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

Coupling with ambiguities and neighbour-dependent dynamics of nucleotidic sequences

Didier Piau

Institut Fourier & Laboratoire d'Écologie Alpine

Université Joseph Fourier, Grenoble

SETTING

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

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 a given site substitution rate $x \to y$ rate$(x \to y) =: q(x,y)$.

Leaves x after an exponential time of parameter $|q(x,x)|$, goes to y with probability $q(x,y)/|q(x,x)|$

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

Q-matrix $Q = (q(x,y))_{x,y}$ characterizes the dynamics:
1) $(X(t))_{t \geqslant 0}$ Markov process with semi-group $(\mathrm{e}^{tQ})_{t \geqslant 0}$
2) Long term behaviour and ergodic behavior both described by the stationary distribution π, solution of

$$\pi Q = 0, \quad \sum_{y} \pi(y) q(y,x) = 0$$

3) Quantitative version:

$$\|\mathbb{P}_{X(t)} - \pi\| \leqslant \mathrm{e}^{-tG(Q)},\ G(Q) := \text{spectral gap of } Q.$$

First consequence: quantitative convergence of the distribution of polynucleotides

For a box of N nucleotides (subadditivity of the TV distance),

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

This is the good behaviour as the construction of the dynamics by Poisson processes show.

Consequence: distance at most e^{-s} at time

$$t_N(s) = G(Q)^{-1}(\log(N) + s)$$

Second consequence: phylogenies

Assume sequences X' and X'' evolved during time t from a MRCA (unknown) sequence X. Estimating the time t elapsed between the ancestral sequence X and the present ones? MLE easy based on e^{tQ} and reversibility.

For Jukes-Cantor, the (normalised) elapsed time is

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

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

TOWARDS MODELS WITH DEPENDENCE

Back to (a tiny bit more of) reality: biologists tell us 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*

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: $(xyz) \approx (xy)\,(yz)/(y)$ (*Mean field*)

Note:
– Called in other contexts Bethe Ansatz, Kikuchi approximation, cluster approximation, etc.
– Exact for (spatial) Markov chains.

Assuming *mean field*, 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 *mean field*?

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 (a wide class of models called) 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.,

and properties (a')–(d') are based on the subcriticality of an underlying branching process.

(A FAIRY TALE BASED ON) POISSON DYNAMICS

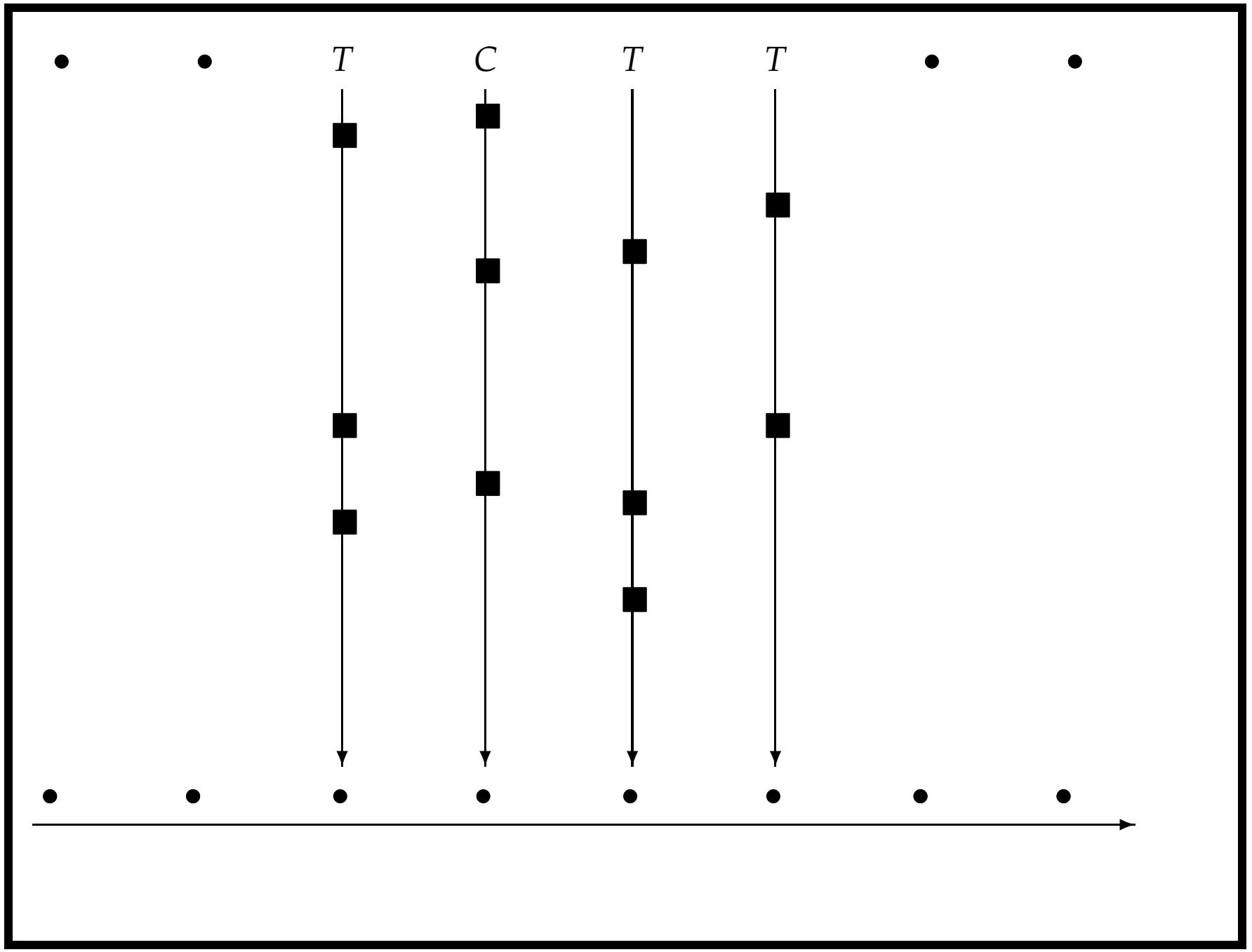
■ = "simple" substitutions.

