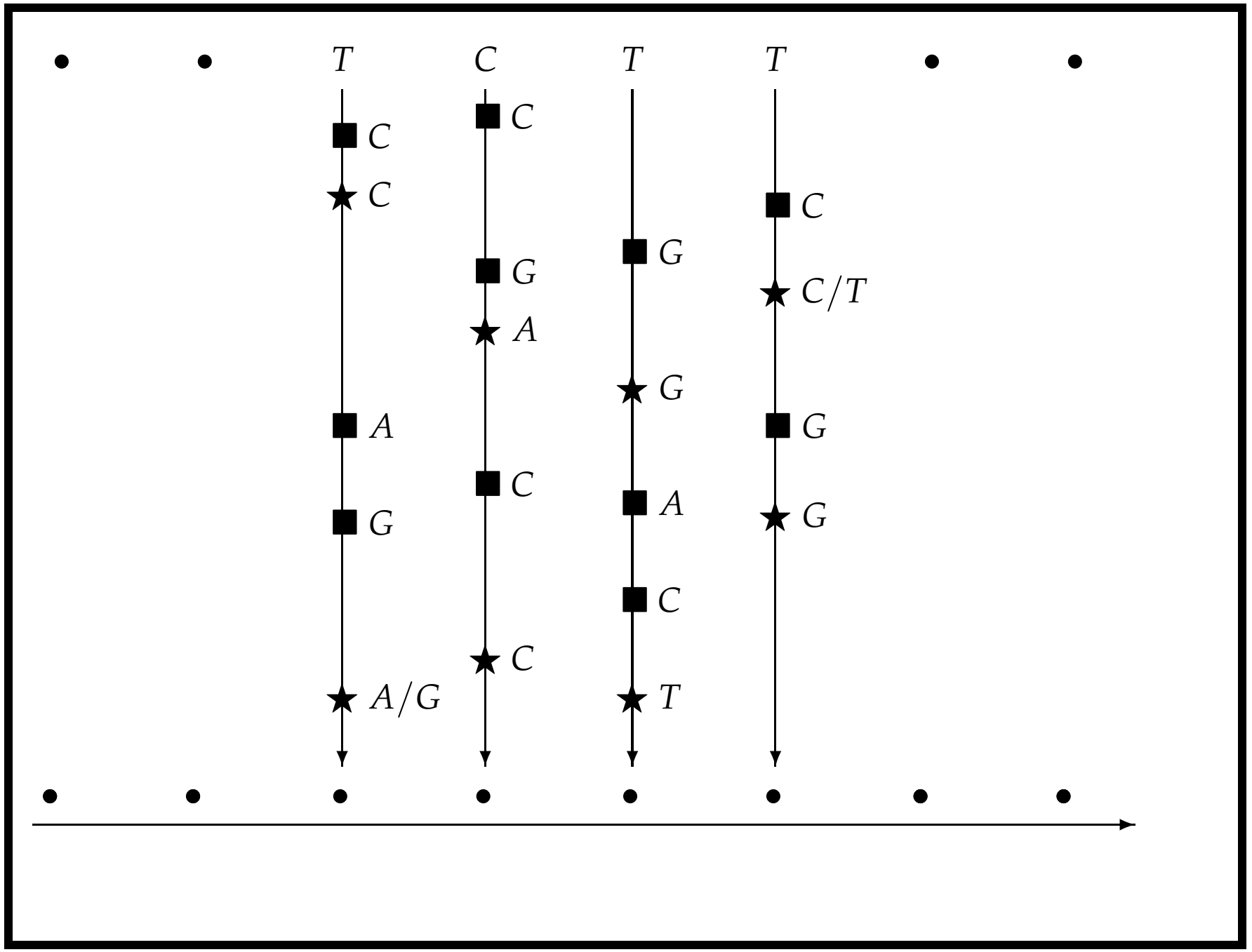

T
C
T
T
C
C
G
C
G
A
C
G
A
C

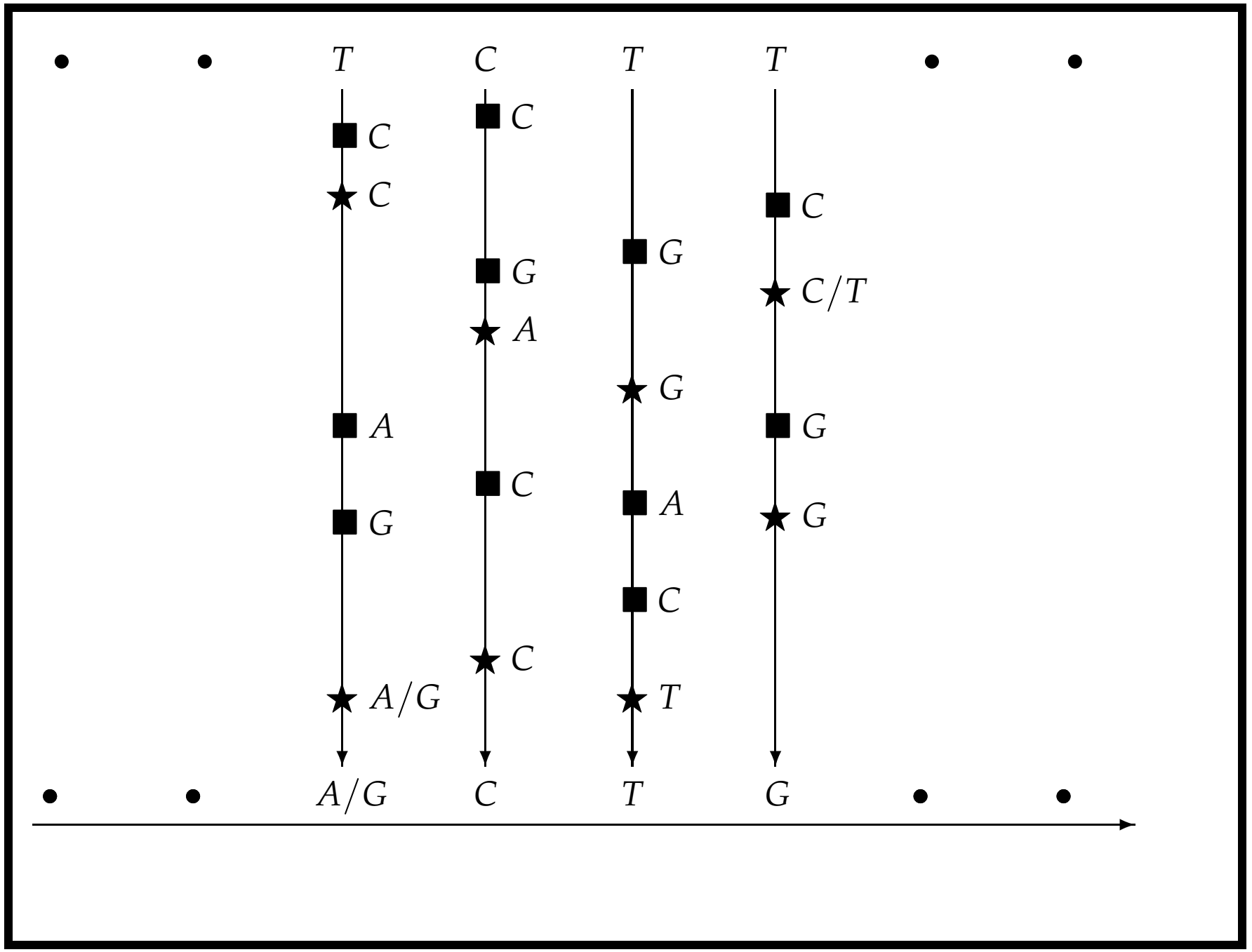

T
C
T
T
C
C
G
C
G
A
C
G
A
C

To decide which “double” substitutions really occur, at each ★ one must read the nucleotide on the left ◄◄◄ or on the right ►►►.


