Combiner démographie et génétique en biologie des populations

Lucas Rey
Collaboration with Camille Coron, Sophie Donnet, Julien Stoehr (MIA PS), Lucille Marescot (CIRAD Montpellier), Raphaël Leblois and Miguel de Navascués (CBGP Montpellier)

MIA Paris-Saclay, INRAE

22/03/26

- **GOAL:** parameter inference in biology of conservation
- 2 research communities: **demographic** data VS **genetic** data
- demography: find the **census size** $N(t)$ =number of individuals
- genetics: capacity of a population to maintain its genetic diversity, measured through **effective population size** $N_e(t)$
- our approach: **joint model** for demography and genetics

Genetic data

locus	1	2	3	4
individual 1	1	0	0	1
individual 2	0	1	1	1
individual 3	0	1	1	0

- sample of n_{gen} individuals
- genotype at L loci
- two possible alleles $0/1$ at every locus
- data $N^{A_1 \dots A_L}$ = number of individuals with genome $A_1 \dots A_L$

typically: ≈ 20 loci, ≈ 5 alleles, $n_{gen} \approx 50$

Demographic data

date	01/03	02/03	01/06	02/06	01/09	02/09
individu 1	1	0	0	0	0	0
individu 2	0	0	1	1	0	1
individu 3	1	1	0	1	1	0

- at times t_i, $i = 1, ..., 6$, every individual alive is captured with probability p_i (**unknown**) and marked.
- **capture history** of an individual $\omega = \{1001\}$: $\omega_i = 1 \iff$ individual captured at i
- data $N_\omega =$ number of individuals with given capture history

typically: 5 years, 5 sessions per year

Figure: Bulbul de Cabanis, Kenya

The computation-intensive approach

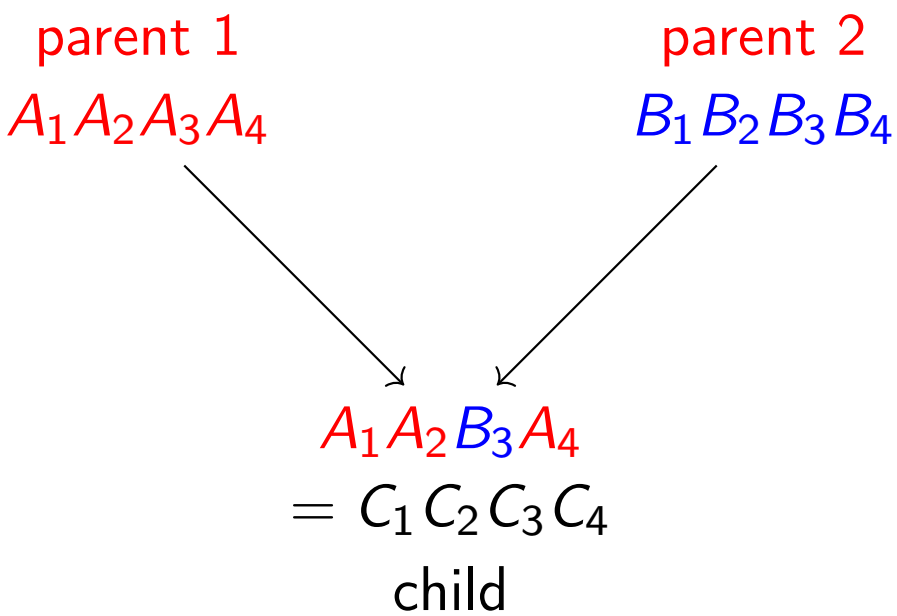
Algorithm (Approximate Bayesian Computing, Tavaré et al., 1997; Beaumont et al., 2002)

- *compute summary statistics* $\mathcal{D} = (\mathcal{D}_{gen}, \mathcal{D}_{dem})$ *from the data*
- *sample* θ_k *according to a prior* $\pi(\theta)$
- *simulate* $\mathcal{D}_k$ *from the model with parameter* θ_k
- *keep* θ_k *iff* $|\mathcal{D}_k - \mathcal{D}| < \varepsilon$

- approach of our collaborators (Miguel de Navascués and Raphaël Leblois): use SLIM to simulate and apply ABC
- our (complementary) approach: understand better the properties of the model; maybe later combine ABC with likelihood-based approaches

