

Models in genomics and biodiversity analysis

Xia Hua

Mathematical Sciences Institute

Australian National University

xia.hua@anu.edu.au

Common mathematical tools in biology

Parameterization

- Stochasticity

Discrete: Markov Chain

Continuous: Diffusion Process

- Complexity:

Confounding factors: Regression

Estimation

- Maximum likelihood
- Bayesian Inference

Markov Chain: A chain of events with Markov property

Random variable X_n

A substitution event can leads to a state of $\{A,T,C,G\}$

The probability of the n^{th} substitution event that leads to A: $P(X_n = A)$

$$X_n = [P(X_n = A), P(X_n = T), P(X_n = C), P(X_n = G)]$$

Markov Chain

A sequence of random variables with Markov Property:

$$P(X_{n+1} = A | X_n = T, X_{n-1} = C, \dots) = P(X_{n+1} = A | X_n = T)$$

Discrete Time Markov Chain

A sequence of random variables with Markov Property:

$$P(X_{n+1} = A | X_n = T, X_{n-1} = C, \dots) = P(X_{n+1} = A | X_n = T)$$

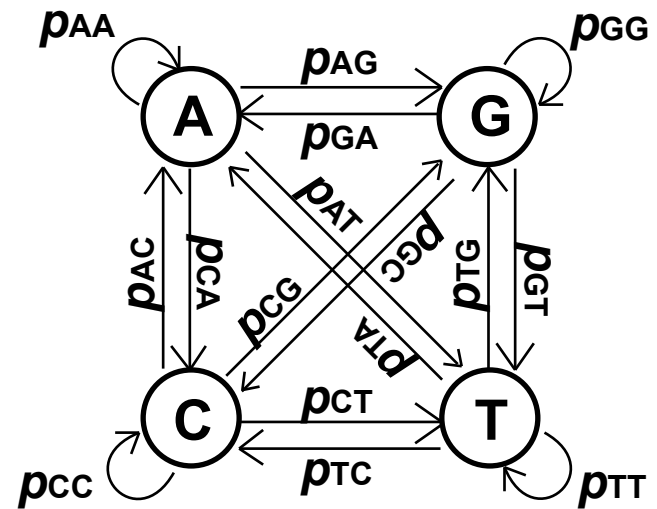
Discrete Time Markov Chain

A sequence of random variables with Markov Property:

$$P(X_{n+1} = A | X_n = T, X_{n-1} = C, \dots) = P(X_{n+1} = A | X_n = T)$$

Transition matrix:

$$P = \begin{bmatrix} p_{AA} & p_{AT} & p_{AC} & p_{AG} \\ p_{TA} & p_{TT} & p_{TC} & p_{TG} \\ p_{CA} & p_{CT} & p_{CC} & p_{CG} \\ p_{GA} & p_{GT} & p_{GC} & p_{GG} \end{bmatrix}$$



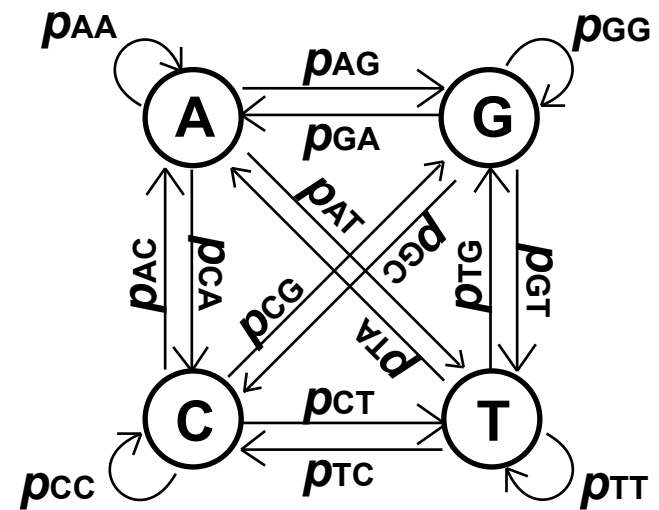
Discrete Time Markov Chain

A sequence of random variables with Markov Property:

$$P(X_{n+1} = A | X_n = T, X_{n-1} = C, \dots) = P(X_{n+1} = A | X_n = T)$$

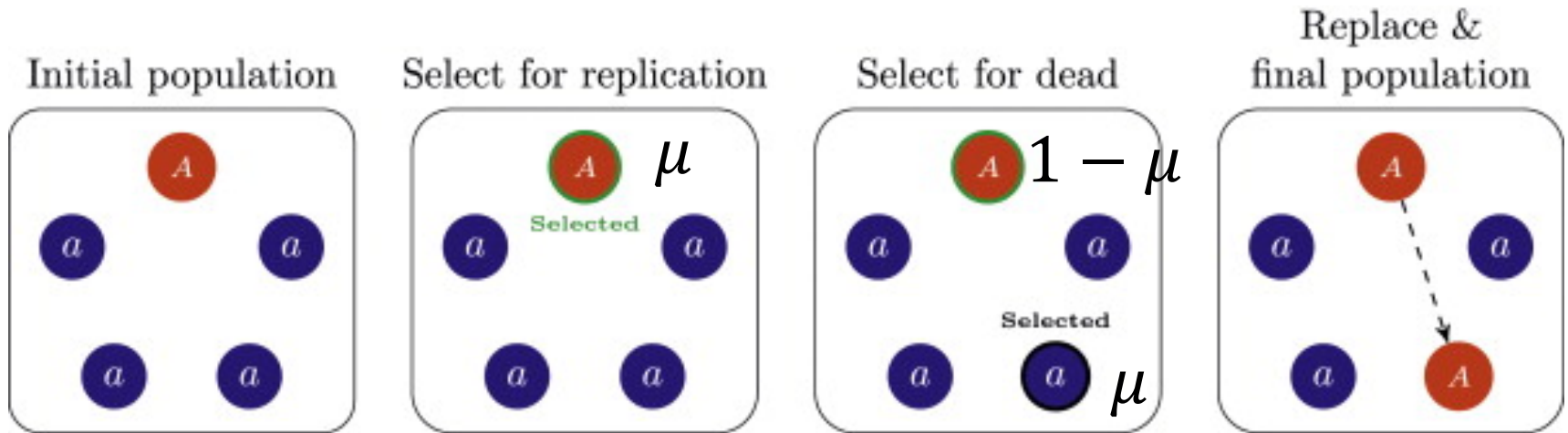
Transition matrix:

$$P = \begin{bmatrix} p_{AA} & p_{AT} & p_{AC} & p_{AG} \\ p_{TA} & p_{TT} & p_{TC} & p_{TG} \\ p_{CA} & p_{CT} & p_{CC} & p_{CG} \\ p_{GA} & p_{GT} & p_{GC} & p_{GG} \end{bmatrix}$$



$$\begin{aligned} X_n &= [P(X_{n-1} = A), P(X_{n-1} = T), P(X_{n-1} = C), P(X_{n-1} = G)]P \\ &= X_{n-1}P = X_{n-2}P^2 = \dots = X_0P^n \end{aligned}$$