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

As a consequence, one has to put some ”wildcards” on the first and last columns. This yields the present sequence, modulo some (unknown) values at the wildcards.

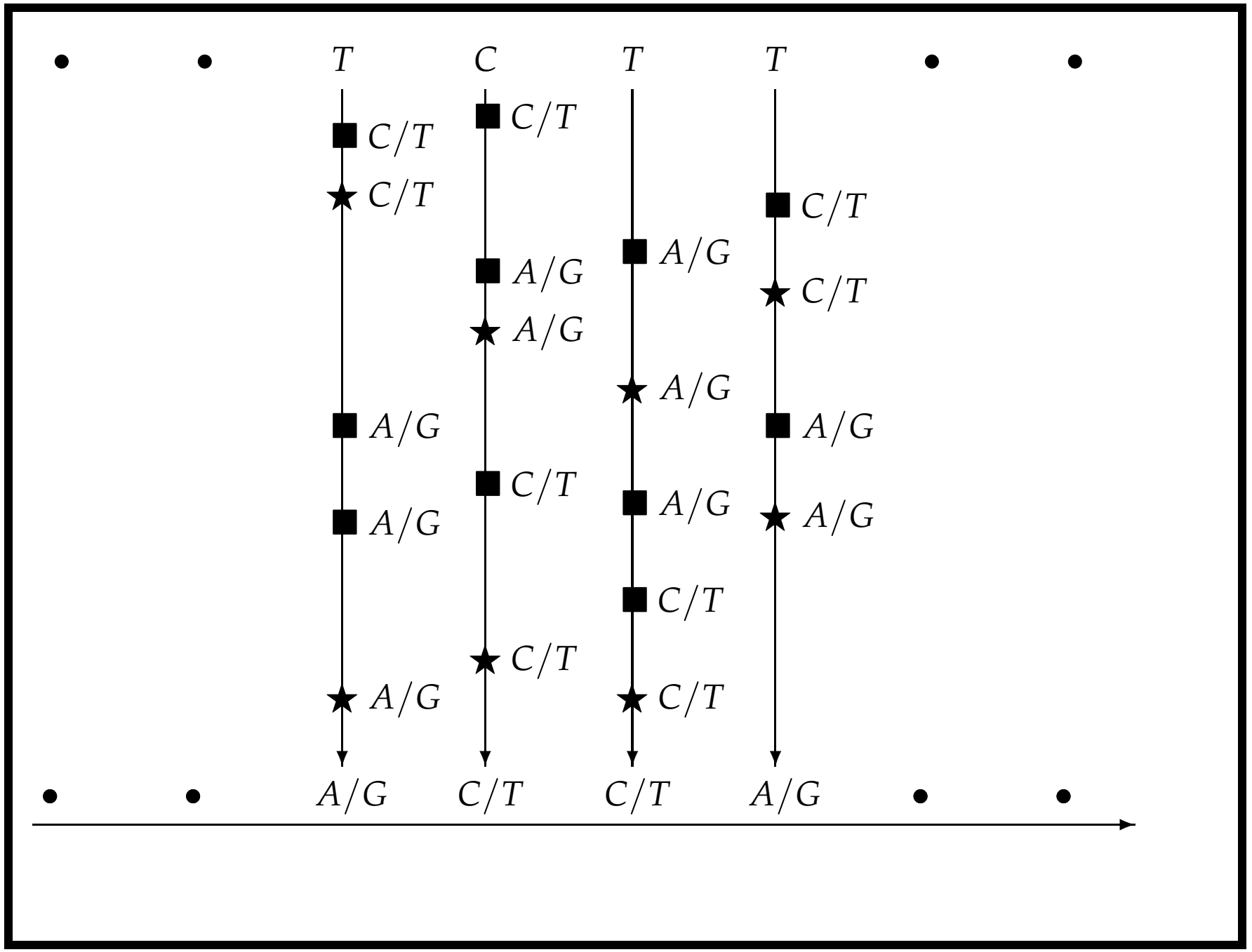

T
C
T
T
C
C
C
C
G
G
A
A
G
G
C
C
A
G
C
C
T
A/G
C/T
G
G

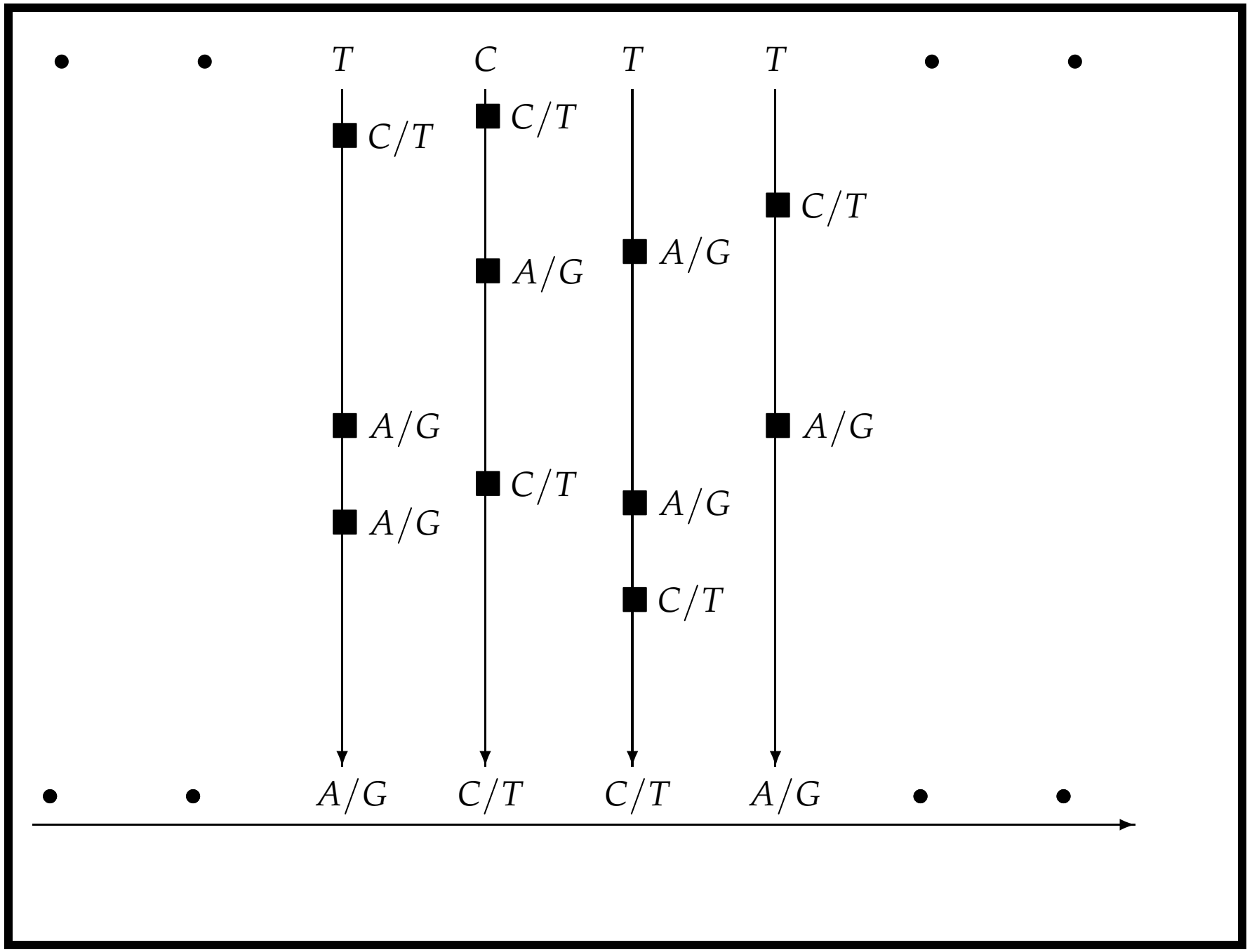