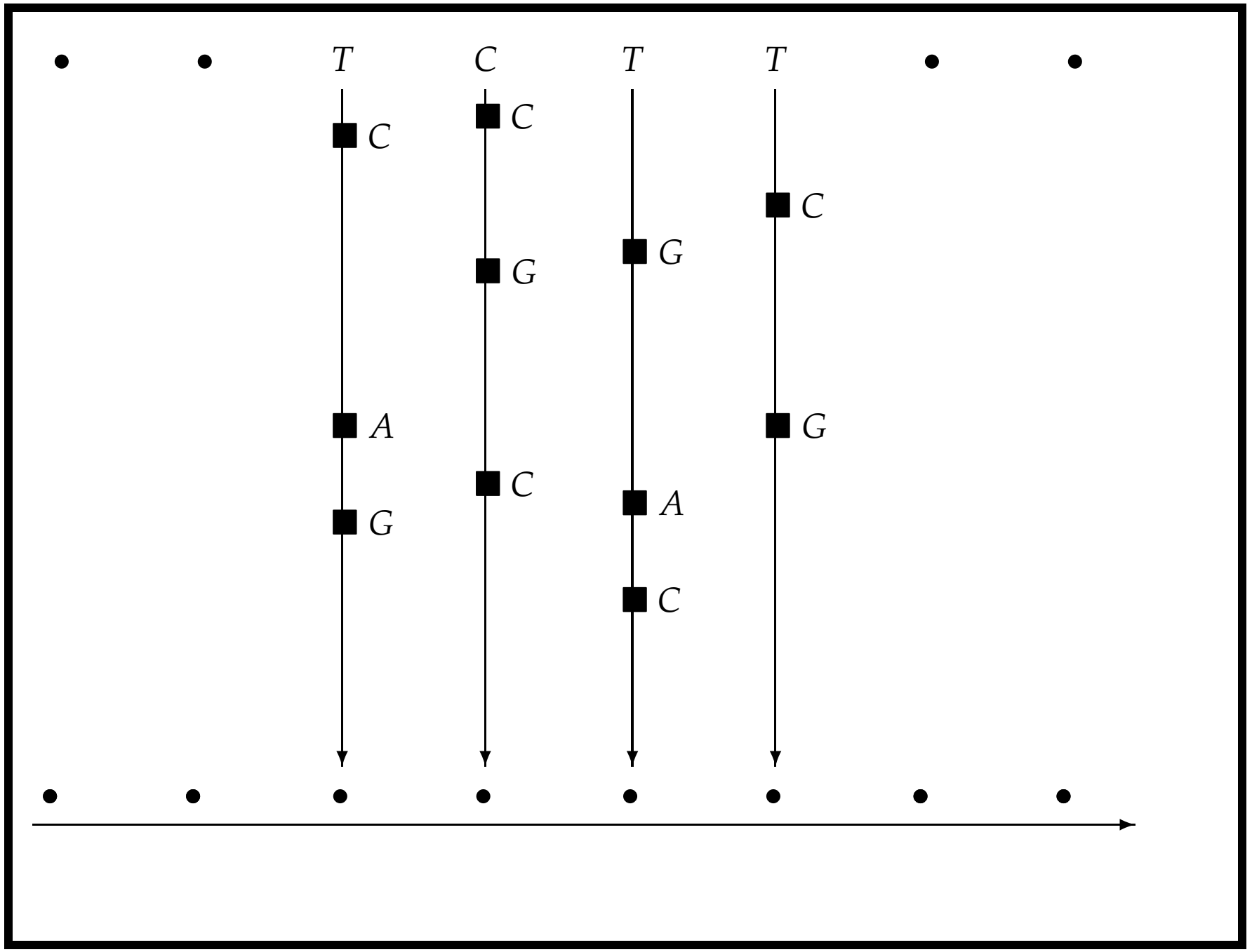
$\tau(x, y)$ = rate of substitution of x by y.

