Introduction to modelling sequence evolution

Bastien Boussau
Bastien.boussau@univ-lyon1.fr
@bastounette

LBBE
BIOMETRIE ET BIOLOGIE EVOLUTIVE
cnrs
Lyon 1

Aims

- Understand the main ideas underlying models of sequence evolution
- To do so, we will:
  - Simulate the evolution of a simple binary character through time
  - Extend to more general alphabets
  - Extend to longer sequences
  - Extend to a tree

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Example

- Is my coin fair?
- Repeated throws=independent *Bernoulli* draws
- Throw coin *N* times
- Estimate probability of heads

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Phylogeny example

- Are transitions as probable as transversions in rodents?
- Sites of alignment=independent Markov chains running along a phylogeny
- Sequence rodents
- Estimate transition/transversion ratio

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Phylogeny example

- Are transitions as probable as transversions in rodents?
- *Sites of alignment=independent Markov chains running along a phylogeny*
- Sequence rodents
- Estimate transition/transversion ratio

Why are we interested in simulations?

- Simulating data forces us to think in terms of a generating process
- By comparing true to simulated data, we can get a sense of how realistic is our model
- Simulations are also central to a lot of inferential problems:
  - Validation of inference methods
  - Posterior predictive tests
  - Approximate Bayesian Computation (ABC)
  - ...

Useful probability concepts

- Conditional probabilities
- Independence/intersection
- Union
- Bayes theorem
- Common distributions that will be useful in this talk:
  - Bernoulli
  - Binomial
  - Poisson
  - Exponential

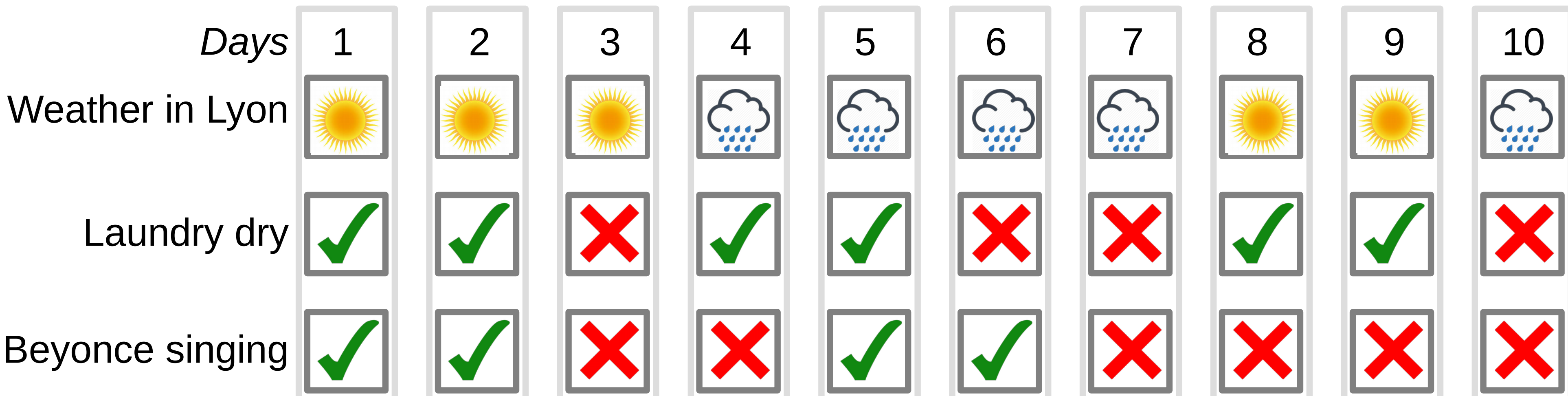
Crash course in probability

Record of various events during 10 days

Days	1	2	3	4	5	6	7	8	9	10
Weather in Lyon	☀	☀	☀	🌧	🌧	🌧	🌧	☀	☀	🌧
Laundry dry	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗
Beyonce singing	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗

Crash course in probability

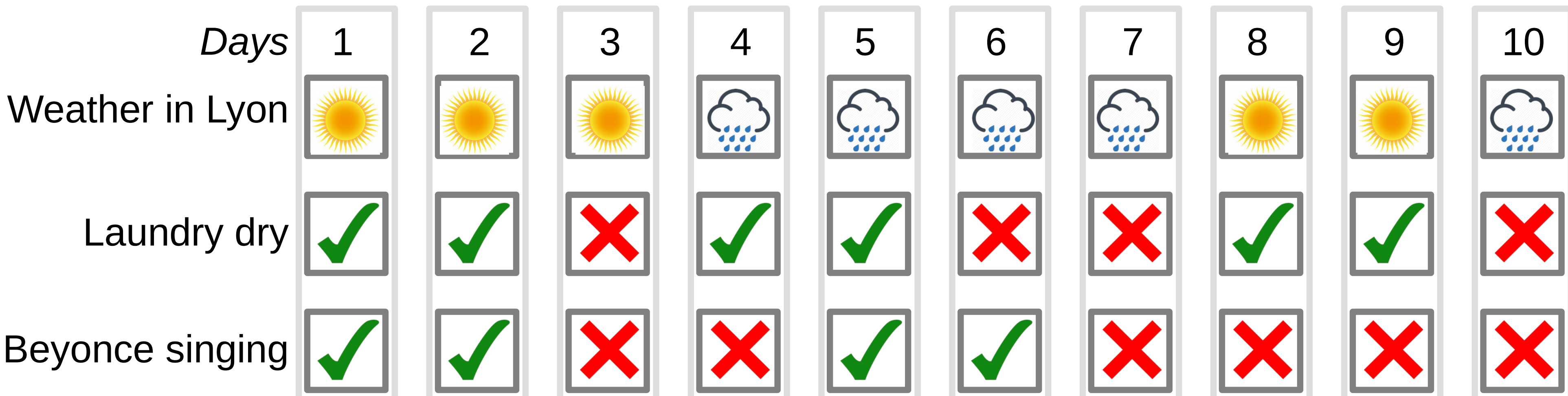
Record of various events during 10 days



$P(rainy) = 0.5 \qquad P(sunny) = 1 - P(rainy) = 0.5$

Crash course in probability

Record of various events during 10 days



$$P(rainy) = 0.5 \qquad P(sunny) = 1 - P(rainy) = 0.5 \qquad P(dry\ laundry) = 0.6$$

Crash course in probability

Record of various events during 10 days

Days	1	2	3	4	5	6	7	8	9	10
Weather in Lyon	☀	☀	☀	🌧	🌧	🌧	🌧	☀	☀	🌧
Laundry dry	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗
Beyonce singing	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗

$P(rainy)=0.5 \qquad P(sunny)=1-P(rainy)=0.5 \qquad P(dry\ laundry)=0.6$

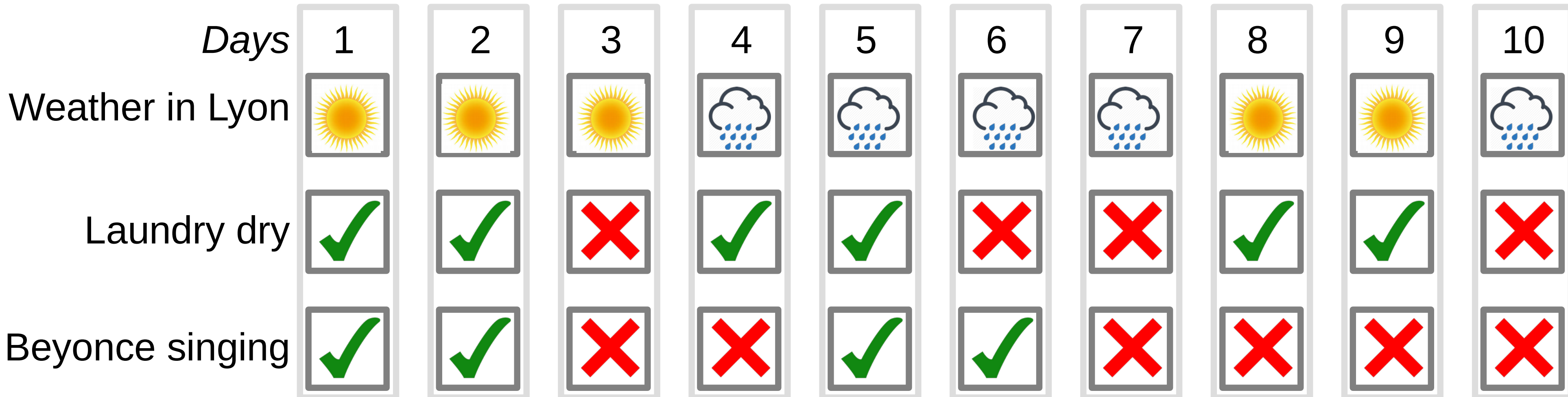
$P(dry\ laundry|sunny)=0.8$

$P(dry\ laundry|rainy)=0.4$

Conditional probability: P(A|B)

Crash course in probability

Record of various events during 10 days



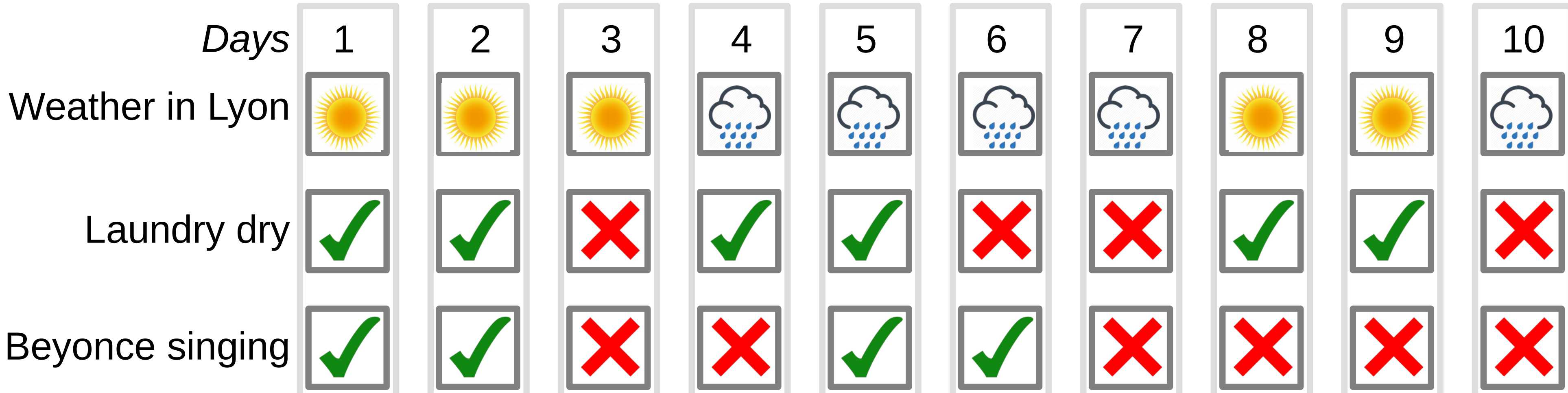
$P(rainy) = 0.5$ $\quad P(sunny) = 1 - P(rainy) = 0.5$ $\quad P(dry\ laundry) = 0.6$

$P(dry\ laundry | sunny) = 0.8$ $\quad P(dry\ laundry \wedge sunny) = 0.4$

$P(dry\ laundry | rainy) = 0.4$

Crash course in probability

Record of various events during 10 days



$P(rainy) = 0.5$ $\quad P(sunny) = 1 - P(rainy) = 0.5$ $\quad P(dry\ laundry) = 0.6$

$P(dry\ laundry | sunny) = 0.8$ $\quad P(dry\ laundry \wedge sunny) = 0.4$

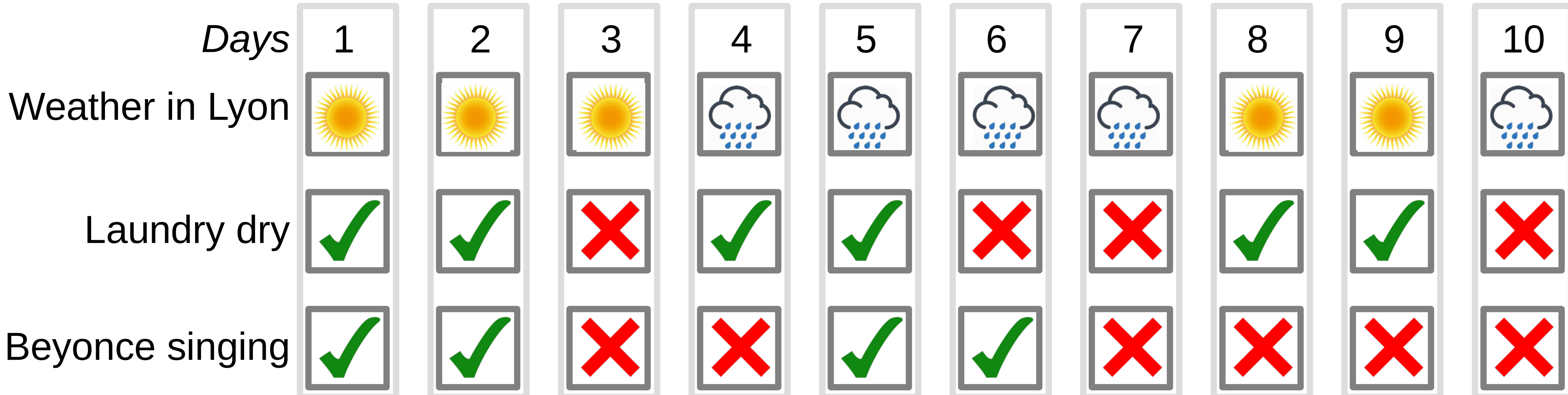
$P(dry\ laundry | rainy) = 0.4$

$P(Beyonce\ singing) = 0.4$

$P(Beyonce\ singing) = P(Beyonce\ singing | rainy) = P(Beyonce\ singing | sunny) = 0.4$