DTMC – Moran Model

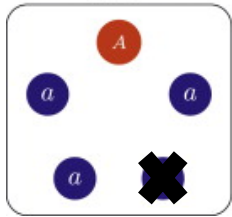


For red individual, states = $\{1, 2, \dots, N\}$, $\mu = 1/N$

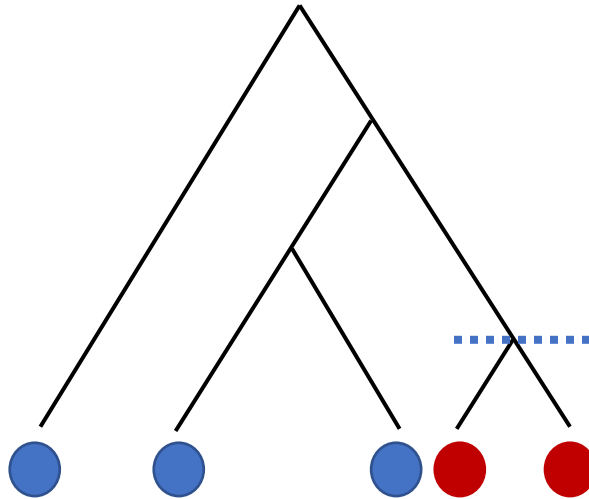
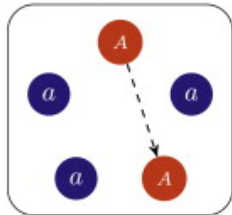
$$P(X_{n+1} = j | X_n = i) = \begin{cases} \mu(1 - \mu) & \text{if } j = i + 1 \\ (1 - \mu)\mu & \text{if } j = i - 1 \\ \mu^2 + (1 - \mu)^2 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

From Moran model to Coalescence

i-1 ancestors



i descendants



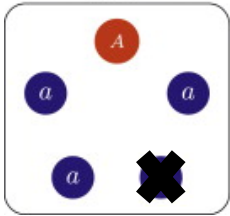
Parent **A** does not die

$$P(i \text{ lineages from } i - 1 \text{ ancestors}) = \left(1 - \frac{1}{N}\right) \left(\frac{i}{N} \frac{i-1}{N-1}\right) = \binom{i}{2} \frac{2}{N^2}$$

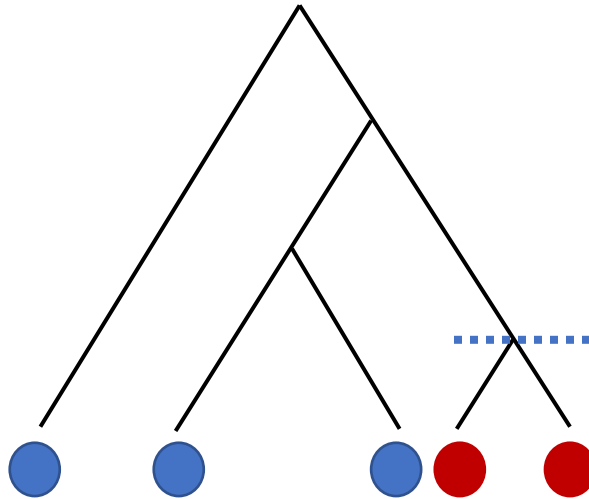
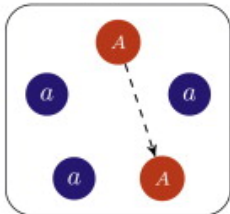
Both parent and offspring **A** are in the *i* lineages

From Moran model to Coalescence

i-1 ancestors



i descendants



$$\text{If } t = \frac{2}{N^2}$$

$$\text{as } N \rightarrow \infty, t \rightarrow 0, P \rightarrow \binom{i}{2}$$

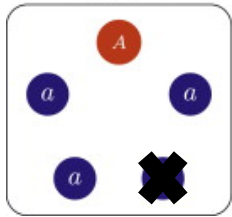
Parent **A** does not die

$$P(i \text{ lineages from } i - 1 \text{ ancestors}) = \left(1 - \frac{1}{N}\right) \left(\frac{i}{N} \frac{i-1}{N-1}\right) = \binom{i}{2} \frac{2}{N^2}$$

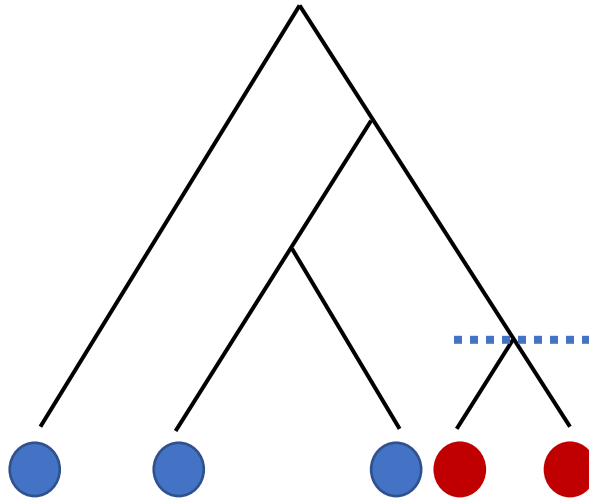
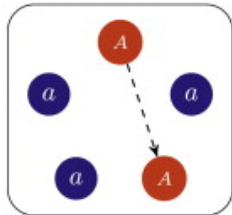
Both parent and offspring **A** are in the *i* lineages

From Moran model to Coalescence

i-1 ancestors



i descendants



$$\text{If } t = \frac{2}{N^2}$$

$$\text{as } N \rightarrow \infty, t \rightarrow 0, P \rightarrow \binom{i}{2}$$

$$f(T = t) = \binom{i}{2} e^{-\binom{i}{2}t}$$

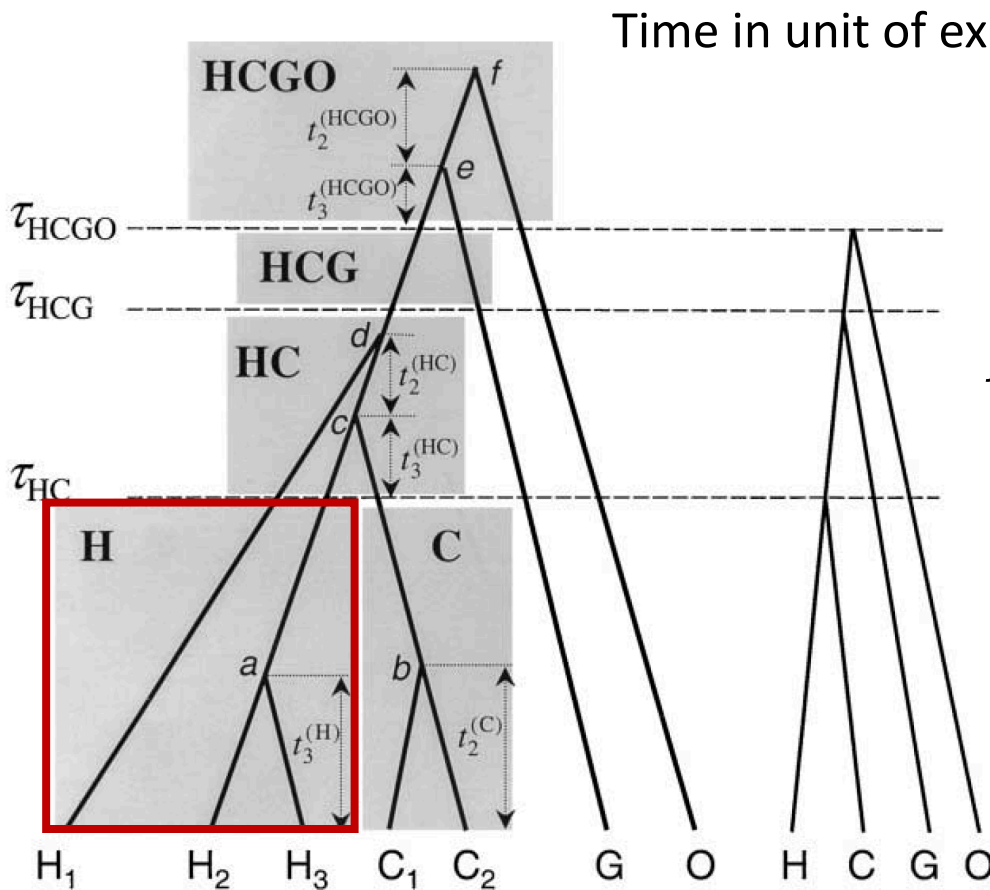
$$p(\text{no coel in } t) = e^{-\binom{i}{2}t}$$