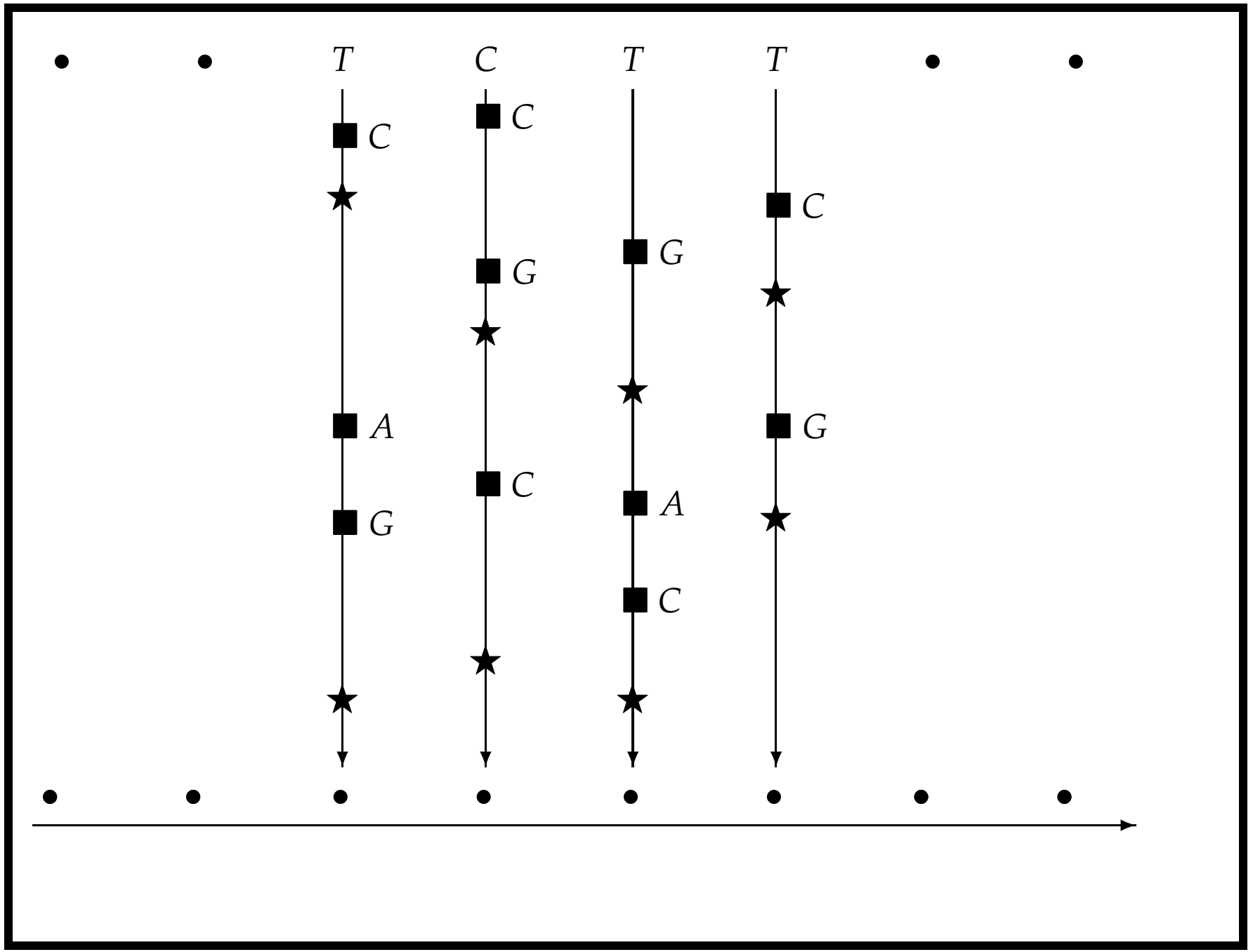
Caution: one authorizes simple (virtual) substitutions of x by x, hence every $\tau(x, x)$ is a free parameter.