Crash course in probability

Record of various events during 10 days



$$P(rainy) = 0.5 \qquad P(sunny) = 1 - P(rainy) = 0.5$$

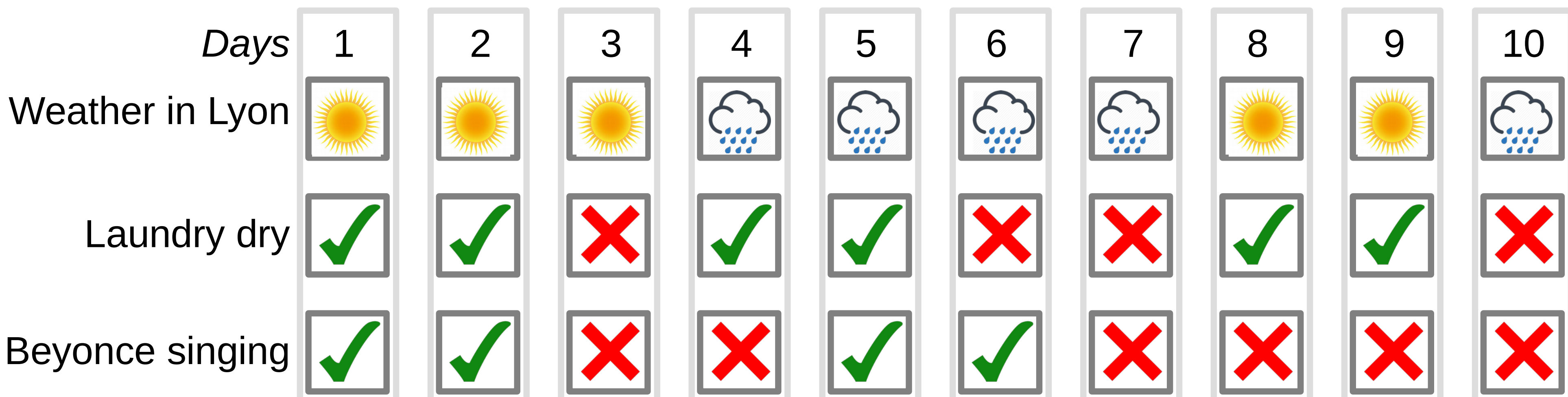
The events "Beyonce singing" and "sunny" are <u>independent</u>

$$P(Beyonce\ singing) = 0.4$$

$$P(Beyonce\ singing) = P(Beyonce\ singing | rainy) = P(Beyonce\ singing | sunny) = 0.4$$

Crash course in probability

Record of various events during 10 days



$P(rainy)=0.5 \qquad P(sunny)=1-P(rainy)=0.5 \qquad P(dry\ laundry)=0.6$

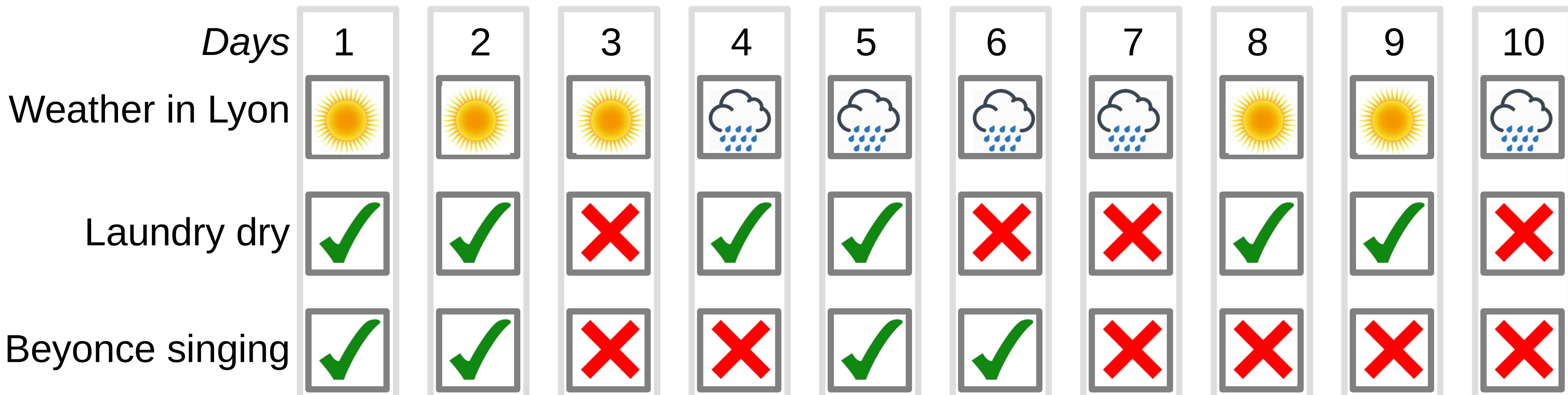
$P(dry\ laundry|sunny)=0.8$

$P(dry\ laundry|rainy)=0.4$

The events "dry laundry" and "sunny" are <u>NOT independent</u>

Crash course in probability

Record of various events during 10 days



$P(rainy)=0.5$ $\quad$ $P(sunny)=1-P(rainy)=0.5$ $\quad$ $P(dry\ laundry)=0.6$

$P(dry\ laundry|sunny)=0.8$ $\quad$ $P(dry\ laundry|rainy)=0.4$

$$\begin{aligned} P(dry\ laundry) &= P(dry\ laundry|sunny)\times P(sunny) \\ &\quad + P(dry\ laundry|rainy)\times P(rainy) \\ &= 0.8\times 0.5+0.4\times 0.5=0.6 \end{aligned}$$

Bayes formula

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B \wedge A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Days	1	2	3	4	5	6	7	8	9	10
Weather in Lyon	☀	☀	☀	🌧	🌧	🌧	🌧	☀	☀	🌧
Laundry dry	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗
Beyonce singing	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗

$P(rainy) = 0.5$ $\quad P(sunny) = 1 - P(rainy) = 0.5$ $\quad P(dry\ laundry) = 0.6$

$P(dry\ laundry|sunny) = 0.8$

$P(dry\ laundry|rainy) = 0.4$

$P(dry\ laundry \wedge sunny) = 0.4$

$$P(sunny|dry\ laundry) = \frac{P(sunny \wedge dry\ laundry)}{P(dry\ laundry)} = \frac{P(dry\ laundry \wedge sunny)}{P(dry\ laundry)}$$

$$P(sunny|dry\ laundry) = \frac{P(dry\ laundry|sunny)\,P(sunny)}{P(dry\ laundry)}$$

Useful distributions

- *Discrete distributions (values in {0,1}, {0,1,2...}):*
  - **Bernoulli**: coin flip: $P(X=1)=p\,;P(X=0)=1-p$
  - **Binomial**: how many heads in several coin flips:

$$Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

  - **Poisson**: how many events of a type over a continuous time: how many meteorites with diameter > 1m in a year:

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- *Continuous distributions (values in $\mathbb{R}$, [0,1]...):*
  - **Exponential**: Time between events in a Poisson process: how much time between two meteorites with diameter >1m:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Aims

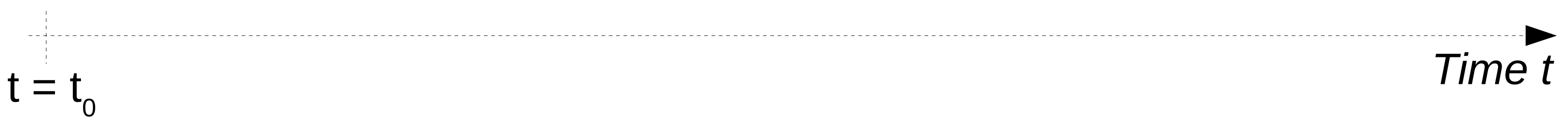
- Understand the main ideas underlying models of sequence evolution
- To do so, we will:
  - Simulate the evolution of a simple binary character through time
  - Extend to more general alphabets
  - Extend to longer sequences
  - Extend to a tree

How would we simulate the evolution of a binary character?