Parent **A** does not die

$$P(i \text{ lineages from } i - 1 \text{ ancestors}) = \left(1 - \frac{1}{N}\right) \left(\frac{i}{N} \frac{i-1}{N-1}\right) = \binom{i}{2} \frac{2}{N^2}$$

Both parent and offspring **A** are in the *i* lineages

Multi-Species Coalescence



Rannala and Yang 2003 Genetics

$$f(T = t) = \binom{i}{2} \frac{2}{\theta} e^{-\binom{i}{2} \frac{2}{\theta} t}$$

$$f(H) = \binom{3}{2}^{-1}$$

A coal event happens

$$\times \binom{3}{2} \frac{2}{\theta_H} e^{-\binom{3}{2} \frac{2}{\theta_H} t_3^{(H)}}$$

And the event happens at $t_3^{(H)}$

$$\times e^{-\binom{2}{1} \frac{2}{\theta_H} (\tau_{HC} - t_3^{(H)})}$$

And no coal event between τ_{HC} and $t_3^{(H)}$

Continuous Time Markov Chain

$$P(X_{n+1} = A | X_n = T) = p_{TA}$$

$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0$$

Continuous Time Markov Chain

$$P(X_{n+1} = A | X_n = T) = p_{TA}$$

$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0 \quad \rightarrow Q$$

Transition rate matrix

$$P(X_{t+\Delta t} = A | X_t = T) = p_{TA}(t; t + \Delta t) \rightarrow P$$

Continuous Time Markov Chain

$$P(X_{n+1} = A | X_n = T) = p_{TA}$$

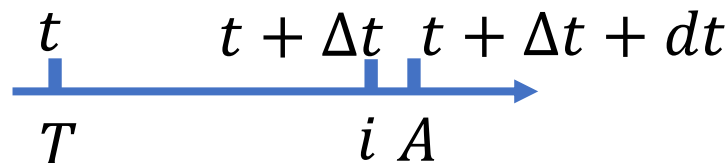
$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0 \quad \rightarrow Q$$

Transition rate matrix

$$P(X_{t+\Delta t} = A | X_t = T) = p_{TA}(t; t + \Delta t) \rightarrow P$$

Forward equation:

$$p'_{TA}(t; t + \Delta t) = \sum_{i \in \{A, T, C, G\}} p_{Ti}(t; t + \Delta t) q_{iA}$$



Continuous Time Markov Chain

$$P(X_{n+1} = A | X_n = T) = p_{TA}$$

$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0 \quad \rightarrow Q$$

Transition rate matrix

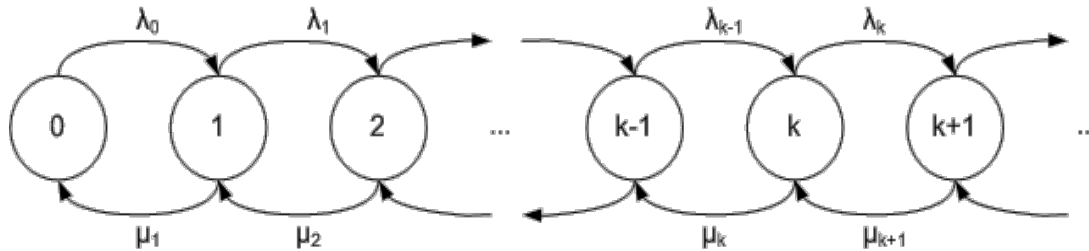
$$P(X_{t+\Delta t} = A | X_t = T) = p_{TA}(t; t + \Delta t) \rightarrow P$$

Forward equation:

$$p'_{TA}(t; t + \Delta t) = \sum_{i \in \{A, T, C, G\}} p_{Ti}(t; t + \Delta t) q_{iA}$$

$$P'(t) = P(t)Q, P(0) = I \quad \rightarrow \boxed{P(\Delta t) = e^{Q\Delta t}}$$