T
C
T
T
C
C
C
G
C/T
G
A
G
G
A
G
C
C
A/G
T
A/G
C
T
G

R-Y quotient

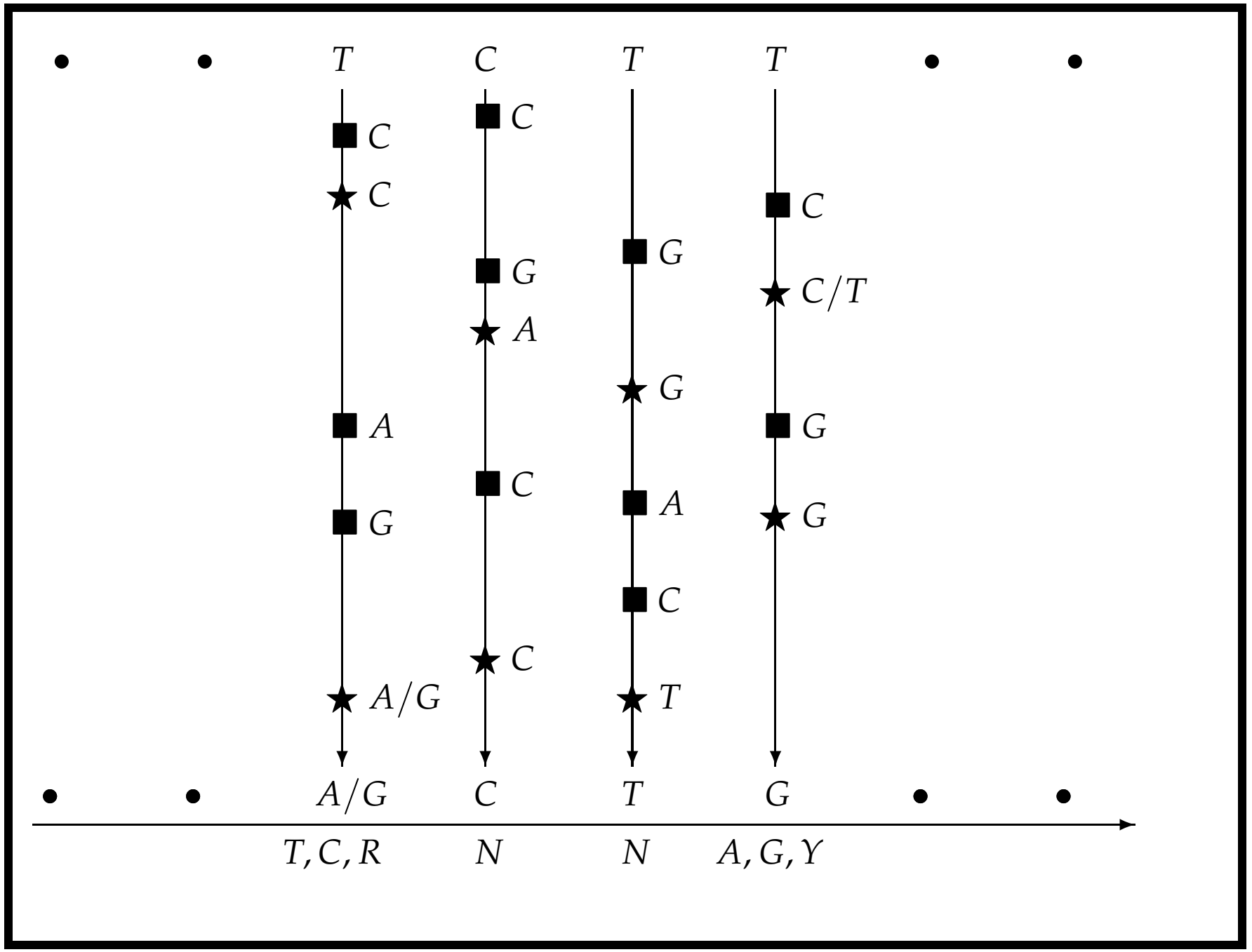
$R = \{G, A\} =$ purines, $Y = \{C, T\} =$ pyrimidines.