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

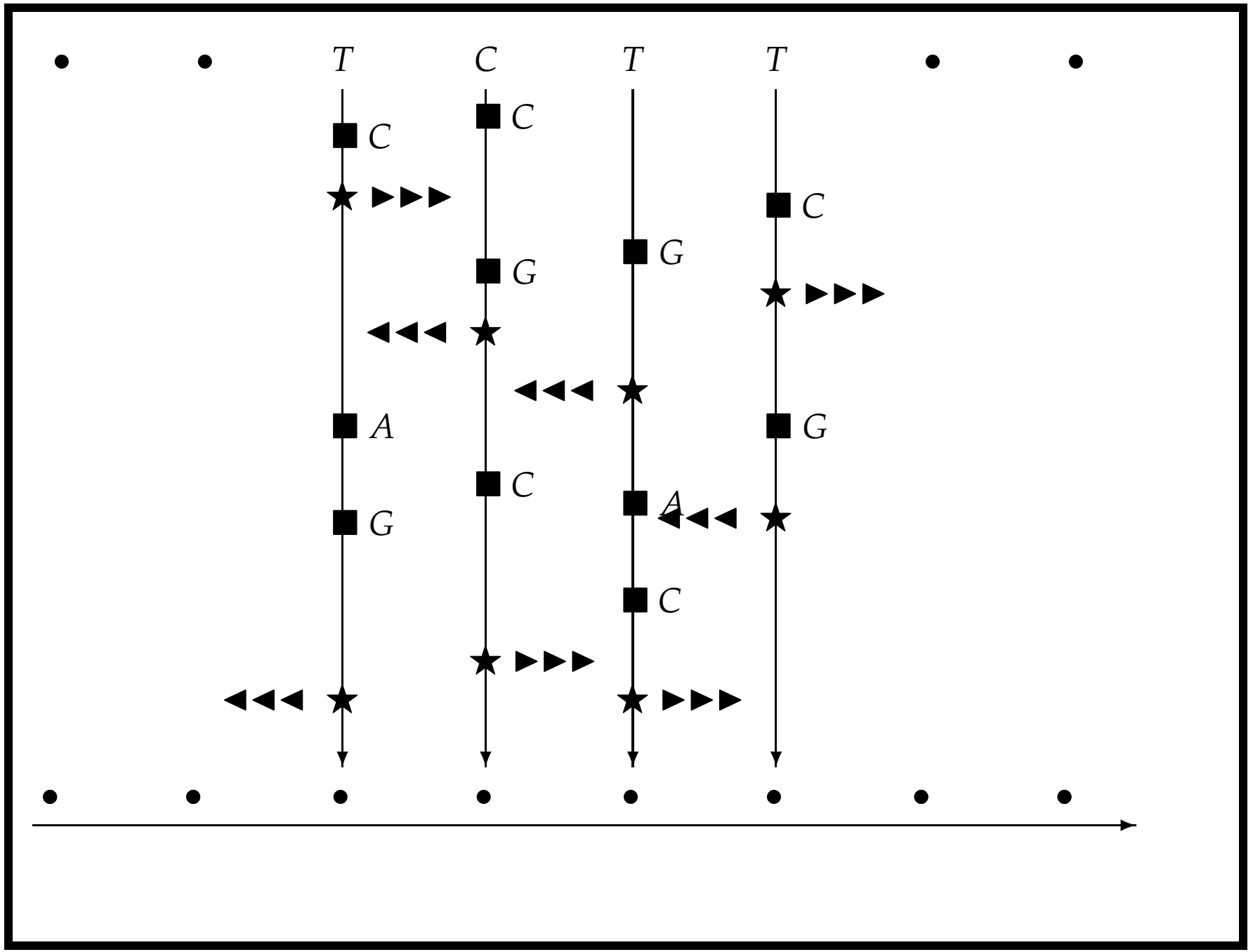
One wants to represent the **evolution of a finite collection of sites**.