CTMC – Birth-death process

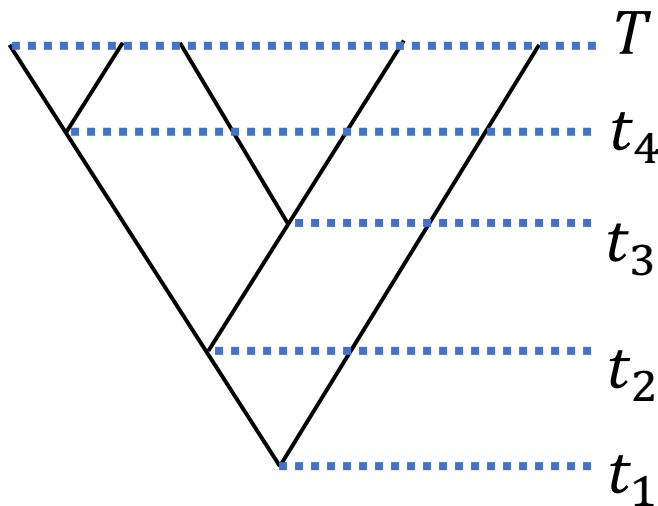


States = $\{0, 1, 2, \dots, N\}$

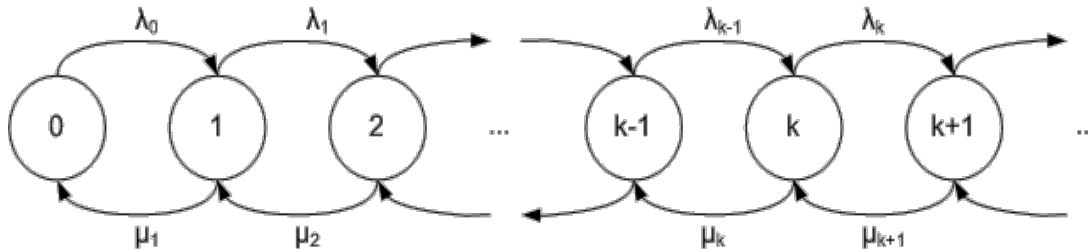
$$q_{k,k+1} = \lambda$$

$$q_{k,k-1} = \mu$$

$$p'_{kk}(0; t) = p_{k,k-1}\lambda + p_{k,k+1}\mu + p_{k,k}(-\lambda - \mu)$$



CTMC – Birth-death process

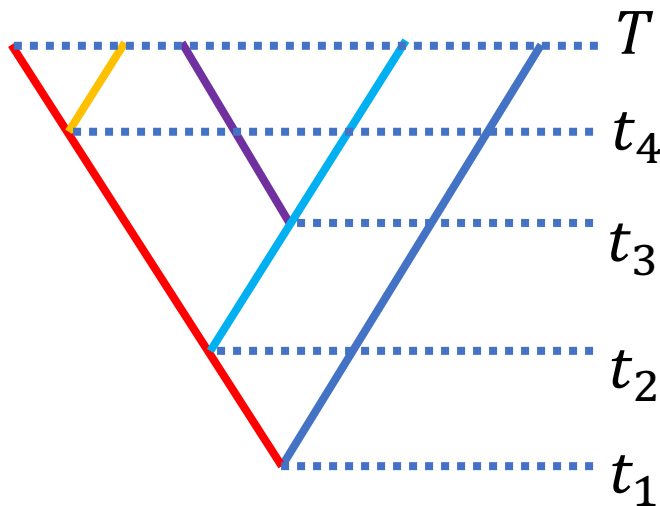


States = $\{0, 1, 2, \dots, N\}$

$$q_{k,k+1} = \lambda$$

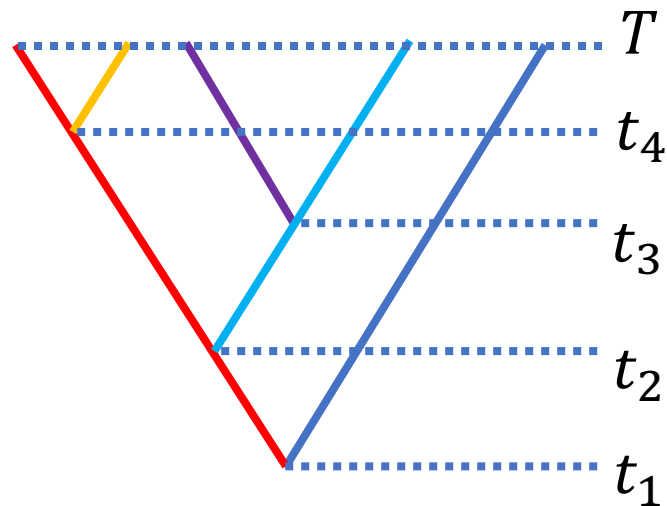
$$q_{k,k-1} = \mu$$

$$p'_{kk}(0; t) = p_{k,k-1}\lambda + p_{k,k+1}\mu + p_{k,k}(-\lambda - \mu)$$



$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

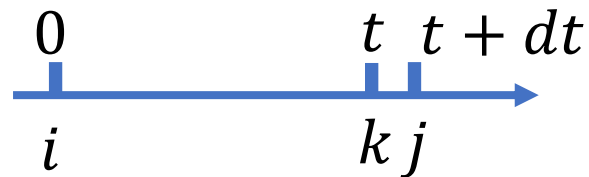
CTMC – Birth-death process



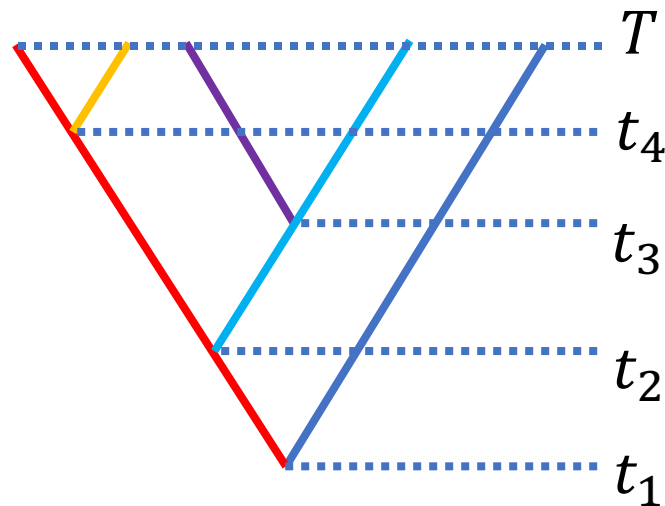
$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

Forward equation:

$$p'_{ij}(0; t) = \sum p_{ik}(0; t) q_{kj}$$



CTMC – Birth-death process



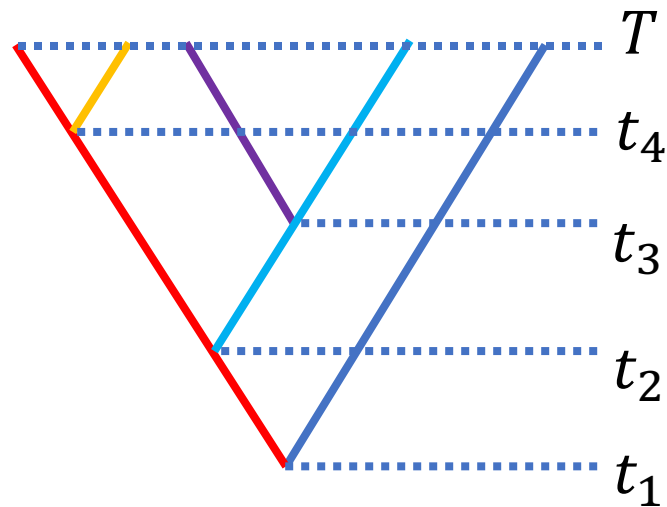
$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$



CTMC – Birth-death process



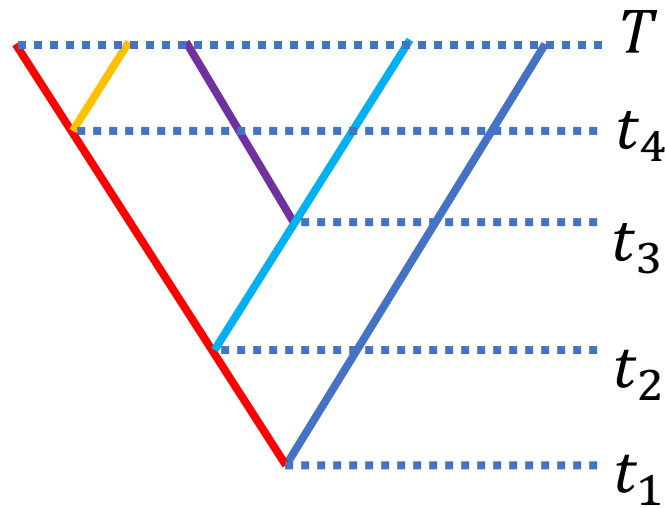
$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

$$p'_{11}(T; T - t_i) = \mu p_{01} - (\lambda + \mu) p_{11} + \lambda p_{21}, p_{11}(T; T) = 1$$

CTMC – Birth-death process



$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

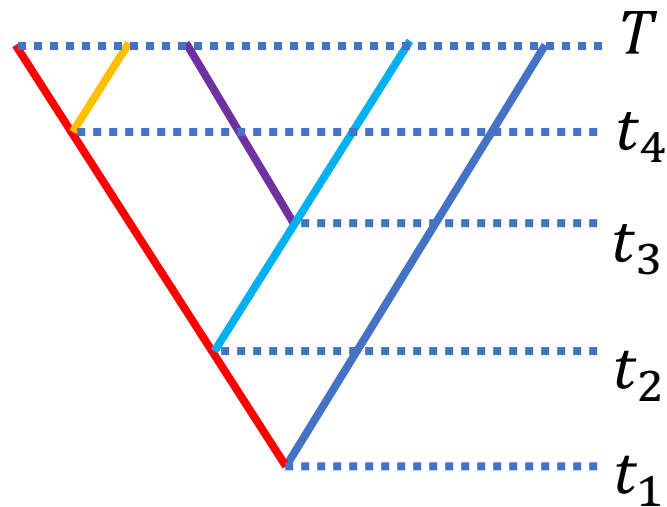
Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