In the $\{R, Y\}$ alphabet, the "double" substitutions become useless.


T
C
T
T
C/T
C/T
C/T
A/G
A/G
A/G
C/T
A/G
A/G
C/T
A/G
C/T
C/T
A/G
A/G
C/T
A/G
A/G
A/G
C/T
C/T
A/G
C/T
C/T
A/G


T
C
T
T
C/T
C/T
C/T
A/G
A/G
A/G
A/G
C/T
A/G
A/G
C/T
A/G
C/T
A/G
C/T
A/G

More on the effect of the wild cards…


T
C
T
T
C
C
C
C
G
G
A
G
A
C
G
C
C
A/G
T
G
A/G
C
T
G
T, C, R
N
N
A, G, Y

When our (fairy tale) construction yields the desired evolution

■ and ★ are events of Poisson processes hence the waiting times are exponentials.

With no ★:
One "leaves" x by a substitution of type ■ after an exponential time of parameter

$$\tau(x) := \sum_y \tau(x, y),$$

and the distribution of the "successor" of x is

$$\sigma_x(\cdot) := \frac{\tau(x, \cdot)}{\tau(x)}.$$

With the ★ :

One replaces (or not) C by T (idem for G and A). For what effect?

(1) The remaining time before leaving C is exponential again, with parameter $\tau(C)$ [bus stop paradox], instead of exponential with parameter $\tau(T)$.

(2) The distribution of the successor is still $\sigma_C(\cdot)$, instead of $\sigma_T(\cdot)$.

This does not matter if

$$\tau(C) = \tau(T), \quad \sigma_C(\cdot) = \sigma_T(\cdot),$$

and

$$\tau(G) = \tau(A), \quad \sigma_G(\cdot) = \sigma_A(\cdot).$$

Finally the (fairy tale) construction works fine if

$$\boxed{\tau(C,\cdot) = \tau(T,\cdot), \qquad \tau(A,\cdot) = \tau(G,\cdot).}$$

Suitable substitution rates

Abstract of the construction

"Double" substitutions : do not change the class R-Y.

"Simple" substitutions of (x in) R by (y in) Y: the rate must not depend on x but only on R and y.

Idem for "simple" substitutions of Y by R.

"Simple" substitutions: RN class (Rzhetsky-Nei)

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} - & v_T & v_C & w_G \\ v_A & - & w_C & v_G \\ v_A & w_T & - & v_G \\ w_A & v_T & v_C & - \end{pmatrix} \end{array}$$

Classic RN examples: Tamura, Tamura-Nei,
Hasegawa-Kishino-Yano, Kimura, Jukes-Cantor, etc.

But not GTR in general.

“Double” substitutions

Every rate between CpG, CpA, TpG and TpA, that is, in the class YpR.

Consequences

R-Y quotient

The evolution of each site encoded by $\{R, Y\}$ is autonomous and Markov, with substitution rates

$$\text{rate}(R \to Y) = v_T + v_C, \quad \text{rate}(Y \to R) = v_A + v_G.$$

Hence, the distribution of the sequence of purines/pyrimidines converges to the product of the Bernoulli distributions

$$(R) = (v_A + v_G)/v, \quad (Y) = (v_T + v_C)/v,$$

with $v = v_A + v_C + v_G + v_T$.

Consequences (cont'd)

Finer quotient

Window of width $n+2$: autonomous evolution of the quotient state in

$$\{R, C, T\} \times \{A, C, G, T\}^n \times \{Y, G, A\}.$$

Hence: stationary distribution of the polynucleotides of length n is a marginal of the stationary distribution of a Markov chain on 9×4^n states, hence one can **compute** it and **simulate** it.

Or better still: discrete circle of size $n+2$ hence 4^{n+2} states, with rotation invariance.

Consequences (cont'd)

Independence

At stationarity, there exists some i.i.d. objects $(C_i)_i$ such that the nucleotide X_i at site i depends on (C_{i-1}, C_i, C_{i+1}) only.