- 2 states: {0,1}

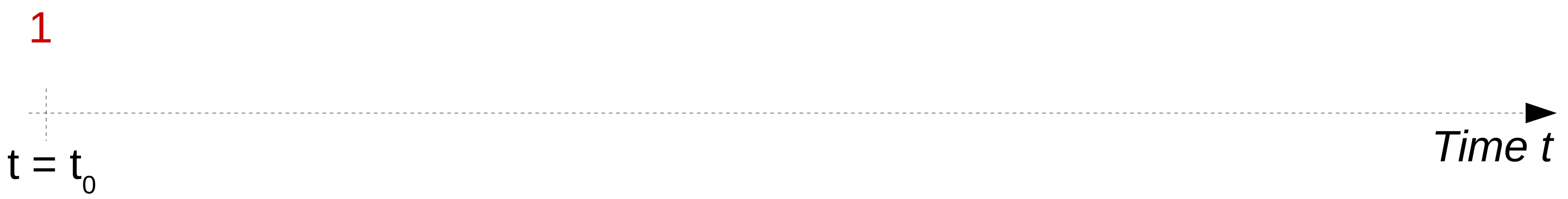
Evolution of a binary character

- 2 states: {0,1}



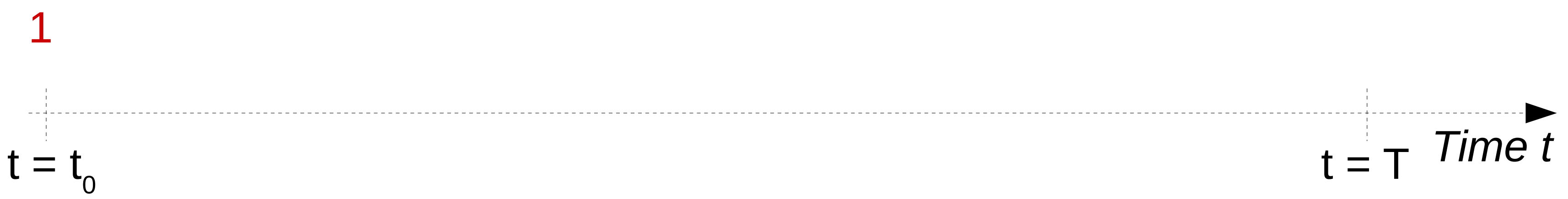
Evolution of a binary character

- 2 states: {0,1}



Evolution of a binary character

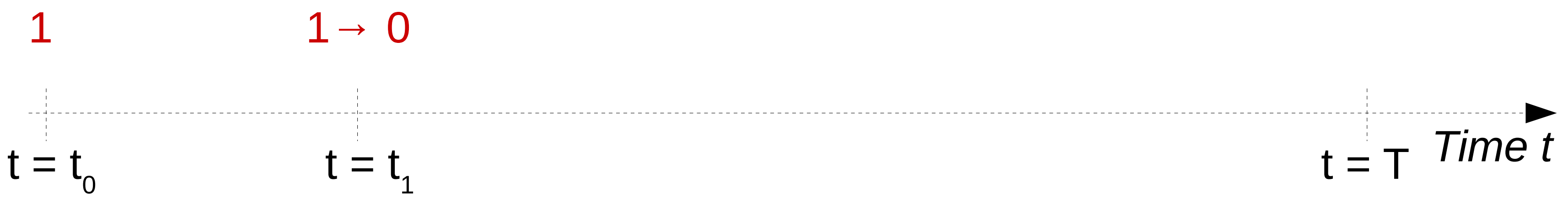
- 2 states: {0,1}



Number of substitutions at time t_0: $N(t_0) = 0$

Evolution of a binary character

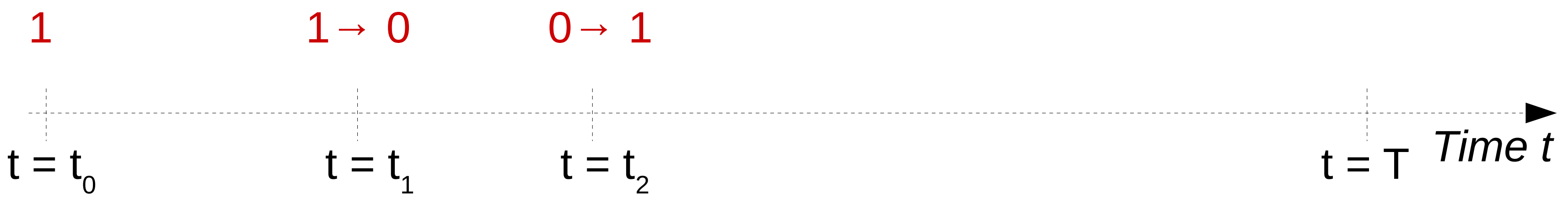
- 2 states: {0,1}



$$N(t_1) = 1$$

Evolution of a binary character

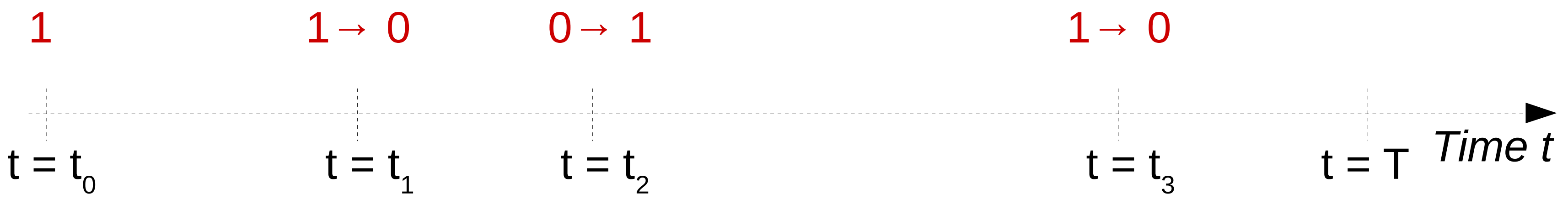
- 2 states: {0,1}



$$N(t_2) = 2$$

Evolution of a binary character

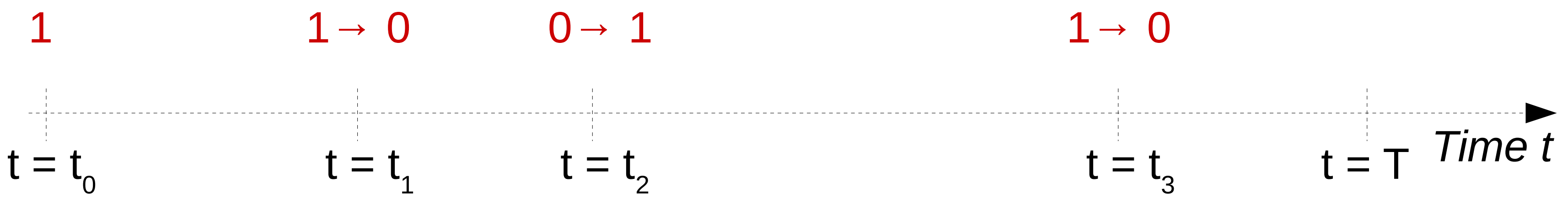
- 2 states: {0,1}



$$N(t_3) = 3$$

Evolution of a binary character

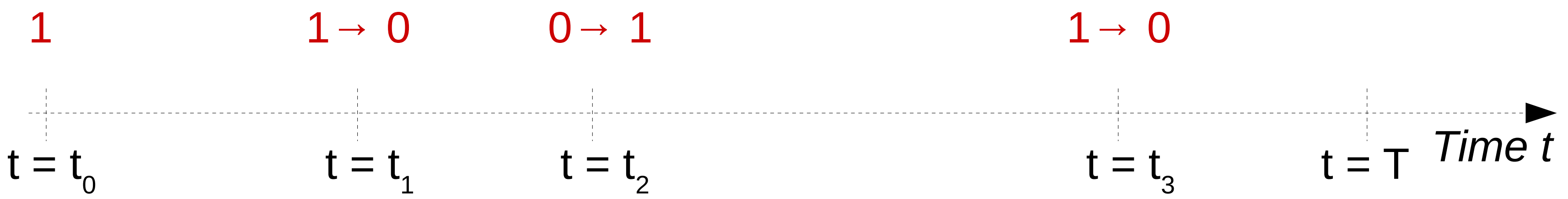
- 2 states: {0,1}



$N(T) = 3$

Evolution of a binary character

- 2 states: {0,1}

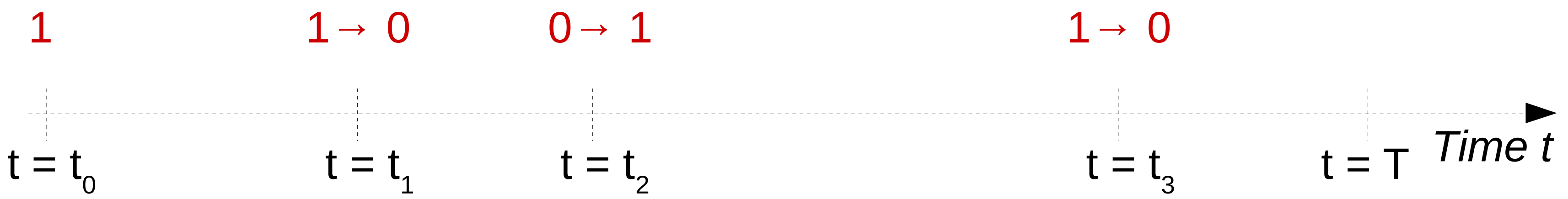


N(T) = 3

Could we simulate this process just with coin flips?