T $p'_{11}(T; T - t) = \mu p_{01} - (\lambda + \mu)p_{11} + \lambda p_{21}, p_{11}(T; T) = 1$

t
 $t + dt$
 t_1

CTMC – Birth-death process



$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

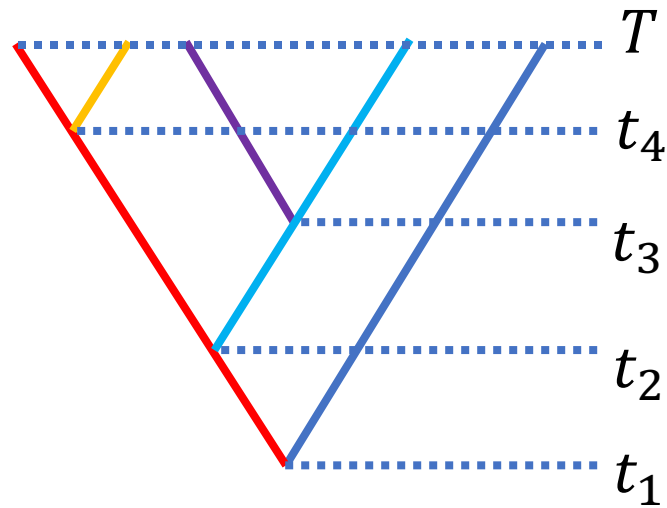
Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

$$p'_{11}(T; T - t) = \mu p_{01} - (\lambda + \mu) p_{11} + \lambda p_{21}, \quad p_{11}(T; T) = 1$$

||
0

CTMC – Birth-death process



$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

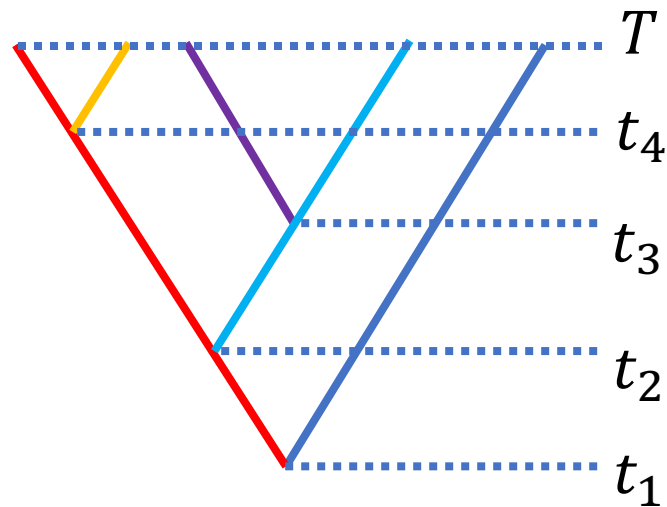
Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

$$p'_{11}(T; T - t) = \mu p_{01} - (\lambda + \mu) p_{11} + \lambda p_{21}, \quad p_{11}(T; T) = 1$$

||
0

CTMC – Birth-death process



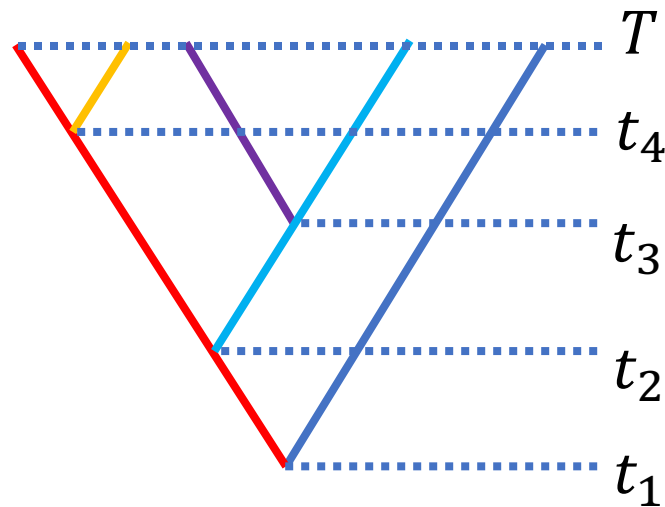
$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

$$p'_{11}(T; T - t) = \underbrace{\mu p_{01}}_{= 0} - (\lambda + \mu) p_{11} + \lambda \underbrace{p_{21}}_{= 2p_{11}p_{10}}, \quad p_{11}(T; T) = 1$$

CTMC – Birth-death process



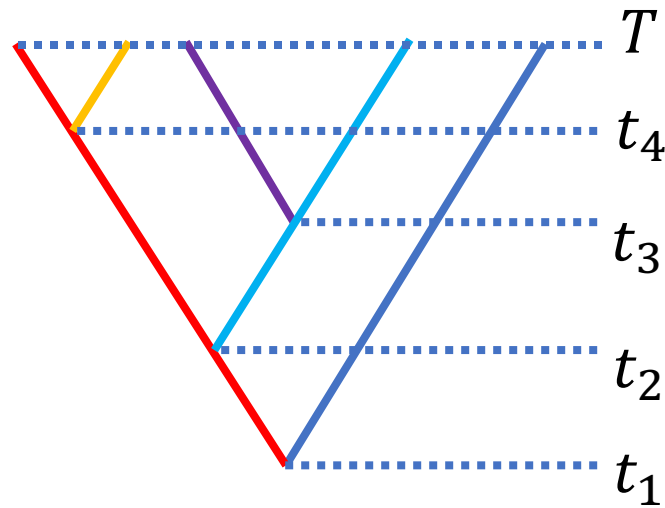
$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

$$p'_{11}(T; T - t_i) = -(\lambda + \mu)p_{11} + \lambda 2p_{11}p_{10}, p_{11}(T; T) = 1$$

CTMC – Birth-death process



$$f(t_1, \dots, t_4) = \prod_{i=1}^5 i \times \lambda(t_i) \times p_{11}(T; T - t_i)$$

Backward equation:

$$p'_{ij}(T; T - t) = \sum q_{jk} p_{ki}(T; T - t)$$

$$p'_{11}(T; T - t_i) = -(\lambda + \mu)p_{11} + \lambda 2p_{11}p_{10}, p_{11}(T; T) = 1$$

$$p'_{10}(T; T - t_i) = \mu p_{00} - (\lambda + \mu)p_{10} + \lambda p_{20}, p_{10}(T; T) = 0$$

\parallel
1

\parallel
 p_{10}^2

$$= \mu - (\lambda + \mu)p_{10} + \lambda p_{10}^2$$

Diffusion Process

$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0$$

$$P(X_{t+dt} = x + dx | X_t = x) = q_x, dt \rightarrow 0, dx \rightarrow 0$$

Diffusion Process

$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0$$

$$P(X_{t+dt} = x + dx | X_t = x) = q_x, dt \rightarrow 0, dx \rightarrow 0$$

$$P(X_{t+\Delta t} = x + \Delta x | X_t = x) = p(x, t)$$


Diffusion Process

$$P(X_{t+dt} = A | X_t = T) = q_{TA}, dt \rightarrow 0$$

$$P(X_{t+dt} = x + dx | X_t = x) = q_x, dt \rightarrow 0, dx \rightarrow 0$$

$$P(X_{t+\Delta t} = x + \Delta x | X_t = x) = p(x, t)$$

Forward equation: $\frac{d}{dt}P = P(t)Q$

$$\frac{\partial}{\partial t} p(x, t) = - \frac{\partial}{\partial x} [\mu(x, t) p(x, t)] + \frac{\partial^2}{\partial x^2} \left[\frac{1}{2} D(x, t) p(x, t) \right]$$


Change in mean of x in dt

Change in variance of x in dt

Diffusion Process – Trait evolution

	$\mu(x, t)$	$D(x, t)$
Brownian Motion	0	σ^2
BM with trend	b	σ^2
Ac/Decelerating	0	$\sigma^2 e^{rt}$
Ornstein-Uhlenbeck	$b(\mu - x)$	σ^2
Peak shift	$b(\mu_t - x)$	σ^2