T
C
T
T

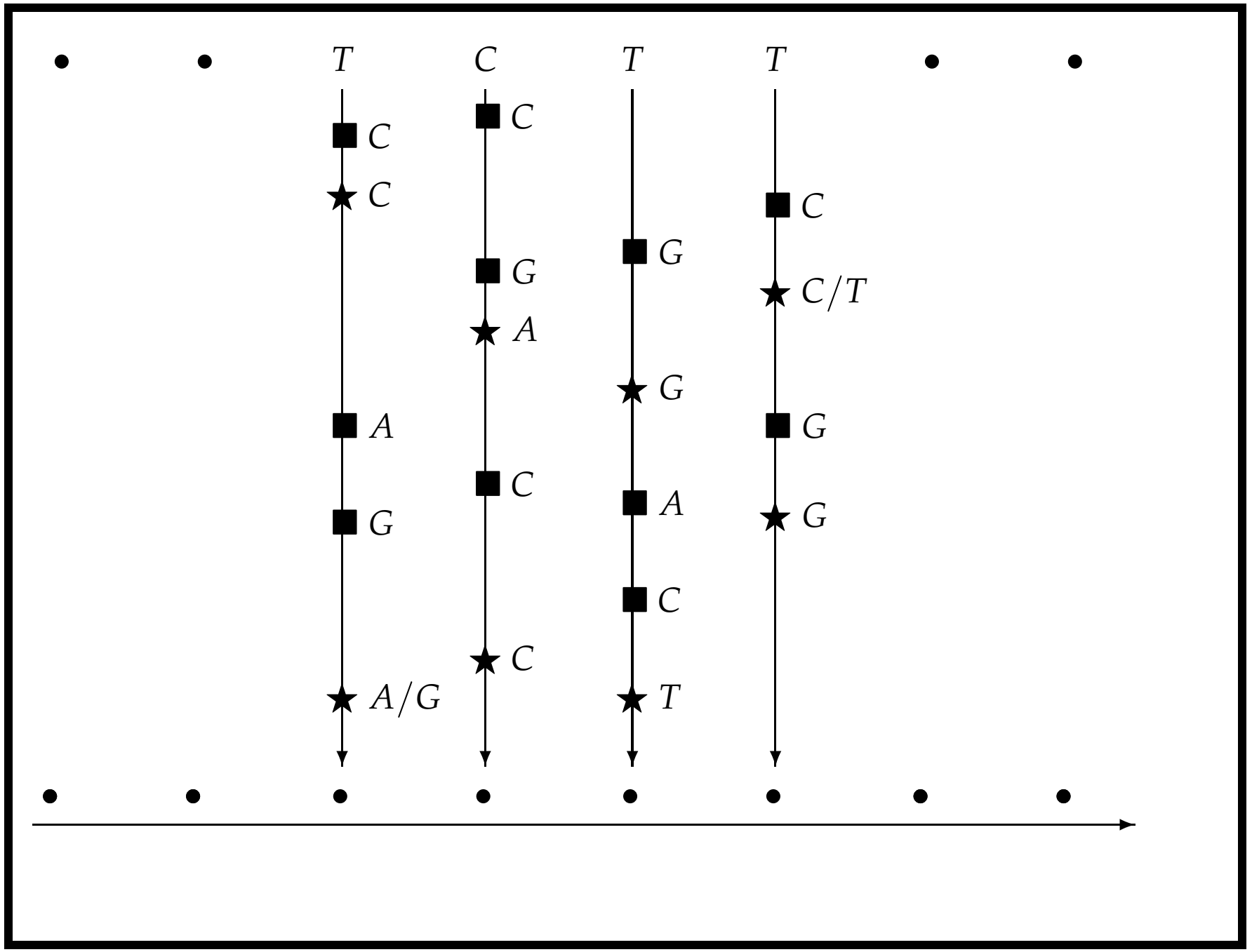

T
C
T
T
C
C
G
C
G
A
C
G
A
C

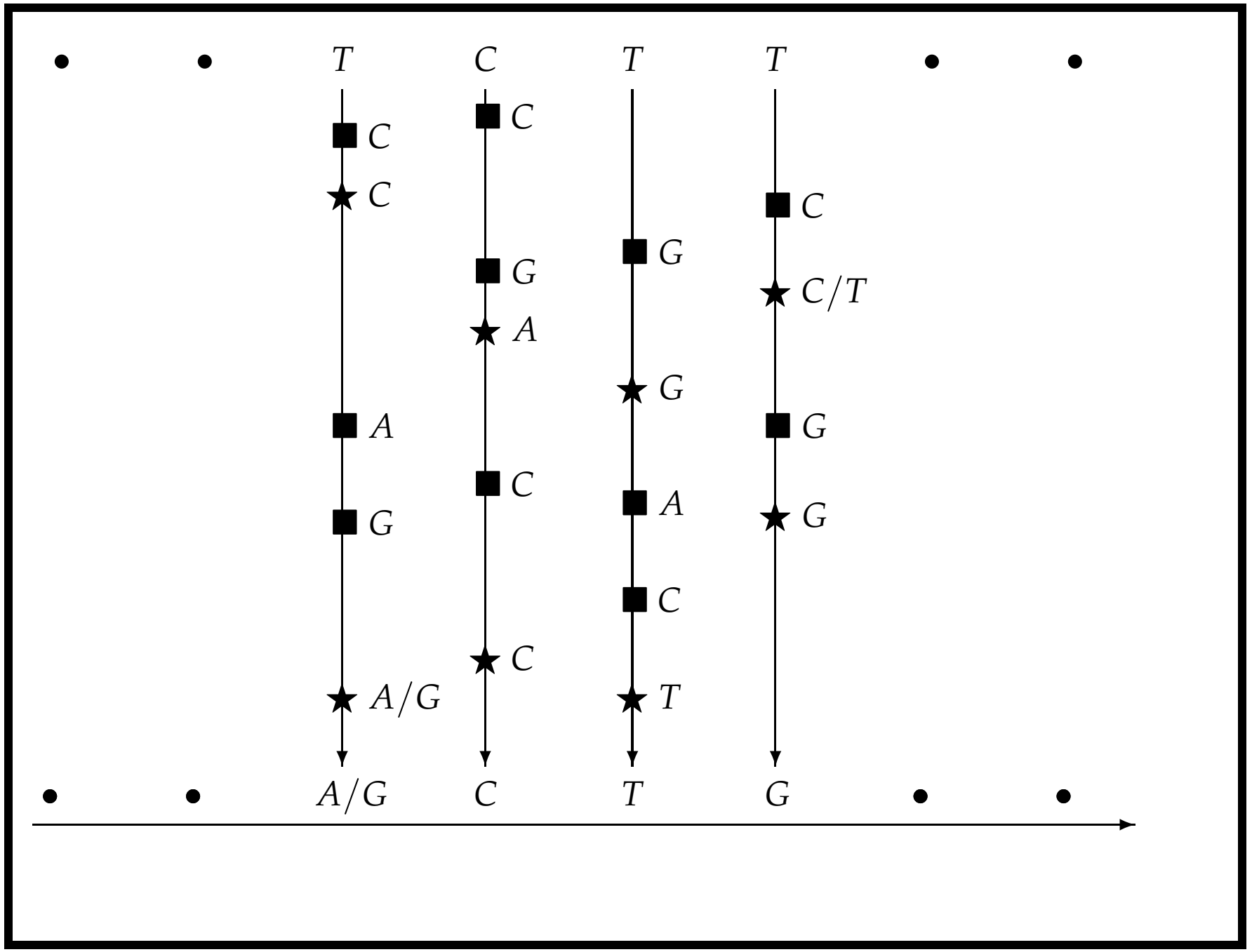

T
C
T
T
C
C
G
C
G
A
C
G
A
C

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.