Evolution of a binary character

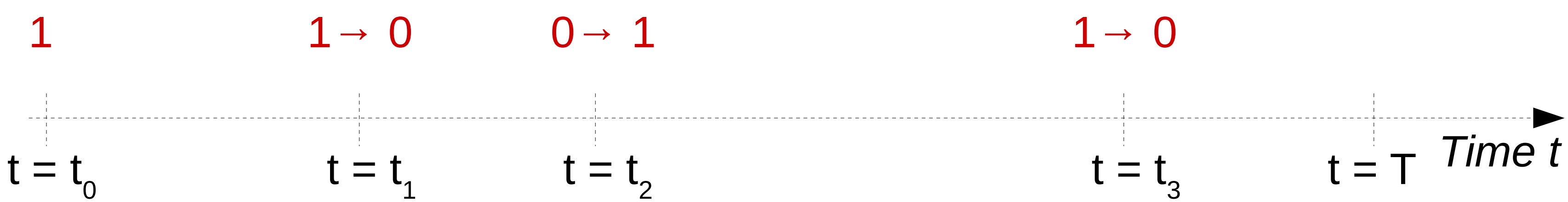
- 2 states: {0,1}



$N(T) = 3$

Evolution of a binary character

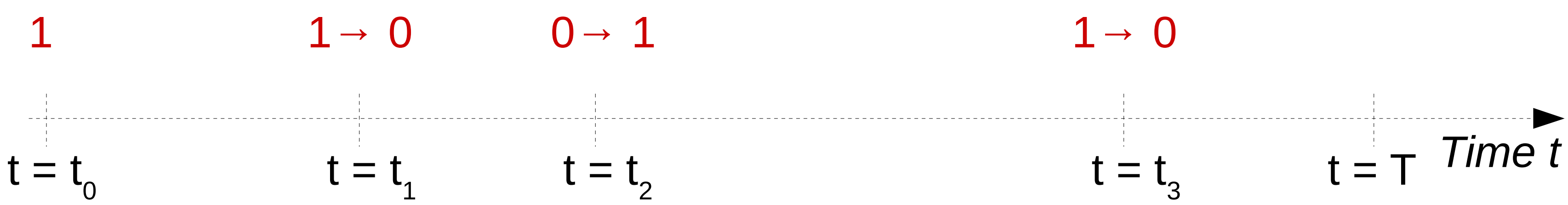
- 2 states: {0,1}



$N(T) = 3$

Evolution of a binary character

- 2 states: {0,1}



$N(T) = 3$

P(H) = ½

Evolution of a binary character

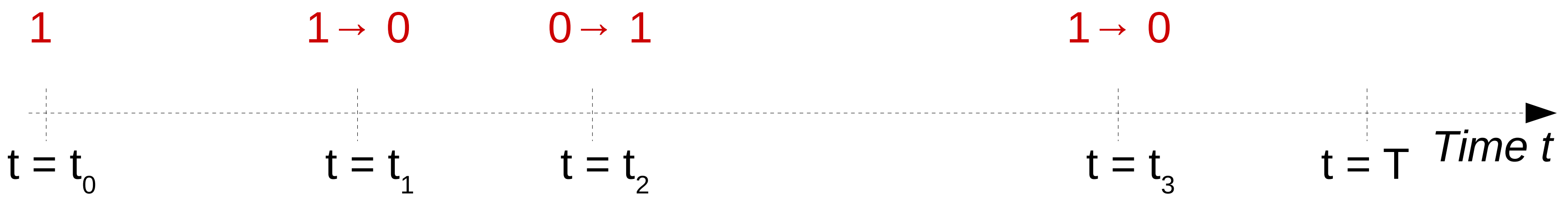
- 2 states: {0,1}

$N(T) = 3$

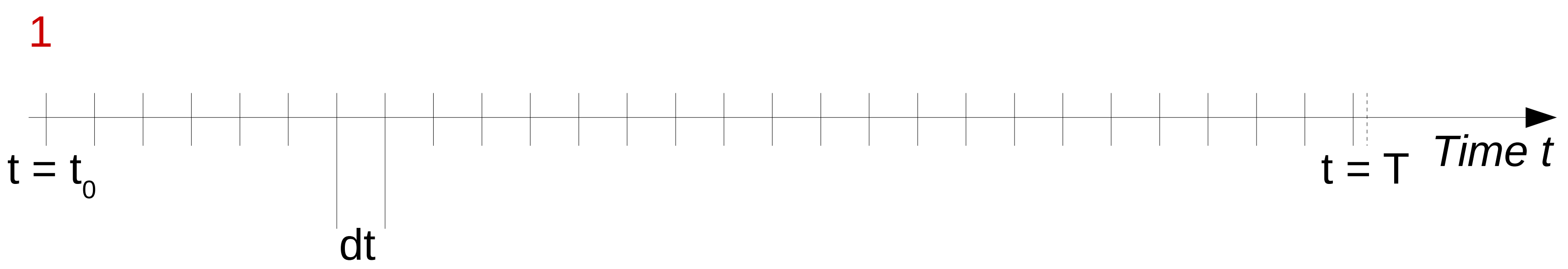
P(H) = ½

Evolution of a binary character

- 2 states: {0,1}

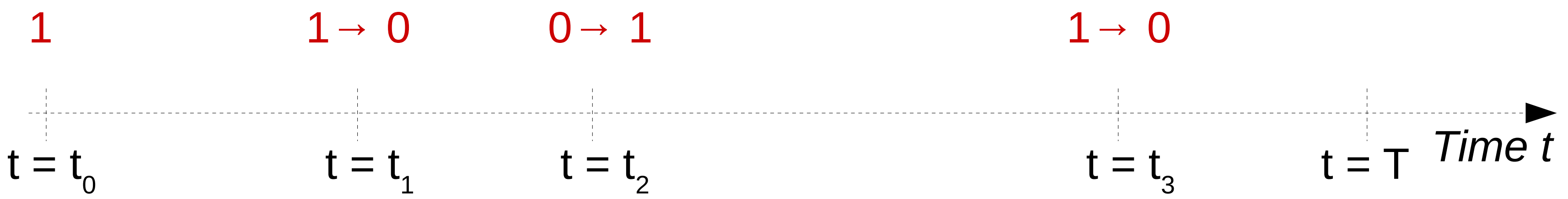


N(T) = 3

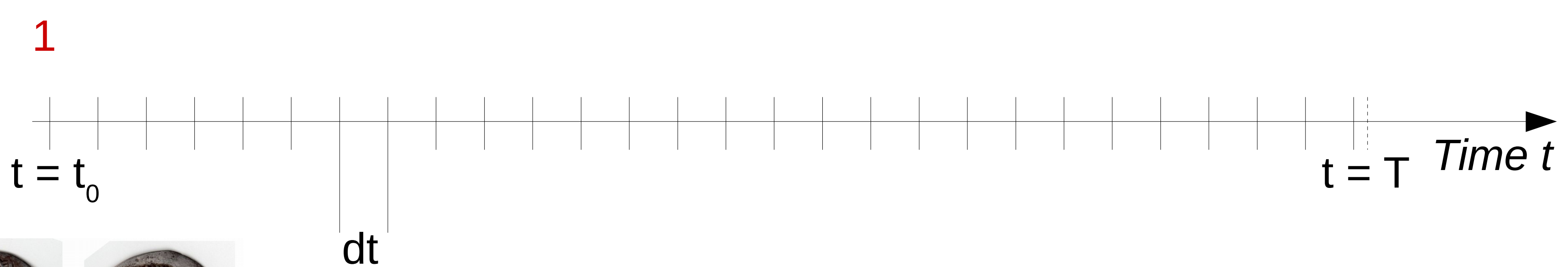


Evolution of a binary character

- 2 states: {0,1}

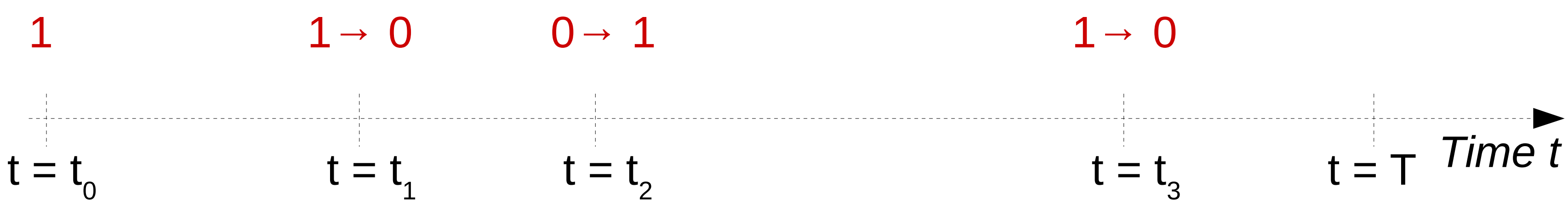


N(T) = 3

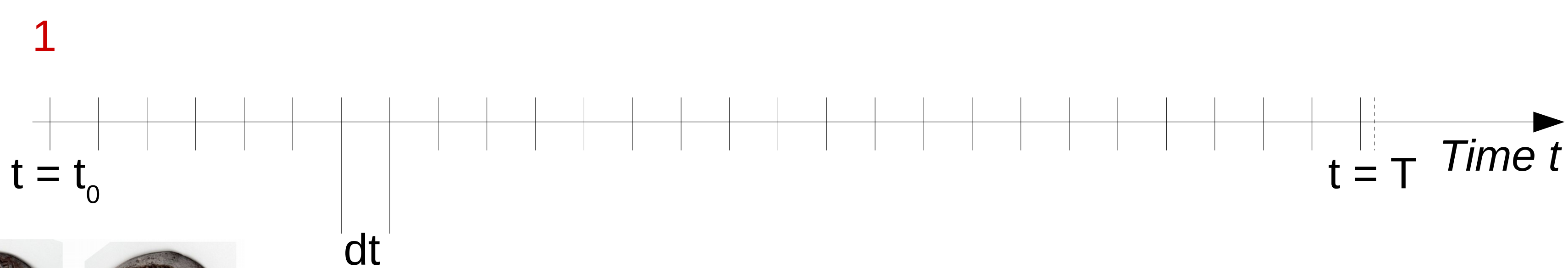


Evolution of a binary character

- 2 states: {0,1}



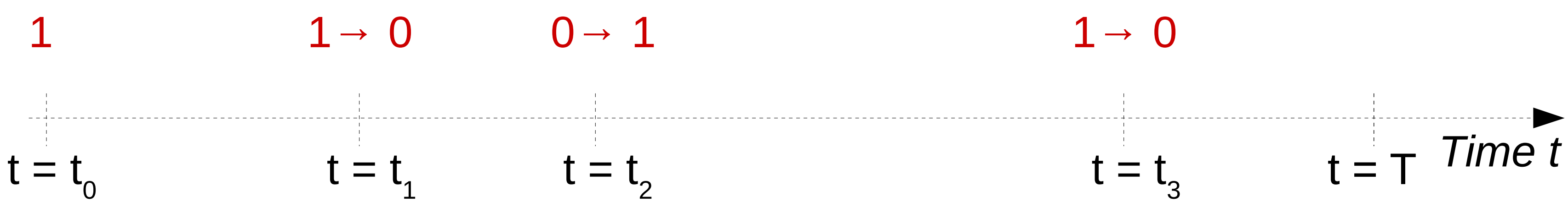
N(T) = 3



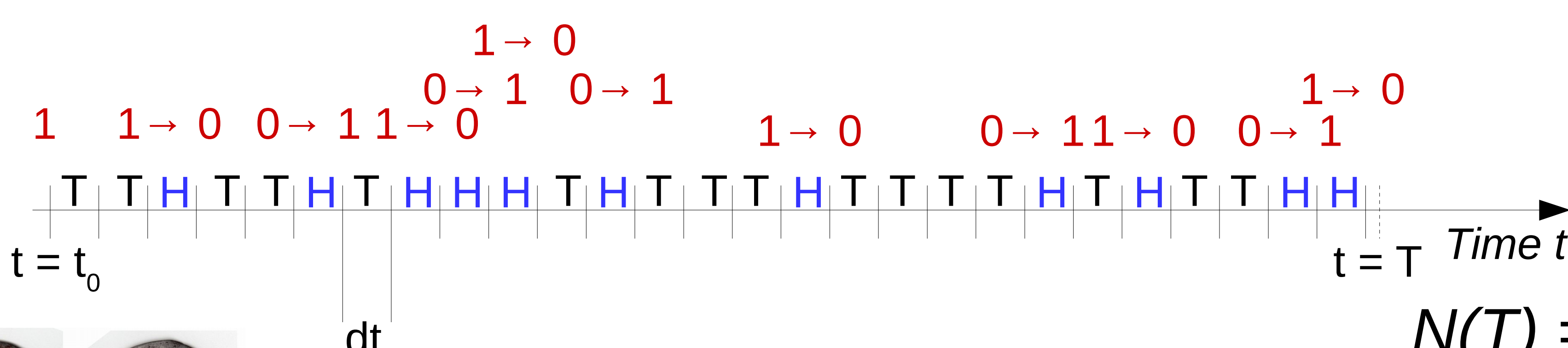
P(H) = ½

Evolution of a binary character

- 2 states: {0,1}



N(T) = 3

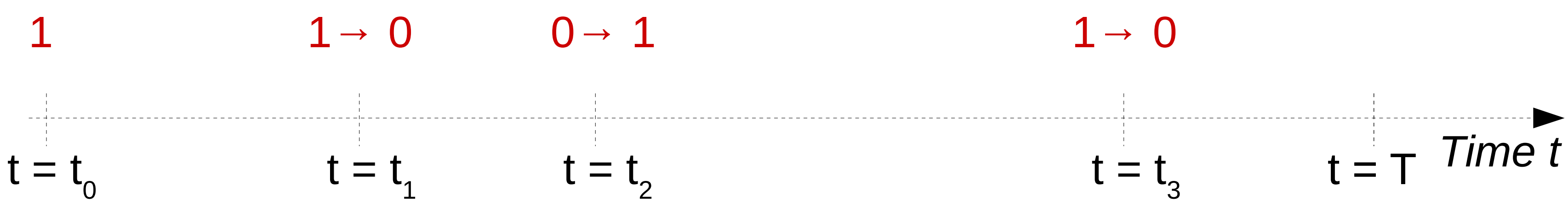


N(T) = 11

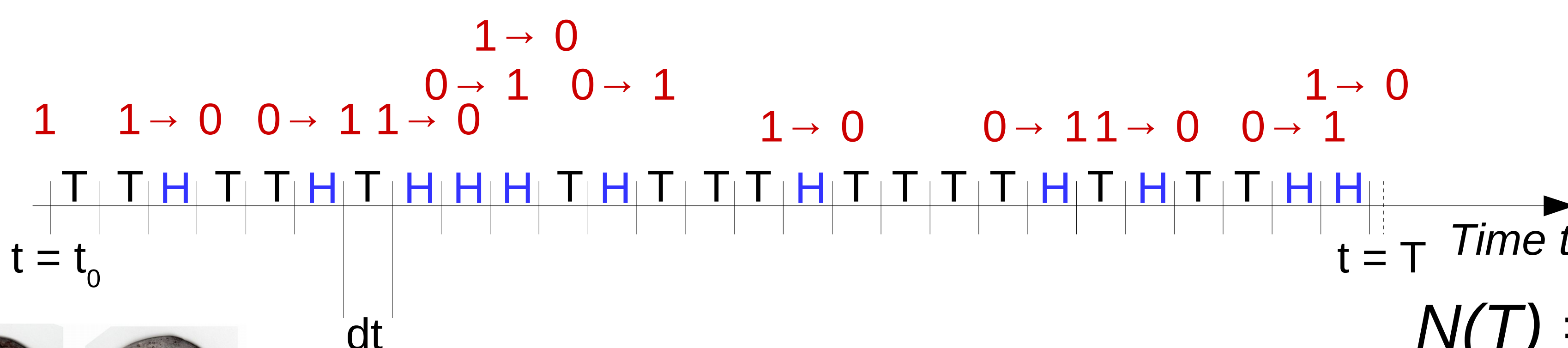
P(H) = ½

Evolution of a binary character

- 2 states: {0,1}



N(T) = 3



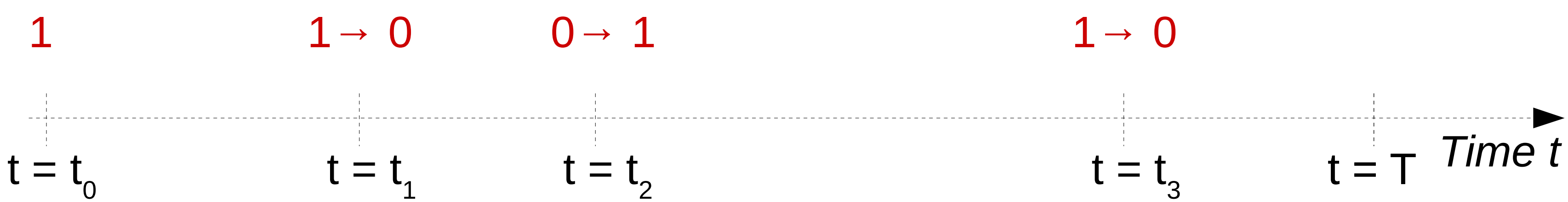
N(T) = 11

P(H) = ½

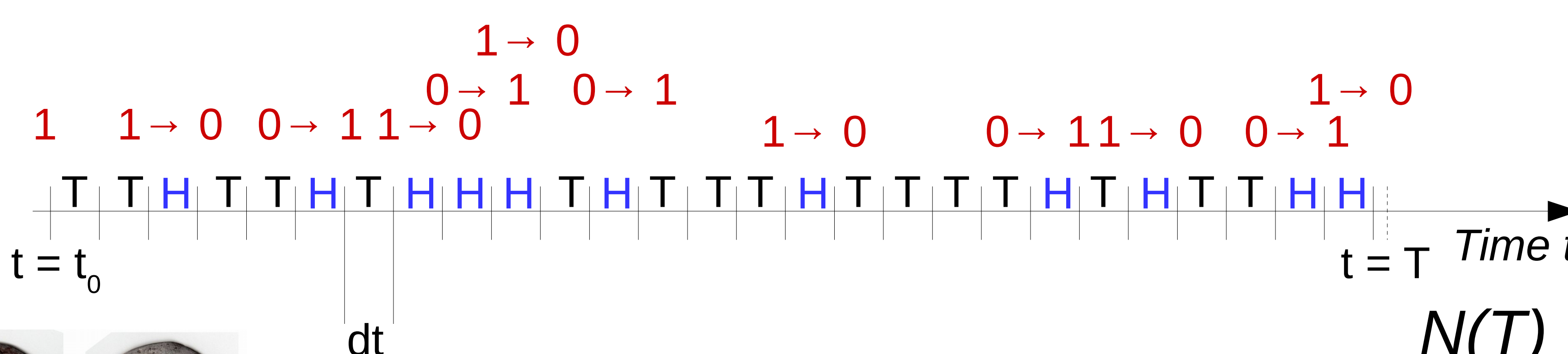
Is our model realistic?

Evolution of a binary character

- 2 states: {0,1}



N(T) = 3

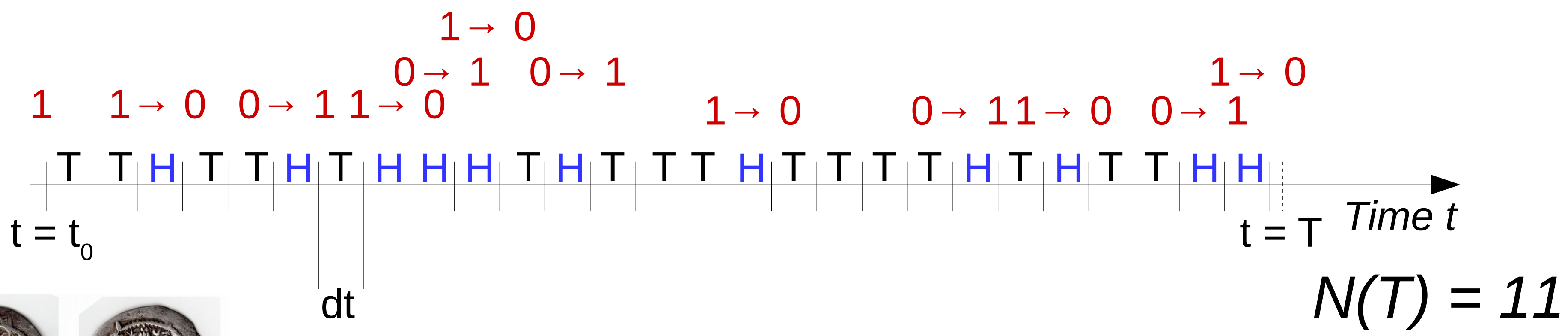


N(T) = 11

P(H) = ~~½~~ **p**

Is our model realistic?