Hence, sites i and j are independent if $|i - j| \geqslant 3$.

Let $\mathbb{X}_0 := (X_{3i})_i$, $\mathbb{X}_1 := (X_{3i+1})_i$, $\mathbb{X}_2 := (X_{3i+2})_i$.

At stationarity, each $\mathbb{X}_k$ is an i.i.d. sample with the same distribution. (But $\mathbb{X}_k$ is not independent on the others.)

Example: Jukes-Cantor + CpG

$\text{rate}(x \to y) = 1$, $\text{rate}(CG \to CA) = \text{rate}(CG \to TG) = r$, no other "double" substitution.

Counting flows in and out of C, G and CG:

$$\frac{\mathrm{d}}{\mathrm{d}t}(C) = 1 - 4(C) - r\,(CG),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(G) = 1 - 4(G) - r\,(CG),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(CG) = (C) + (G) - (8 + 2r)\,(CG).$$

Autonomous system for (C), (G) and (CG).

Jukes-Cantor + CpG (statics)

$$(C) = (G) = \frac{1}{4}\left(1 - 2\frac{r}{32+10r}\right),$$

$$(CG) = \frac{1}{16}\left(1 - 10\frac{r}{32+10r}\right).$$

And also (A), (T) and every dinucleotide, such as

$$(AA) = \frac{1}{16}\left(1 + \frac{r}{32+10r}\left(3 + \frac{3r}{96+19r}\right)\right).$$

Excursus: Possible non analyticity in r signaling an accumulation of poles (conjecture).

Simulation and convergence rate

Coupling time T_n for n sites such that $\mathbb{P}(T_n \geqslant t) \leqslant \exp(-s)$ with

$$t = \alpha \cdot (\log(n) + 1 + s),$$

and α explicit function of the "simple" rates v_x and w_x.

Total variation distance between the distribution at time t and the stationary distribution: at most $\exp(-s)$.

To sum up

One converges in distribution after a time of the order of $\log(n)$, while the number of substitutions to perform is

$$\text{Constant} \times n \times \log(n).$$

TOWARDS RN+YpR PHYLOGENIES

Building a phylogeny for the simplest model: (independent) Jukes-Cantor

(i) One compares an ancestral sequence and a present one. MLE yields an estimator of the elapsed time

$$\widehat{T} = -\frac{3}{4}\log\left(1 - \frac{4}{3}D\right),$$

where D is the proportion of different sites.
(ii) By reversibility, the age of the MRCA of two present sequences is $\frac{1}{2}\widehat{T}$.

With influence

Two major problems : (i) Likelihood not computable; (ii) Non reversible.

$(x, y)_t$ = proportion of sites occupied by x at time 0 and by y at time t.

The dynamics of the quantities $(x, y)_t$ solve a finite size linear system.

Hence (several) estimator(s) of time T, solving

$$\widehat{T}_{x,y} \text{ solution de } (x, y)_t = (x, y)_{\text{obs}}.$$

For instance, $\widehat{T}_{C,C}$ is based on the evolution of $(C, C)_t$, which requires the evolutions of $(C, T)_t$, $(C*, *G)_t$, $(C*, *A)_t$, $(C*, CG)_t$, $(C*, CA)_t$, $(C*, TG)_t$ and $(C*, TA)_t$.

Consequences : the “true” divergence times are larger than the times based on the formula valid fo the independent case; and one has rigorous confidence intervals.

What about several present sequences? For two sequences:

$[x, y]_t$ = proportion of sites occupied by x and y in two present sequences which diverged t time ago.

Evolution of the quantities $[x, y]_t$ described by equations similar to those used for the equations $(x, y)_t$.

Unsolved: Find a sound statistical principle to build an estimator based (for example) on the 16 observed frequencies $(x, y)_t$.

((Lacking an idea))

PERTURBED MODELS

Ergodic model $\mathfrak{M}$ + perturbations $\mathfrak{P}$: ergodic?

((Remember: Gács/Gray.)) ((Critical 2D Ising.))