Common mathematical tools in biology

Parameterization

- Stochasticity

Discrete: Markov Chain

Continuous: Diffusion Process

- Complexity:

Confounding factors: Regression

Estimation

- Maximum likelihood
- Bayesian Inference

Regression Models

$$Y_i = f(X_i, \beta) + e_i$$

General Linear Model

$$Y_i = X_i\beta + e_i, e_i \sim N(0, \Omega)$$

Autoregressive Model

$$Y_t = X_t\beta + \sum_{j=1}^p Y_{t-j} + e_t, e_t \sim N(0, \sigma^2)$$

Linear Mixed Model

$$Y_i = X_i\beta + Zu + e_i, u \sim N(0, \Omega), e_i \sim N(0, \sigma^2)$$

Generalized Linear Model

$$g(Y_i) = X_i\beta + e_i, e_i \sim N(0, \Omega)$$

Genome-Wide Association Studies

Fisher's polygenic model:

$$y_i = \mu + \sum_{l=1}^L G_{il} + \varepsilon_i$$



Phenotype of
individual i



Genotype of
locus l of
individual i

Kinship Coefficient Matrix

$$Var(y) = 2\sigma_a^2\phi + \sigma_e^2I$$



Additive
genetic
variance



Environmental
variance

Linear mixed model:

$$g(y_i) = \mu + \sum_j \beta_j X_j + \beta_k G_{ik} + \sum_{l \neq k}^L \beta_l G_{il} + \varepsilon_i$$



Confounding factors



Test SNP

\parallel

$$\eta_i \sim N(0, 2\sigma_a^2\phi + \sigma_e^2I)$$

Accounting for genetic relatedness

Population admixture:

$$E(G_{i,j}) = 2 \sum_{k=1}^K \phi_{i,k} P_{k,j}$$

Most restricted: $\sum_{k=1}^K \phi_{i,k} = 1, \phi_{i,k} \geq 0$

 Admixture proportion

Principle component analysis:

$$E(G_{i,j}) = (\phi F)_{ij}$$

No restriction

 loadings factors

Accounting for genetic relatedness

Population admixture:


$$E(G_{i,j}) = 2 \sum_{k=1}^K \phi_{i,k} P_{k,j}$$

Most restricted: $\sum_{k=1}^K \phi_{i,k} = 1, \phi_{i,k} \geq 0$

 Admixture proportion

Sparse factor analysis:

$$E(G_{i,j}) = (\phi F)_{ij}$$

 Encouraging sparsity in ϕ
by giving prior $\phi_{i,k} \sim N(0, \sigma_{i,k}^2)$

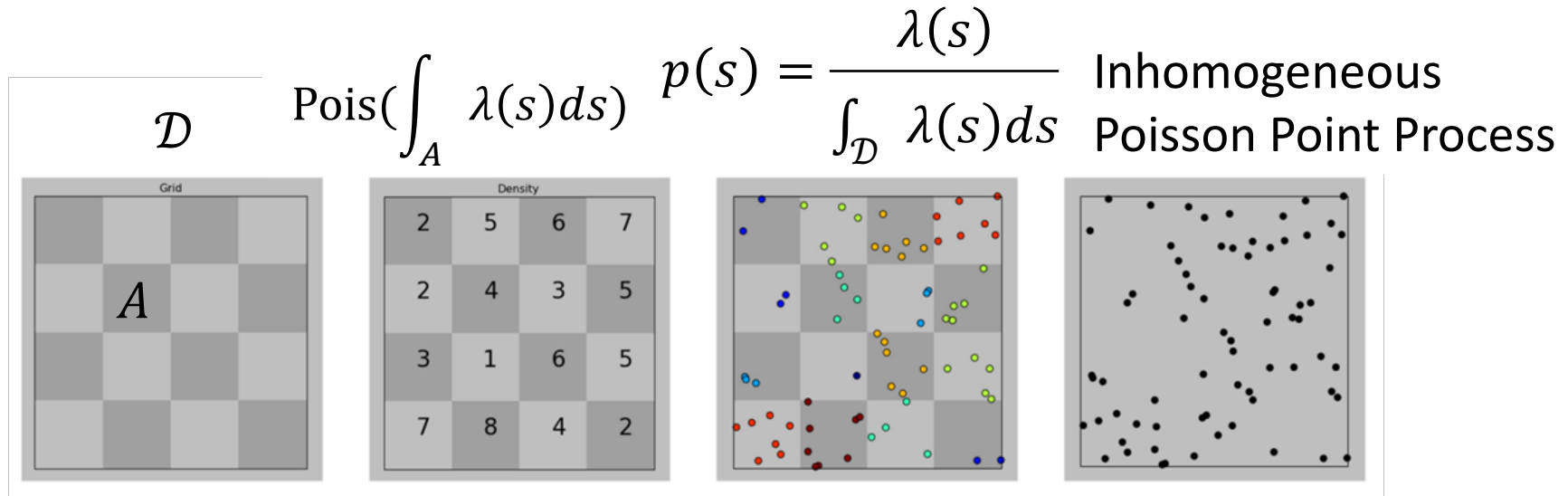
Principle component analysis:

$$E(G_{i,j}) = (\phi F)_{ij}$$

No restriction

 loadings factors

Species distribution modeling



$$\ln \lambda(s) = \alpha + X(s)\beta$$

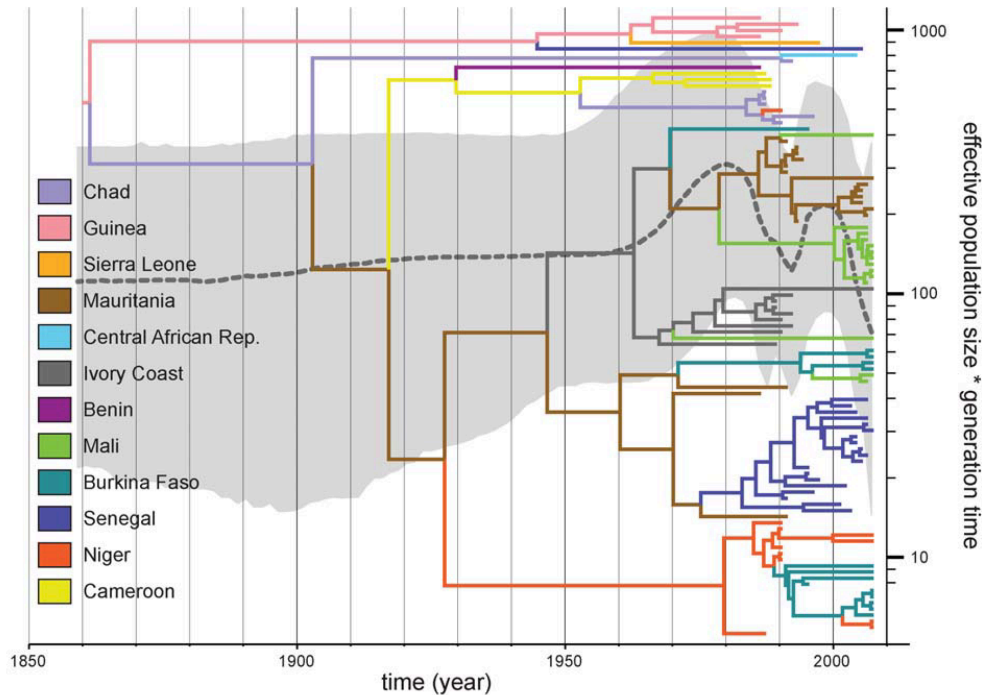
Presence-absence in quadrat A : $N_A \approx \text{Pois}(\lambda(s)A)$