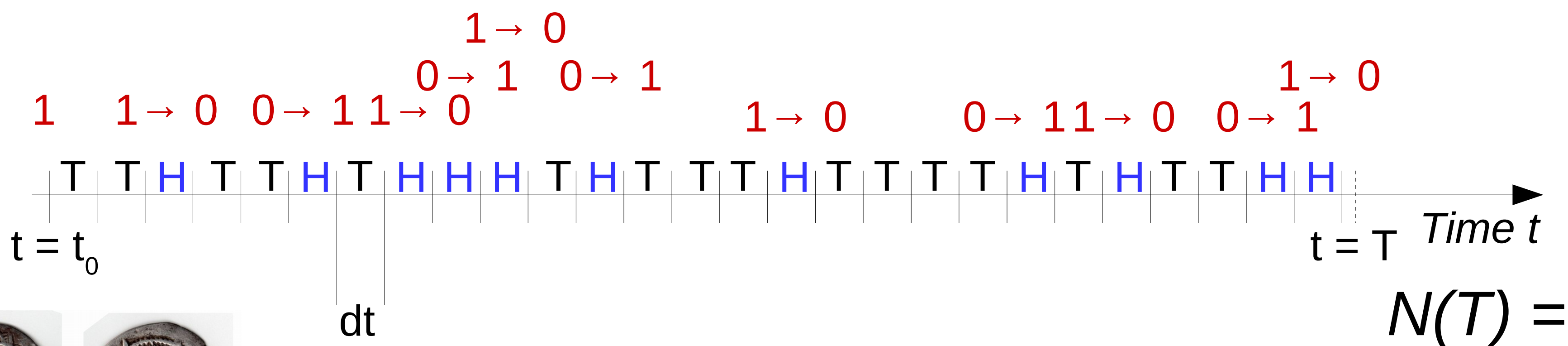
Our model so far



P(H) = p

- Discretization of time into n short intervals of length dt
- Initial state: draw from a Bernoulli(p)
- Substitutions: in each interval, draw from a Bernoulli(p)

Our model so far



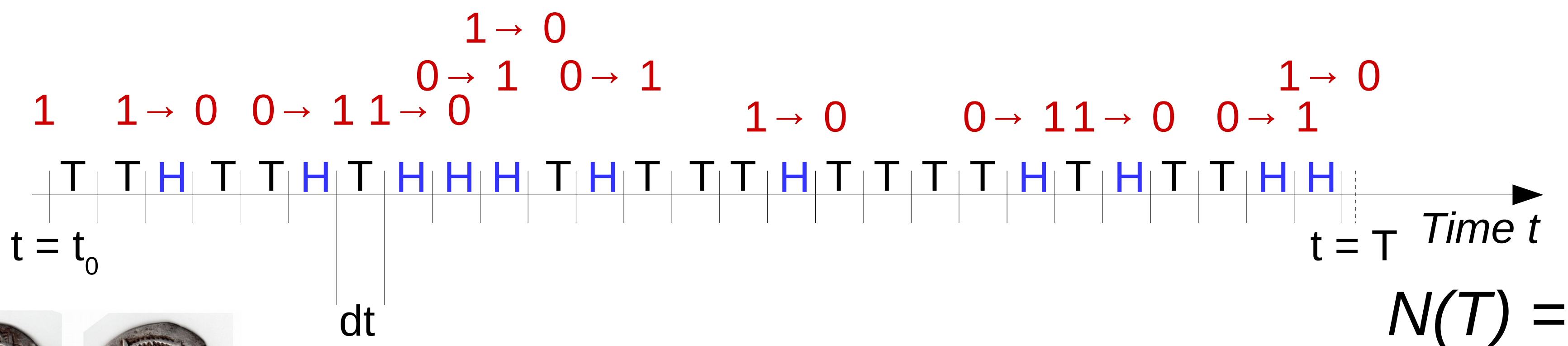
$N(T) = 11$

P(H) = p

- Discretization of time into n short intervals of length dt
- Initial state: draw from a Bernoulli(p)
- Substitutions: in each interval, draw from a Bernoulli(p)

What is the distribution of N(T)?

Our model so far



N(T) = 11

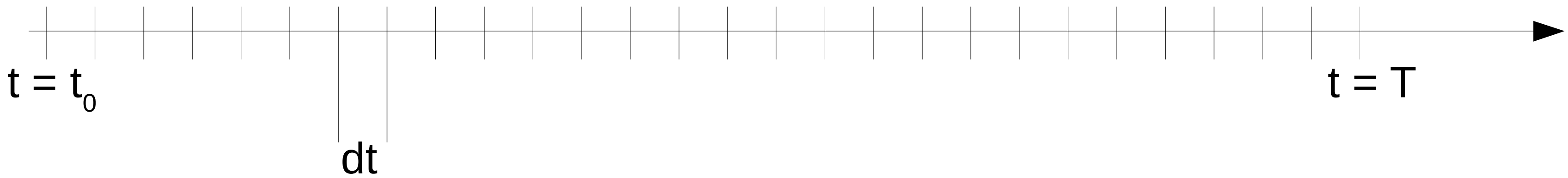
P(H) = p

- Discretization of time into n short intervals of length dt
- Initial state: draw from a Bernoulli(p)
- Substitutions: in each interval, draw from a Bernoulli(p)

N(T)~Binomial(n, p)

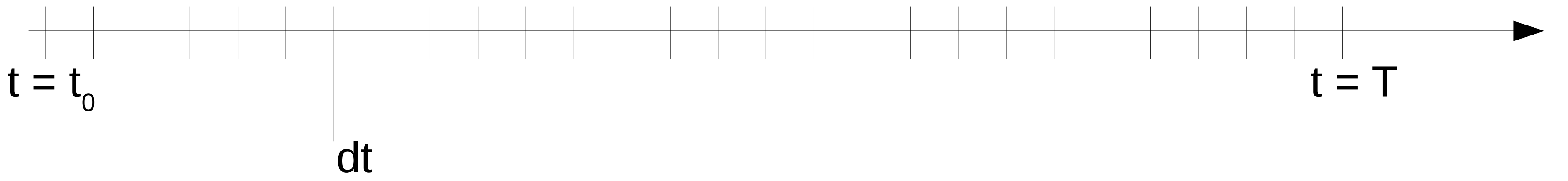
$$P(N=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

From a discrete to a continuous model



If *dt* → *0* or, equivalently, $n \to \infty$, towards what distribution tends *N(t)* ?

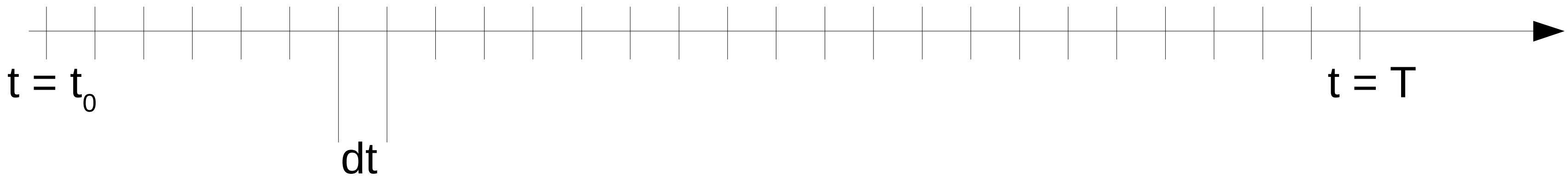
From a discrete to a continuous model



If *dt* → *0* or, equivalently, $n \to \infty$, towards what distribution tends *N(t)* ?

$$\lim_{n \to \infty} Binomial(n, p) = Poisson(np)$$

From a discrete to a continuous model



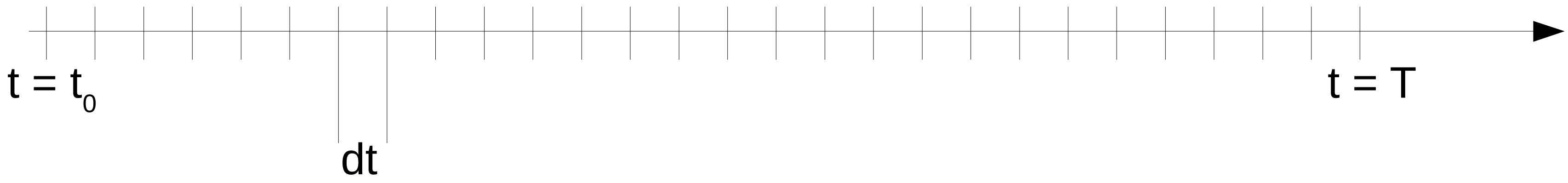
If *dt* → *0* or, equivalently, $n \to \infty$, towards what distribution tends *N(t)* ?

$$\lim_{n \to \infty} Binomial(n, p) = Poisson(np)$$

$$Si\ p = \lambda\, dt$$

$$\lim_{n \to \infty} Binomial(n, \lambda\, dt) = Poisson(\lambda\, t)$$

From a discrete to a continuous model



If *dt* → *0* or, equivalently, $n \to \infty$, towards what distribution tends *N(t)* ?

$$\lim_{n \to \infty} Binomial(n, p) = Poisson(np)$$

$$Si\ p = \lambda\, dt$$

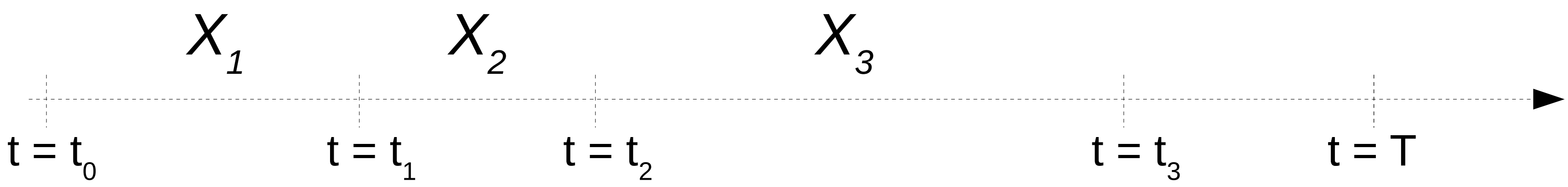
$$\lim_{n \to \infty} Binomial(n, \lambda\, dt) = Poisson(\lambda\, t)$$

$$Poisson_{\lambda}(k\ substitutions\ during\ \tau) = \frac{(\lambda\,\tau)^k e^{-\lambda\,\tau}}{k!}$$

The Poisson process

- Let $\lambda>0$. The counting process $\{N(t),t\in[0,\infty)\}$ is a Poisson process of rate λ if all the following conditions apply:
  - $N(0)=0$;
  - $N(t)$ has independent increments;
  - The number of events in any interval $\tau>0$ has distribution $Poisson(\lambda\tau)$.

Waiting times in Poisson processes

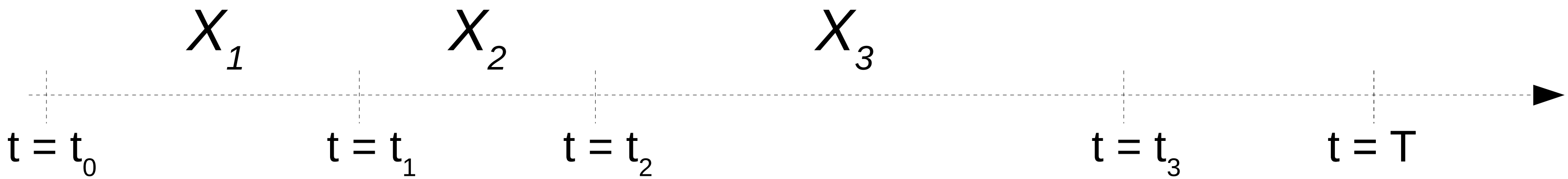


Waiting time:

- Time between the beginning of the process and the first event
- Time between 2 events.

Let's call a waiting time *X*.

Waiting times in Poisson processes



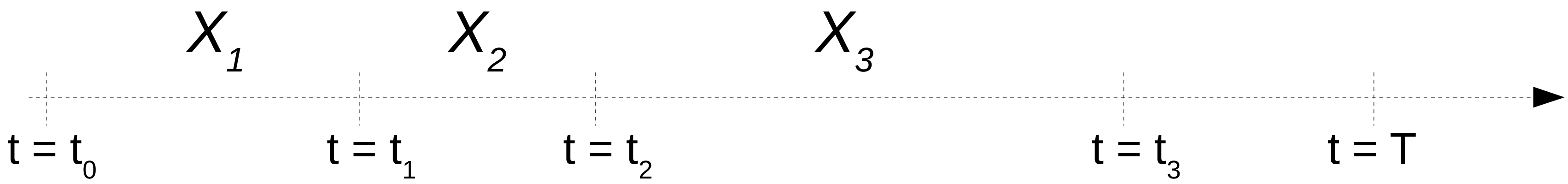
$$P(X > t) = P(no\ event\ during\ t) = \frac{(\lambda t)^0 \mathrm{e}^{-\lambda t}}{0!}$$

$$P(X > t) = \mathrm{e}^{-\lambda t}$$

$$F(X) = \begin{cases} 1 - e^{-\lambda t} \ldots if\ t > 0 \\ 0 \ldots otherwise \end{cases}$$

$$X \sim Exponential(\lambda)$$

Waiting times in a Poisson process



The X_i variables are the waiting times between events. They are all independent and follow the same distribution :

$$X_i \sim Exponential(\lambda)$$

Modelling the evolution of a binary trait: summary

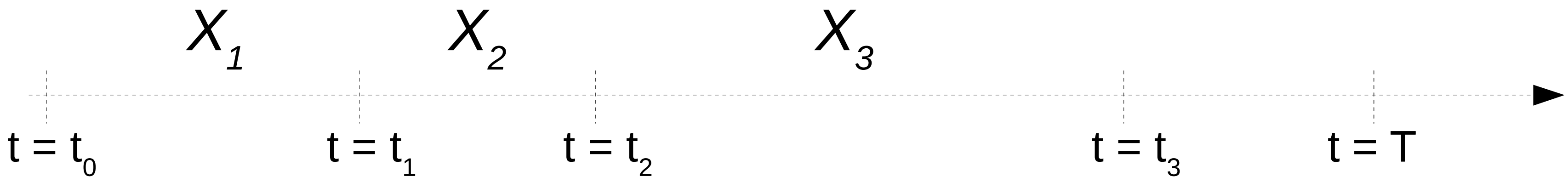
- We can simulate its evolution by repeating n Bernoulli draws during time intervals *dt*
- *N(t)* follows a Binomial distribution
- When *dt* becomes very small, *N(t)* follows a Poisson distribution of parameter $\lambda = n.dt$
- Waiting times between events follow an exponential distribution of same rate parameter λ

Modelling the evolution of a binary trait: summary

- We can simulate its evolution by repeating n Bernoulli draws during time intervals *dt*
- *N(t)* follows a Binomial distribution
- When *dt* becomes very small, *N(t)* follows a Poisson distribution of parameter $\lambda = n.dt$
- Waiting times between events follow an exponential distribution of same rate parameter λ

Can you think of a way to simulate the evolution of a binary trait in continuous time?