Ingredients based on CFTP ideas (coupling from the past):

The existence of a coupling time allows to forget the starting configuration; being less ambitious, one defines a kind of "weak" coupling time: defects of the coupling are due to some ambiguities, which are structured as a tree, if the tree is almost surely finite everything will work fine.

Control by a Galton-Watson process with mean $m \leqslant C(\mathfrak{M})D(\mathfrak{P})$, where $C(\mathfrak{M})$ depends on the non perturbed model only and $D(\mathfrak{P})$ describes the overall size of the perturbations

$$D(\mathfrak{P}) = \sum_{A \in \mathfrak{P}} \text{rate}(A)\,|\text{context}(A)|.$$

INTERACTING PARTICLE SYSTEMS AS SETS OF ACTIONS

State space S finite, configuration $\eta \in S^{\mathbb{Z}}$.
Action A: rate r & context (B, ℓ, s) with $r \geqslant 0$, $B \subset \mathbb{Z}$ finite box, $\ell \subset S^B$ and state $s \in S$.
Configuration η compatible with action A at site x: $\eta(x + B) \in \ell$.
Configuration $\eta^{x,s}$: $\eta^{x,s}(y) = \eta(y)$ if $y \neq x$, $\eta^{x,s}(x) = s$.

Finite collection $\mathfrak{A}$ of actions, Poisson process Ψ on $\mathfrak{A} \times \mathbb{R} \times \mathbb{Z}$ with intensity

$$\sum_{A \in \mathfrak{A}} \sum_{x \in \mathbb{Z}} r(A) \delta_A \otimes \mathrm{d}t \otimes \delta_x.$$

For each (A, t, x) in Ψ, $\eta_t = \eta_{t-}^{x,s}$ iff η_{t-} is compatible with A at x.

For $T < 0$, η_0 is measurable with respect to η_T and the collection Ψ_T of proposals made at times between T and 0.

((Exercise: Encode Ising by a collection of such actions.))

Ambiguities

A proposal (A, t, x) in Ψ_T is ambiguous if there exists two initial conditions at time T such that A is compatible with η_{t-} for one initial condition and not for the other.

Coupling time with ambiguities: (T, H) with $T \leqslant 0$ almost surely finite, $H \subset \Psi_T$, H has the *stopping property* and $\eta_{0-}(0)$ coincide for every pair of initial configurations *compatible* with (T, H).

> Stopping property: $H \cap \Psi_t$ measurable for Ψ_t.
> Compatibility: each proposal in H is applied either for both initial configurations or for none.

Growth: $$\mathfrak{g}_H = \mathbb{E}\left(\sum_{(A,t,x)\in H} \#B(A)\right).$$

RESULTS

If there exists a coupling time with ambiguities (T, H) with growth $\mathfrak{g}_H < 1$ (subcritical), then

– ergodicity (call π the stationary distribution),
– explicit control of the distance in total variation between the marginals of η_t and π,
– exponential convergence in distribution if T is exponentially integrable,
– decay of correlations bounded by an explicit power of the mean $m < 1$ of the underlying branching process.

A "numerical" application

Perturbed JC & CpG: simple rates $1 + \varepsilon(x, y)$ for $x \to y$ and rates r for CpG $\to$ CpA and for CpG $\to$ TpG.

Size of the perturbation : $|\varepsilon| =$ sum of $|\varepsilon(x, y)|$.

Sub-critical branching process as soon as $|\varepsilon| < \varepsilon_* \approx .3143$.

Decay in $p^{|x-y|/4}$ with $p = |\varepsilon|/\varepsilon_*$.

A different coupling time with ambiguities yields the sufficient condition $|\varepsilon| < \varepsilon_{**}(r)$.

Merci de votre attention

With the participation of

Jean Bérard and Jean-Baptiste Gouéré, and much help and patience from Jean Lobry, Laurent Duret, Manolo Gouy and Laurent Guéguen.
Then Mikael Falconnet, Yves Vandenbrouck, Oscar Gaggiotti, then Marielle Simon, and others.

Merci de votre attention