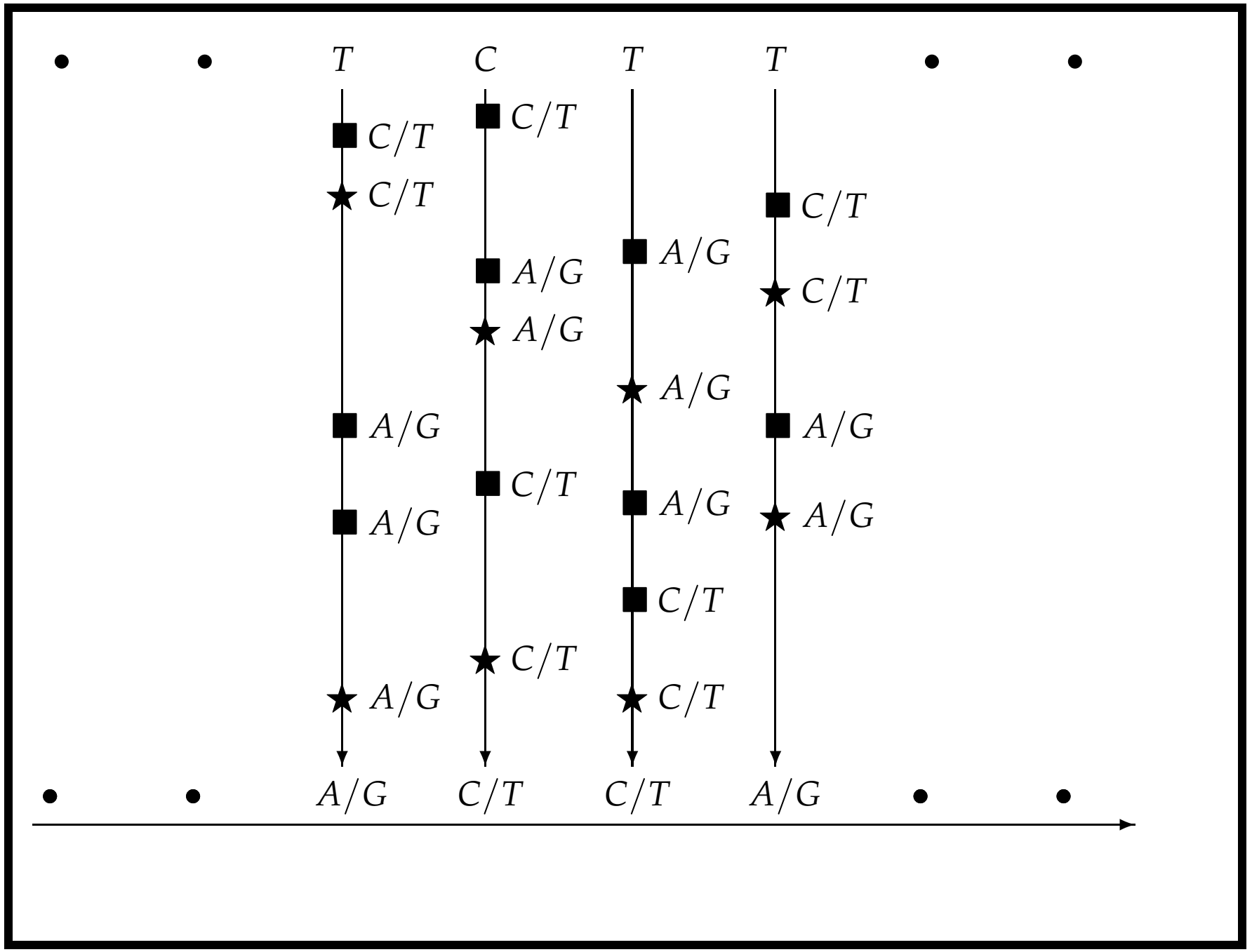

T
C
T
T
C
C
C
C
G
G
A
A
G
G
C
C
A
G
C
C
T
A/G
C/T
G
G

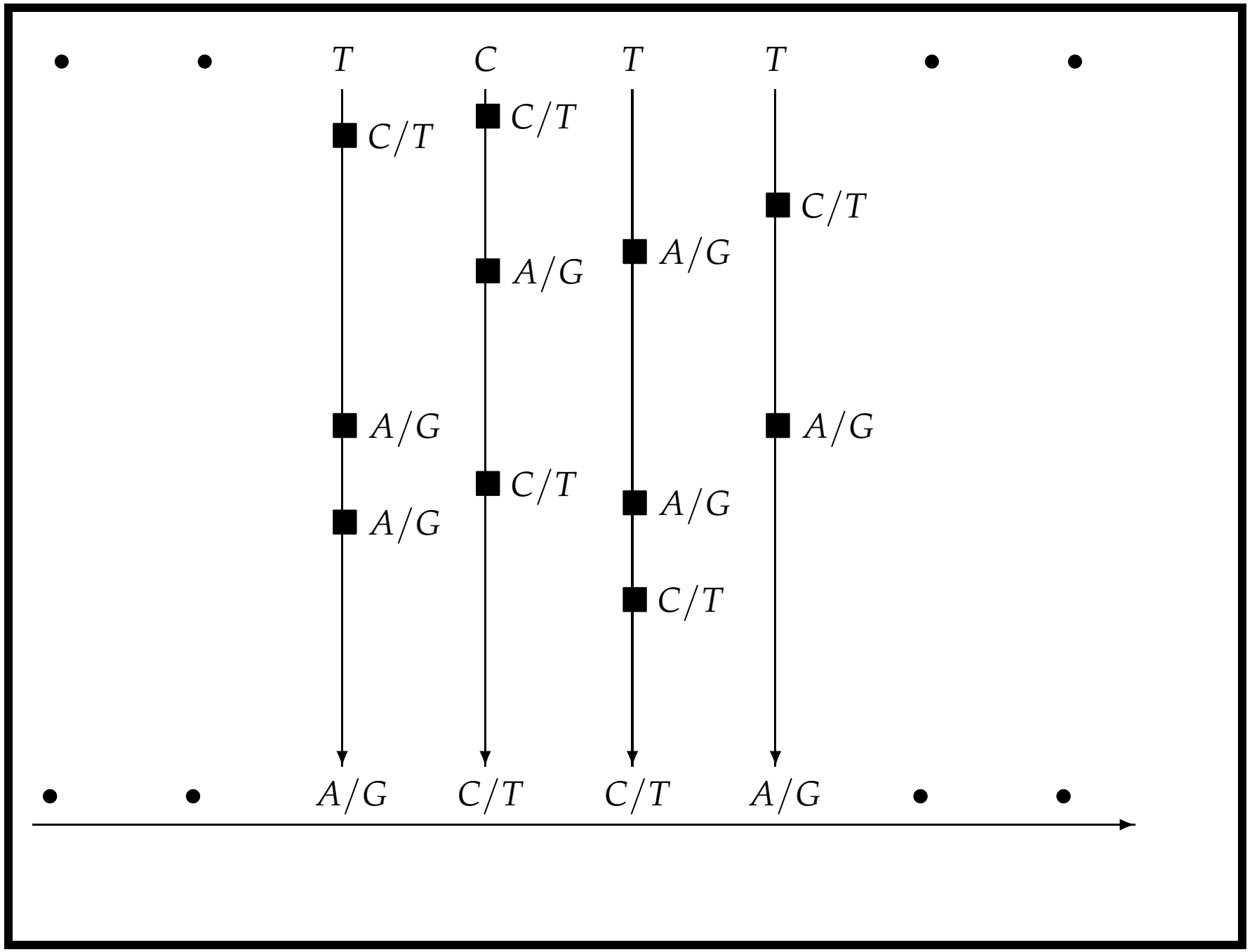

T
C
T
T
C
C
C
G
C/T
G
A
G
G
A
G
C
C
A/G
T
A/G
C
T
G

R-Y quotient

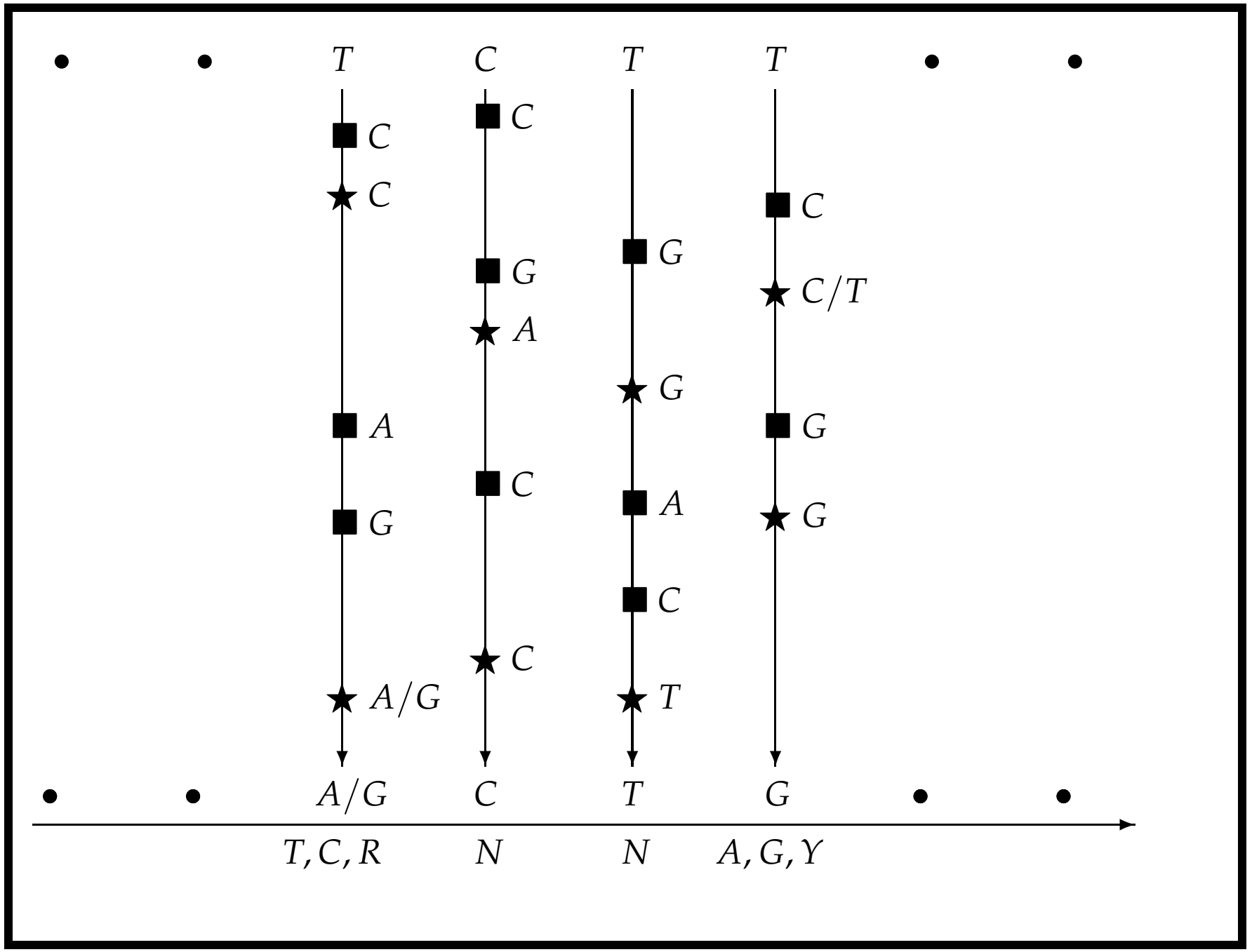
$R = \{G, A\} =$ purines, $Y = \{C, T\} =$ pyrimidines.

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

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

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

"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 hence the distribution of the sequence of purines/pyrimidines converges to a product of Bernoulli distributions.

Independence At stationarity, boxes at distance 3 are independent since there exists i.i.d. objects $(C_i)_i$ such that X_i at site i depends on (C_{i-1}, C_i, C_{i+1}) only.

Finer quotient For a box 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.

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 Dependent model but many things work as in the independent case, for instance the number of substitutions to perform to simulate an n box at stationarity like

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

And using this one can compute evolutionary distances and construct phylogenies based on RN+YpR dynamics

Trying to go out of Paradise (but not too far)

What about the behaviour of models not in the RN+YpR (measure zero) submanifold?

Take-home message In a neighborhood of RN+YpR, one can use (and modify) coupling-from-the-past ideas to prove ergodicity of the perturbed model, provided a branching process representing ambiguities of the dynamics dies out; additional information about the stationary measure and the dynamics then comes for free.

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.

"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 $(|\varepsilon|/\varepsilon_*)^{|x-y|/4}$.

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