Modelling the evolution of a binary trait in continuous time

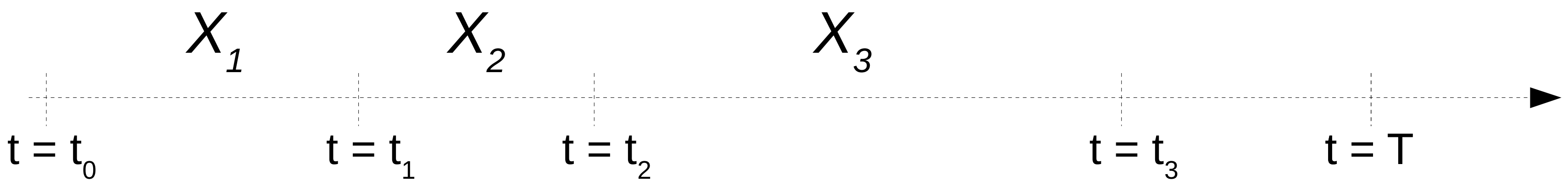


- Draw an initial state from a Bernoulli distribution:

```
p=0.3; state=rbinom(1,1,p)
```

- $t = t_0$; $N = 0$; $\lambda=0.1$
- While $t < T$:
  - Draw from an exponential distribution a waiting time X_i until the next event; $t = t + X_i$
    - If $t < T$, change the state of the variable
    - (Else ($t \geq T$): we stop)

Our model of DNA evolution in continuous time is a *Markov process*

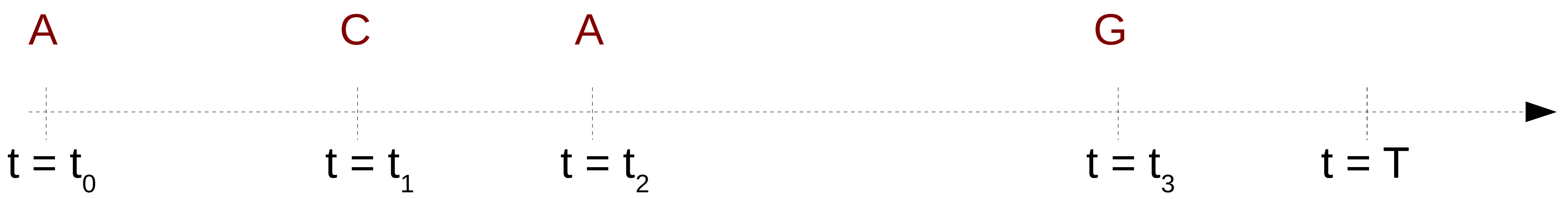


- At any given time, the next state only depends on the current state, not on the previous states.
- Therefore, we have defined a *Markov chain*.

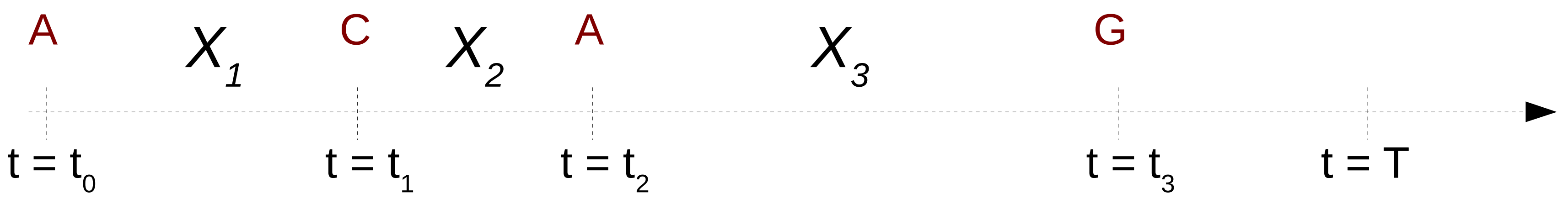
R function to simulate the evolution of a binary trait

```
simulate <- function (T, p, lambda) {
  N = 0
  t = 0.0
  state = rbinom(1,1,p)
  states = c(state)
  waitingTimes = c()
  while (t < T) {
    X = rexp(n=1, lambda)
    t = t+X
    if (t < T) {
      N=N+1
      if (state == 0) {
        state = 1
      }
      else {
        state = 0
      }
      states=c(states, state)
      waitingTimes = c(waitingTimes, X)
    }
  }
  return (list(N, states, waitingTimes))
}
```

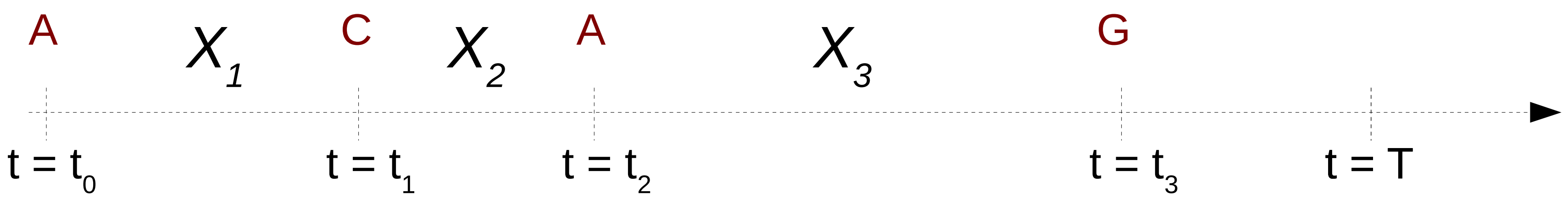
Modelling the evolution of a DNA character in continuous time



Modelling the evolution of a DNA character in continuous time



Modelling the evolution of a DNA character in continuous time

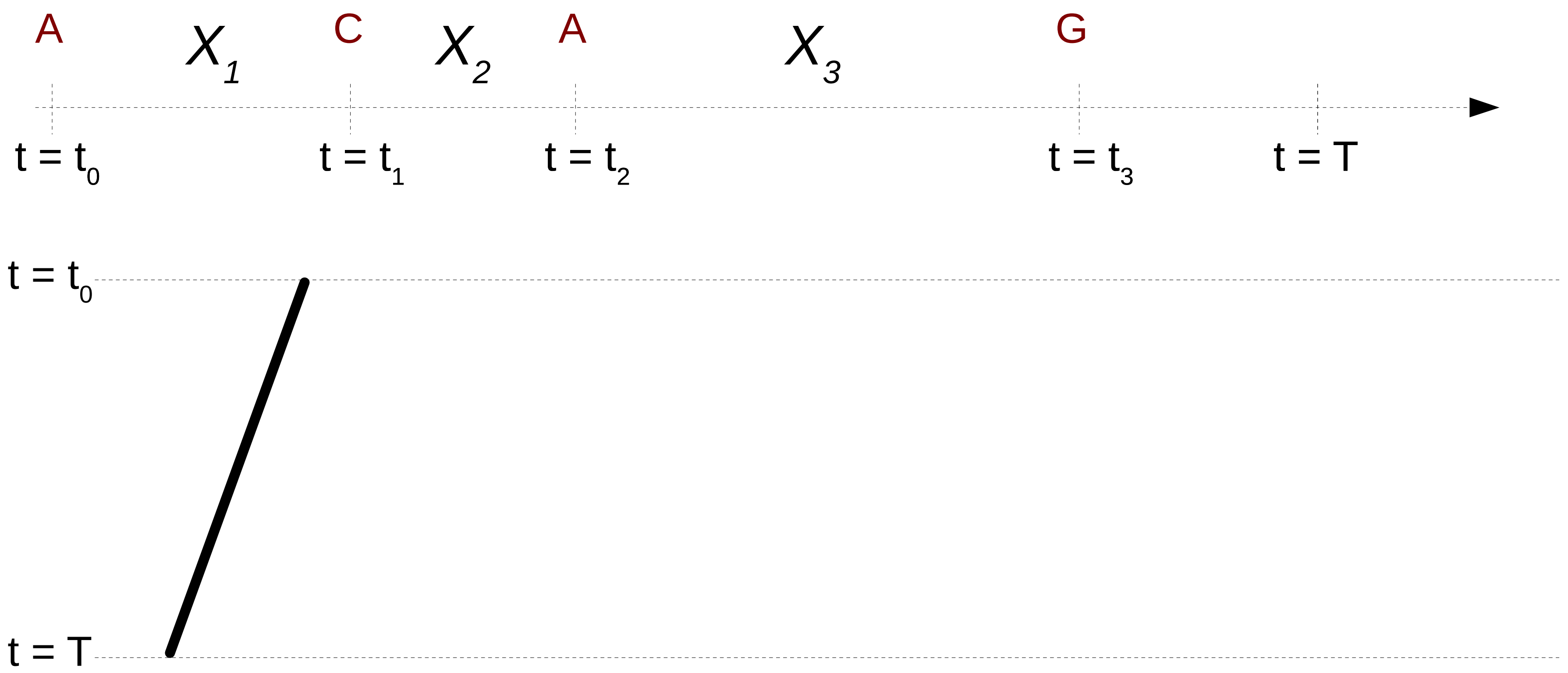


- Draw an initial state from a Multinomial distribution:

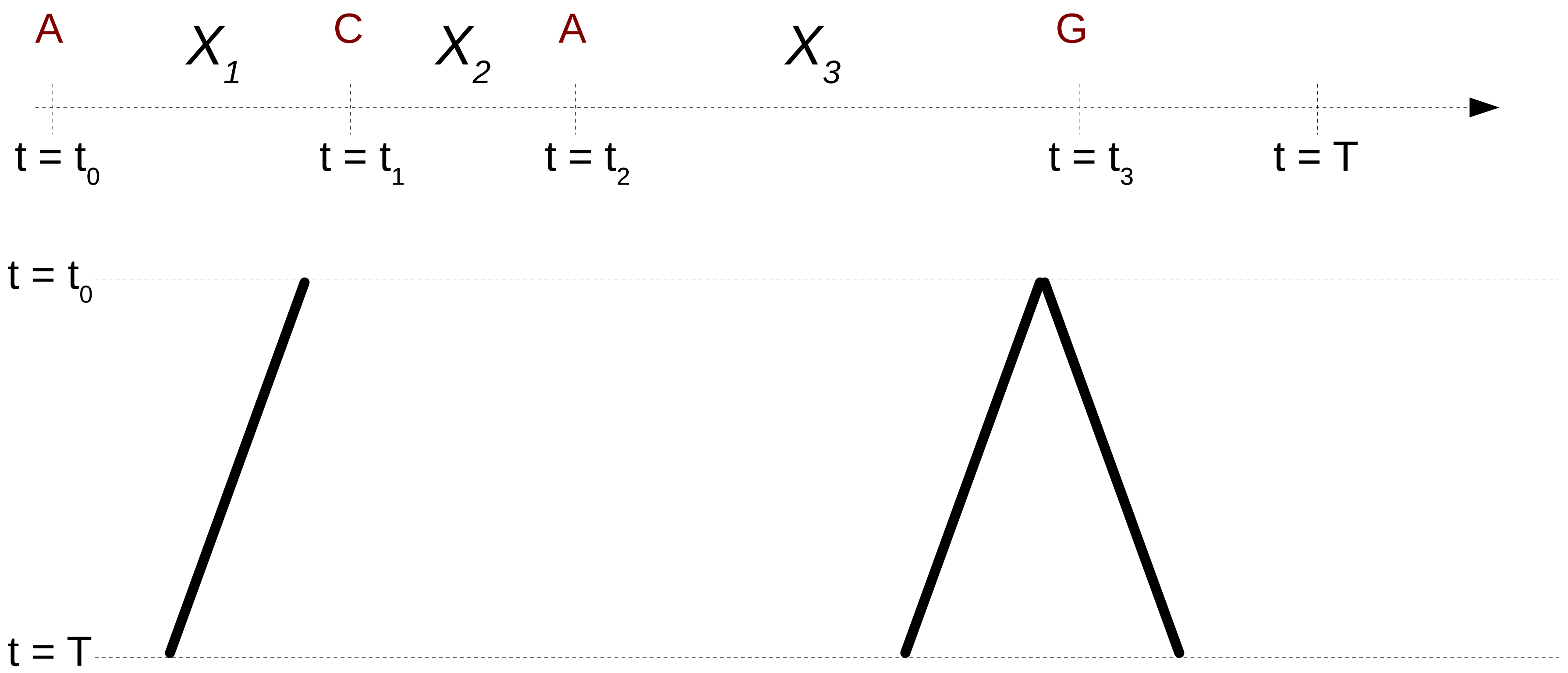
```
p=c(0.25, 0.25, 0.25, 0.25); state=rmultinom(n=1, p=p, size=1)
```

- $t = t_0$; $N = 0$; $\lambda=0.1$
- While $t < T$:
  - Draw from an exponential distribution a waiting time X_i until the next event; $t = t + X_i$
    - If $t < T$, change the state of the variable: `state=rmultinom(n=1, p=p, size=1)`
    - (Else ($t \geq T$): we stop)