Absence: $P(N_A = 0) = e^{-Ae^{\alpha + X(s)\beta}}$

Presence: $P(N_A > 0) = 1 - e^{-Ae^{\alpha + X(s)\beta}} \approx Ae^{\alpha + X(s)\beta}$

Presence-only at s : $\ln p(S) = \alpha + X(s)\beta - \sum_{i \in BG} \frac{\mathcal{D}}{n_{BG}} e^{\alpha + X(i)\beta}$

Pathogen Phylodynamics



Lemey et al. 2009 PLoS Computational Biology

States: {Chad, Guinea,...}

$$P(T) = e^{Qt}$$

$$\log Q_{ij} = \sum_k \delta_k \beta_k X_{ijk}$$

X_{ijk} : value of predictor k
between location i and j

β_k : effective size of predictor k

δ_k : (0,1)-indicator of the
inclusion of predictor k

Common mathematical tools in biology

Parameterization

- Stochasticity

Discrete: Markov Chain

Continuous: Diffusion Process

- Complexity:

Confounding factors: Regression

Estimation

- Maximum likelihood
- Bayesian Inference


Maximum Likelihood

Likelihood: $P(\text{Data}|\text{Parameter})$

Maximum Likelihood

Likelihood: $P(\text{Data}|\text{Parameter})$

E.g., $Y_i = X_i\beta + e_i, e_i \sim N(0, \sigma^2)$



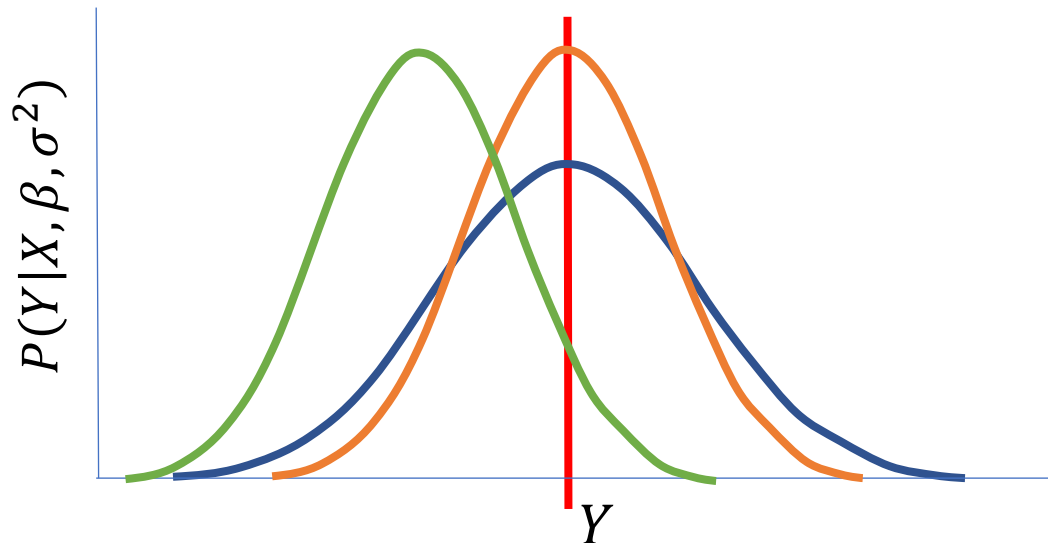
$$P(Y|X, \beta, \sigma^2) = N(X\beta, \sigma^2)$$

Maximum Likelihood

Likelihood: $P(\text{Data}|\text{Parameter})$

E.g., $Y_i = X_i\beta + e_i, e_i \sim N(0, \sigma^2)$

$$P(Y|X, \beta, \sigma^2) = N(X\beta, \sigma^2)$$



Maximum Likelihood

$$L = \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}}$$

$$-\ln L = \sum_{i=1}^n \frac{1}{2} \ln(2\pi\sigma^2) + \frac{(Y_i - X_i\beta)^2}{2\sigma^2}$$

```
f <- function (p, X, Y) {  
  beta <- p[1]  
  sigma <- exp(p[2])  
  n <- length(Y)  
  nll <- n/2*log(2*pi*sigma) + sum((Y-X*beta)^2)/2/sigma  
  nll  
}  
mle <- optim(x0=c(0,0),fun=f)
```

Bayesian Inference

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

$$P(\text{Parameter}|\text{Data}) = \frac{P(\text{Data}|\text{Parameter})P(\text{Parameter})}{P(\text{data})}$$

Bayesian Inference

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

Likelihood Prior Probability

↓ ↓

$$P(\text{Parameter}|\text{Data}) = \frac{P(\text{Data}|\text{Parameter})P(\text{Parameter})}{P(\text{data})}$$

↑

Posterior Probability

Bayesian Inference

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

Likelihood Prior Probability

↓ ↓

$$P(\text{Parameter}|\text{Data}) = \frac{P(\text{Data}|\text{Parameter})P(\text{Parameter})}{P(\text{data})}$$

↑

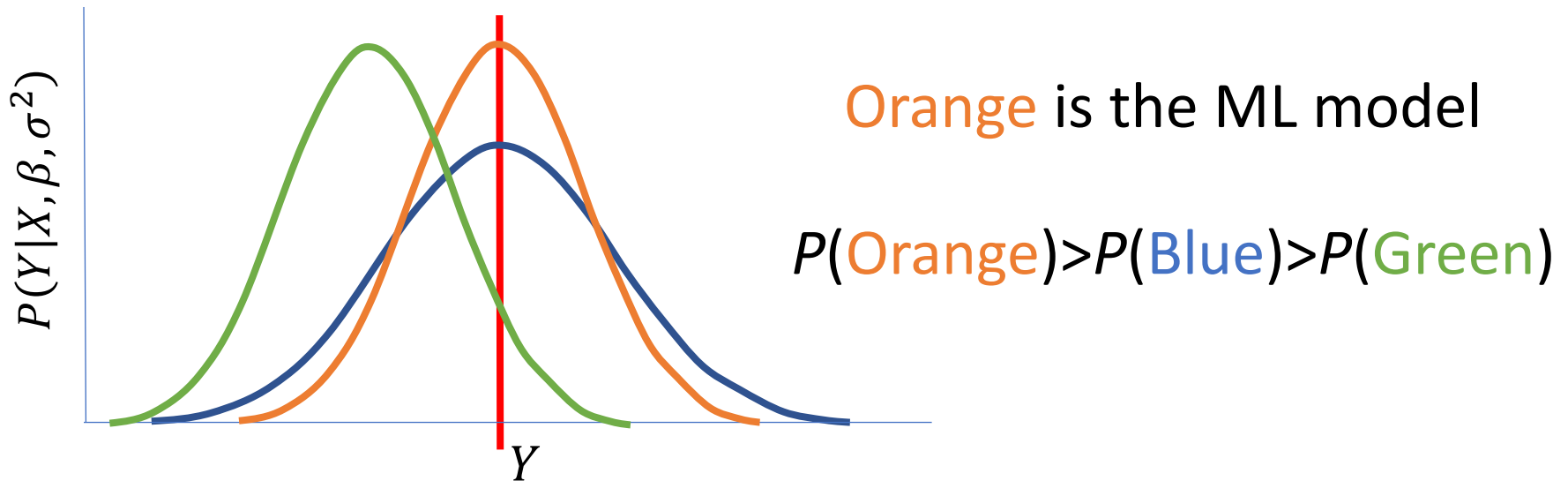
Posterior Probability

$$\propto P(\text{Data}|\text{Parameter})P(\text{Parameter})$$

Bayesian Inference

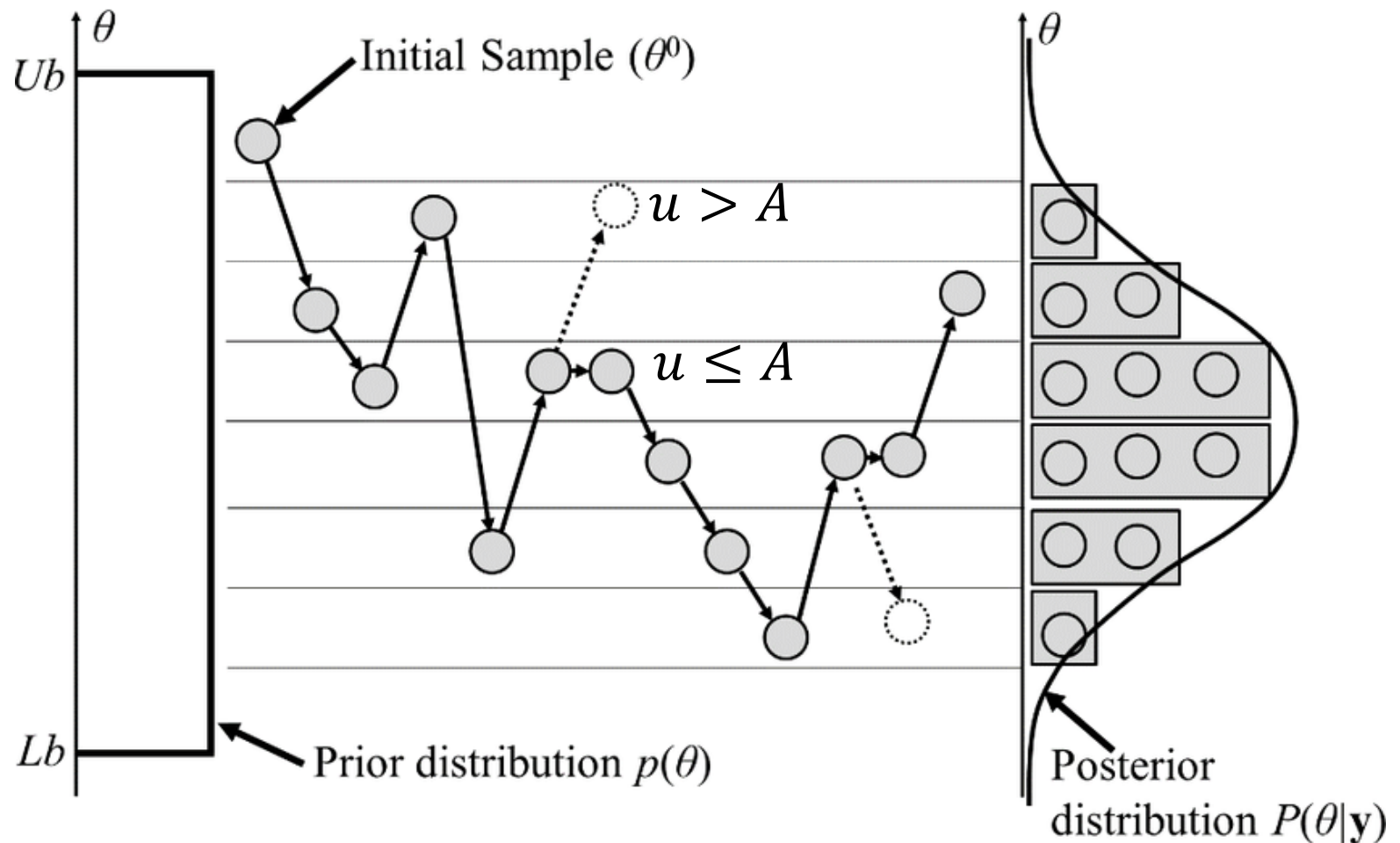
$$\text{Likelihood: } P(Y|X, \beta, \sigma^2) = N(X\beta, \sigma^2)$$

$$\text{Posterior: } P(\beta, \sigma^2 | Y, X) \propto N(X\beta, \sigma^2) P(\beta, \sigma^2)$$



MCMC: Metropolis–Hastings algorithm

$$R = \min \left(1, \frac{L(x') P(x')}{L(x) P(x)} \frac{g(x|x')}{g(x'|x)} \right)$$

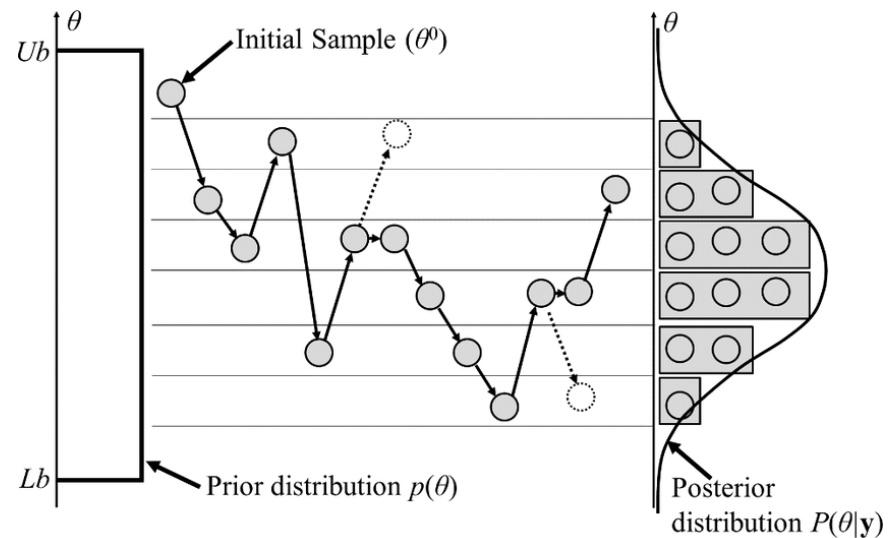


Metropolis–Hastings algorithm

$$L = \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}}, \quad \beta \sim N(0,10), \quad \sigma^2 \sim \text{Gamma}(0.1,0.1)$$

Initialization:

```
pr <- function (beta,sigma) {  
  dnorm(beta, 0, sqrt(10), log=T)+dgamma(sigma, 0.1, 0.1, log=T)  
}  
beta <- rnorm(1, 0, sqrt(10))  
sigma <- rgamma(1, 0.1, 0.1)  
ll <- -f(c(beta,sigma),X,Y)  
lp <- pr(beta,sigma)  
beta_up <- 1  
sigma_up <- 1
```



Metropolis–Hastings algorithm

Start MCMC:

updating beta:

```
niter <- 5000
for (i in 1:niter) {
  beta_new <- beta +
    runif(1, -beta_up, beta_up)
  ll_new <- -f(c(beta_new, sigma), X, Y)
  lp_new <- pr(beta_new, sigma)
  R <- ll_new + lp_new - ll - lp
  if (R > runif(1, 0, 1)) {
    beta <- beta_new
    ll <- ll_new
    lp <- lp_new
  }
}
```

updating sigma:

```
u <- runif(1, 0, 1)
sigma_new <- sigma *
  exp(sigma_up * (u - 0.5))
ll_new <- -f(c(beta, sigma_new), X, Y)
lp_new <- pr(beta, sigma_new)
R <- ll_new + lp_new - ll - lp
R <- R * exp(sigma_up * (u - 0.5))
if (R > runif(1, 0, 1)) {
  sigma <- sigma_new
  ll <- ll_new
  lp <- lp_new
}
}
```