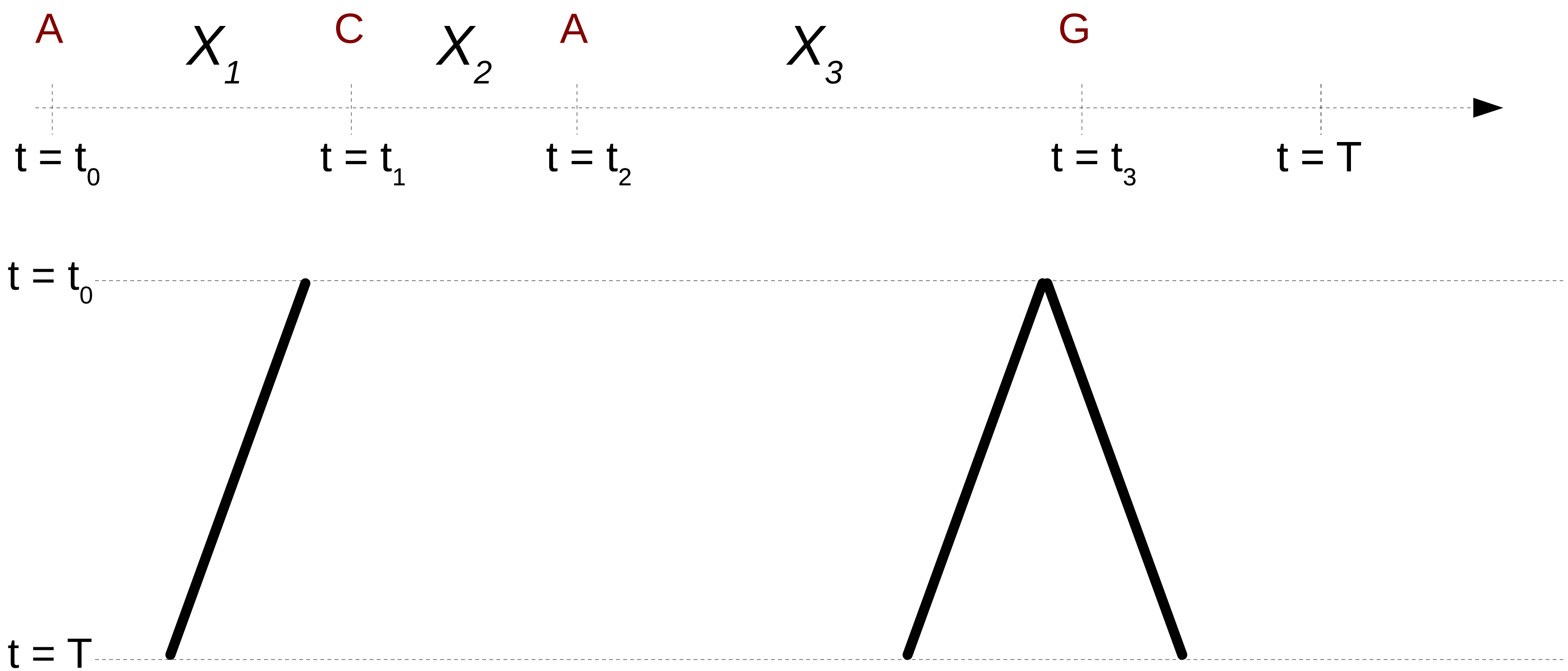
From one branch to two



From one branch to two



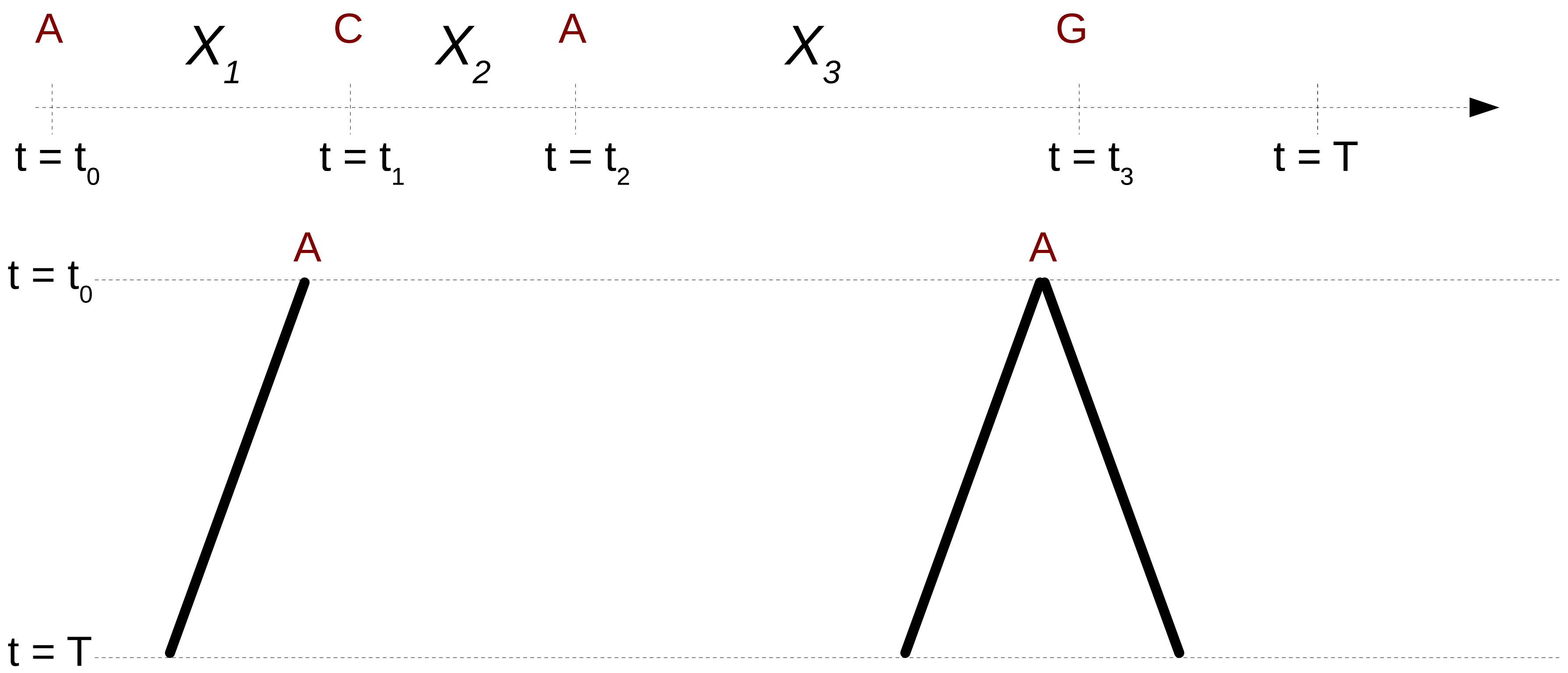
From one branch to two



Draw an initial state from a Multinomial distribution

- Left branch = simulate(state, T)
- Right branch = simulate(state, T)

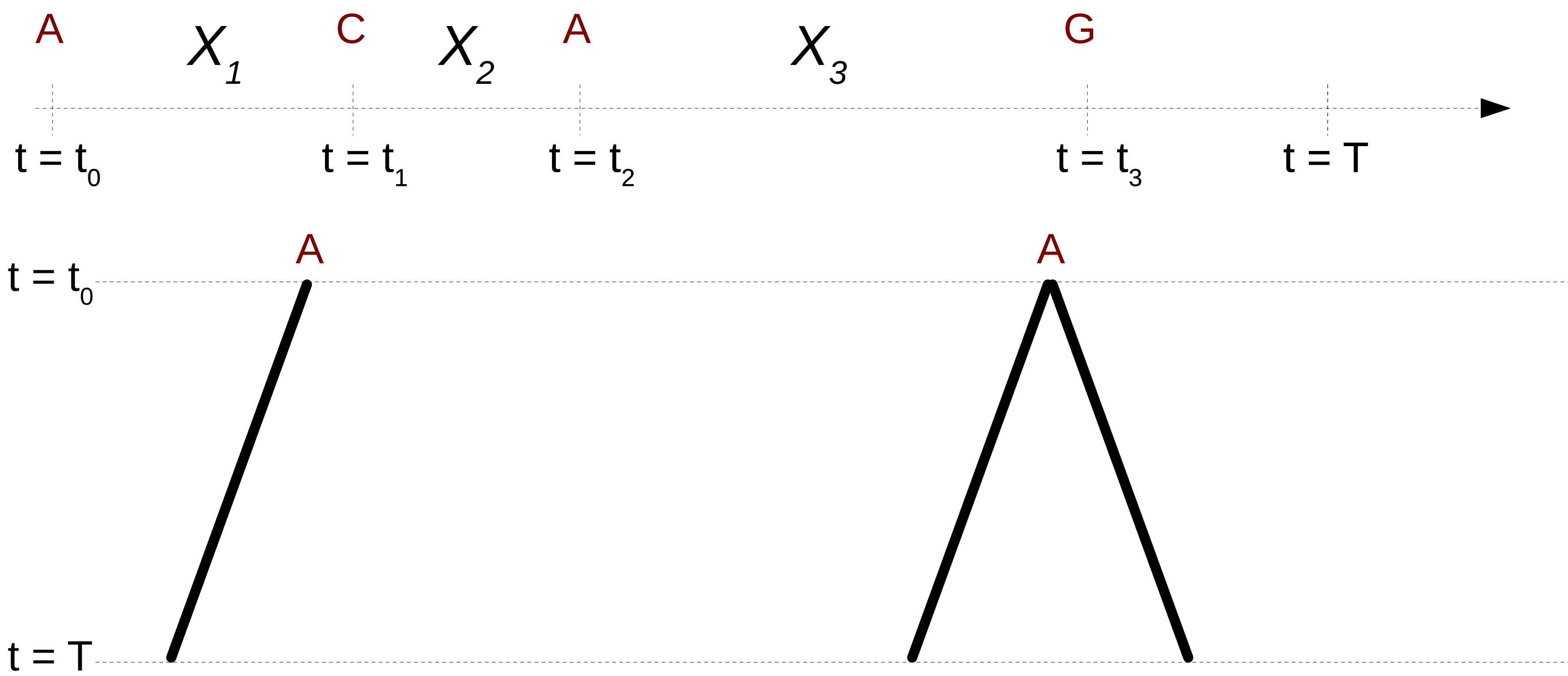
From one branch to two



Draw an initial state from a Multinomial distribution

- Left branch = simulate(state, T)
- Right branch = simulate(state, T)

From one branch to two

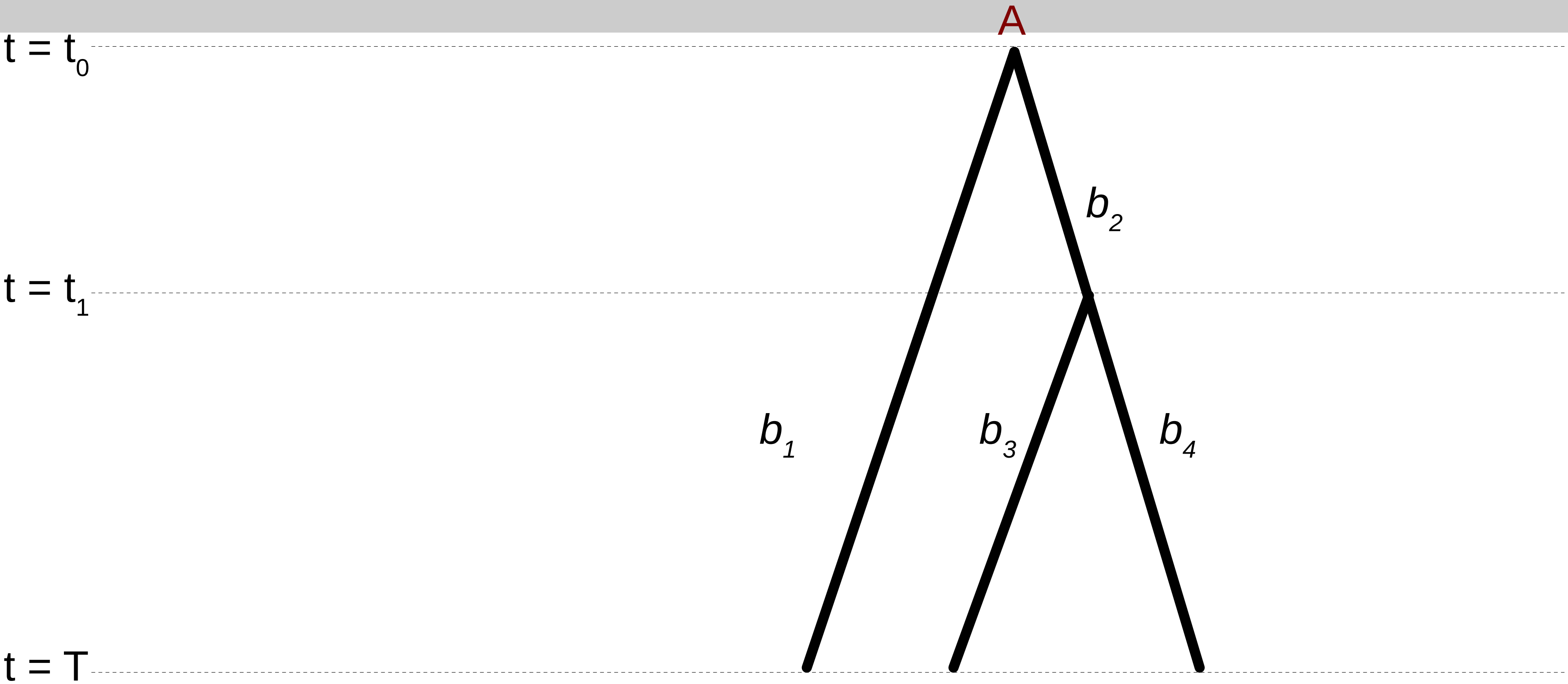


Draw an initial state from a Multinomial distribution

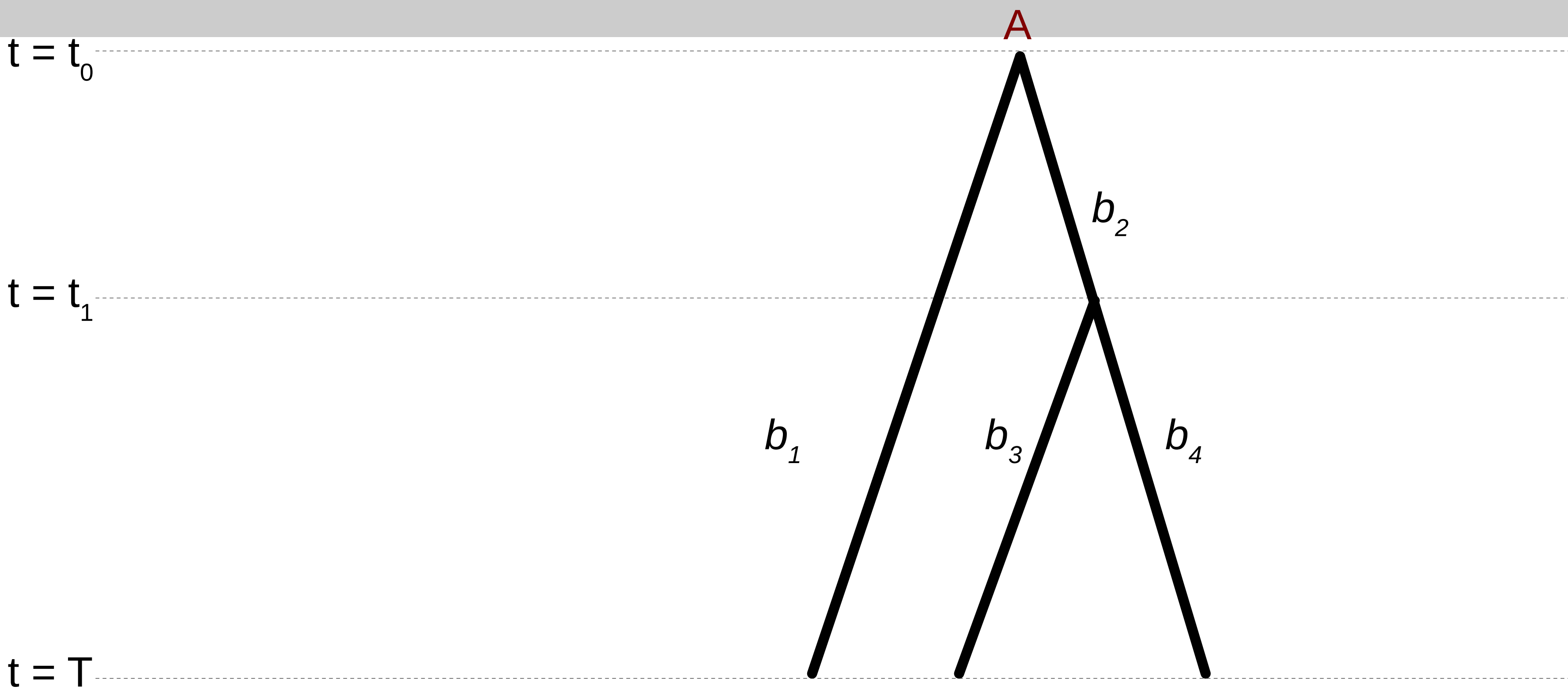
- Left branch = simulate(state, T)
- Right branch = simulate(state, T)

We assume independence between the two branches

Simulating on a tree



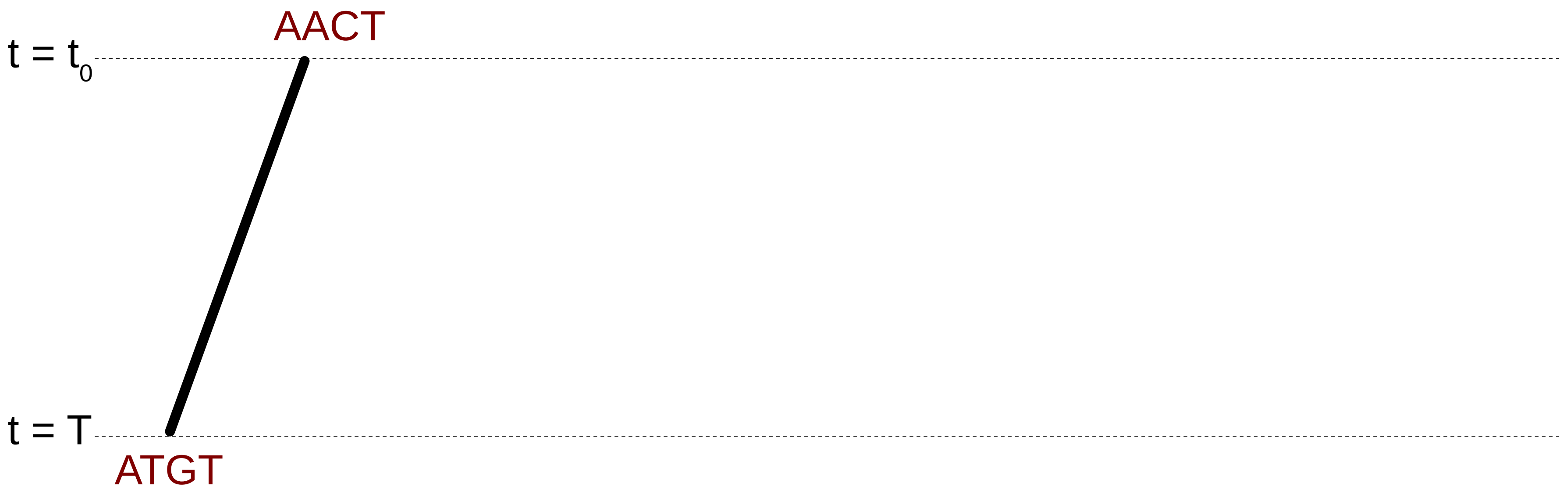
Simulating on a tree



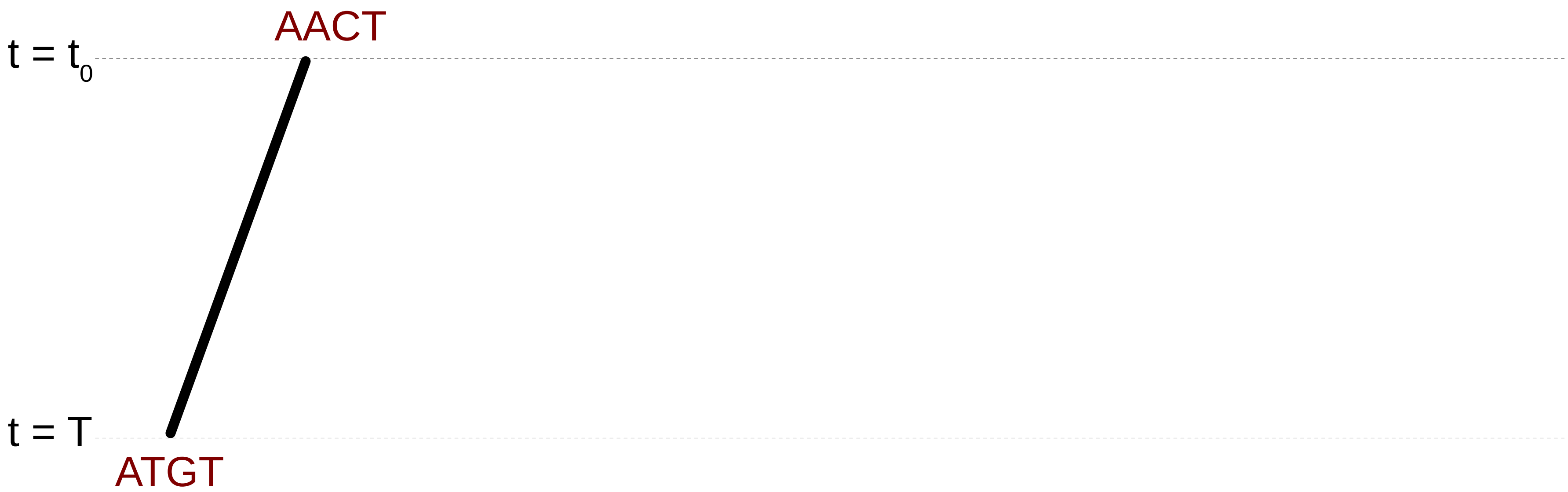
Draw an initial state from a Multinomial distribution

- Branch b_1 = simulate(state, T)
- Branch b_2 = simulate(state, t_1)
- Branch b_3 = simulate(state(end Branch b_2), T-t_1)
- Branch b_4 = simulate(state(end Branch b_2), T-t_1)

From one site to several



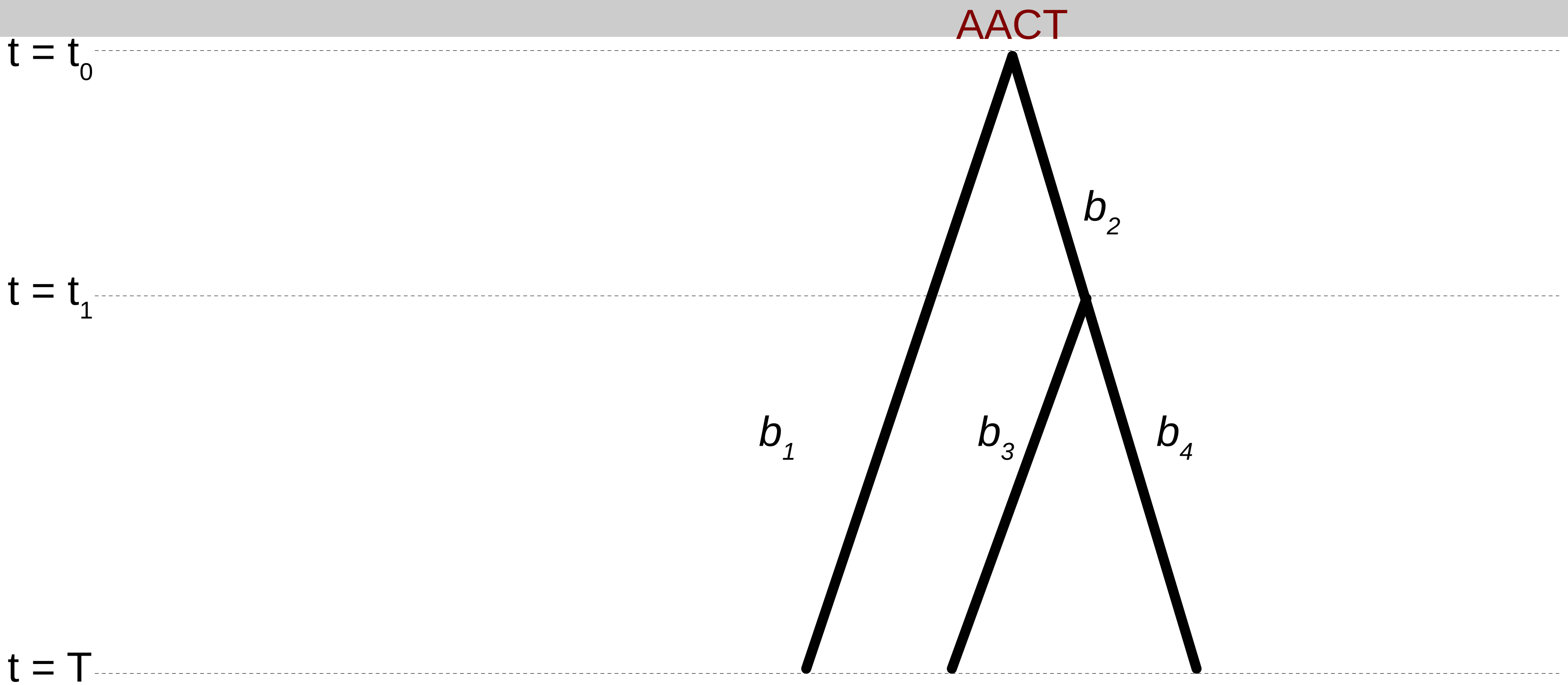
From one site to several



For i in {1..Number of sites}

- Draw an initial state *$state_i$* from a Multinomial distribution
- $Site_i$ = simulate(*$state_i$*, T)

From one site to several, on a tree



For i in {1..Number of sites}

- Draw an initial state *state*$_i$ from a Multinomial distribution
- Site$_i$ = simulate_along_tree(*state*$_i$)

Summary

- We can simulate the evolution of a DNA character by drawing an initial state, then waiting times between substitutions
- We can simulate on a tree by taking as initial state for child branches the terminal state of parent branches
- We can simulate many independent sites
- As long as all substitutions are equally likely, this model is the Jukes-Cantor model (1969)

Other models of sequence evolution

Rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

Other models of sequence evolution

Rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

Jukes and Cantor 1969

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

1 free parameter (0 if we impose one substitution per unit time)

Kimura 1980

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

1 transition/transversion ratio : 1 free parameter

Hasegawa, Kishino, Yano 1985

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

1 transition/transversion ratio
4 equilibrium frequencies:
4 free parameters

Other models of sequence evolution

Rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

Jukes and Cantor 1969

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

1 free parameter (0 if we impose one substitution per unit time)

Kimura 1980

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

1 transition/transversion ratio : 1 free parameter

Hasegawa, Kishino, Yano 1985

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

1 transition/transversion ratio
4 equilibrium frequencies:
4 free parameters

All those are particular cases of the GTR model

General Time Reversible model of sequence evolution

Rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

Lanave et al. 1984; Tavaré, 1986

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

4 ***equilibrium frequencies***: 3 parameters
6 ***exchangeability parameters***: 5 parameters (if we impose one substitution per unit time)

More general models do not assume reversibility (e.g. Barry-Hartigan model)