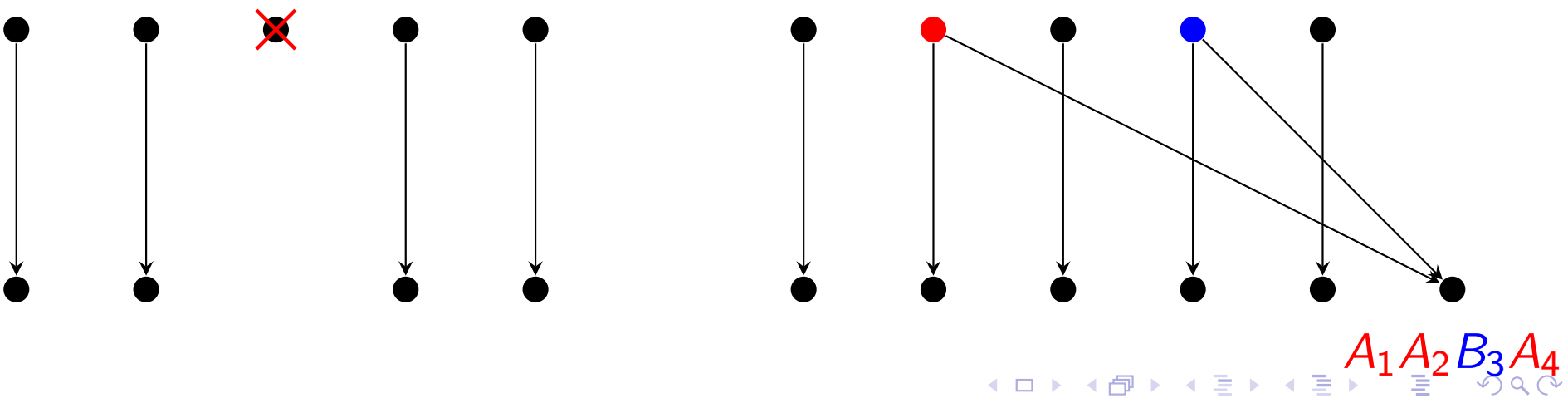
Genetic transmission

- genotype: $A_1 \dots A_L$
- loci are assumed **independant** (e.g. on different chromosomes)
- when two parents $A_1A_2A_3A_4$ and $B_1B_2B_3B_4$ have a child: for every allele, child takes an allele at random

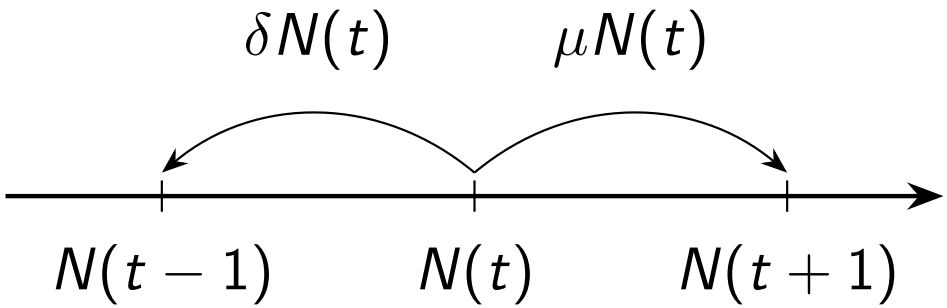


Joint model

- parameters: $N(0), \mu, \delta$
- $N(t) =$ number of individuals at time t
- with rate $\delta N(t)$, an individual is chosen at random and dies
- with rate $\mu N(t)$, two individuals are chosen at random and reproduce



- $N(t)$ = number of individuals is a **birth and death process**



- $(N(t), (X^{A_1 \dots A_L})_{A_1 \dots A_L})$ = continuous time Markov chain
- genetics resembles a **Wright-Fisher** or **Moran** model with recombination, but with varying population size

Remark

We loose some information by forgetting the pedigree and the life history of individuals

Diffusion limit

When $N(0)$ is large, time scale of $\asymp 1$ generation: **diffusion approximation**

Proposition

- $N(t+h) \approx N(t) + h(\mu - \delta)N(t) + h(\mu + \delta)N(t)\mathcal{N}(0,1)$
- *at a given locus and alele* $X(t) = X^{A_i}(t)$ *is approximately a Wright-Fisher diffusion*

$$X(t+h) \approx X(t) + \sqrt{\frac{\mu+\delta}{N(t)} X(t)(1-X(t))\mathcal{N}(0,h)}$$

- *correlation between loci vanish*

$$D^{A_iA_j}(t) = X^{A_iA_j}(t) - X^{A_i}(t)X^{A_j}(t) \approx 1/N(t)$$

$N_e(t) = \frac{\mu+\delta}{N(t)}$ is called the **effective population size**

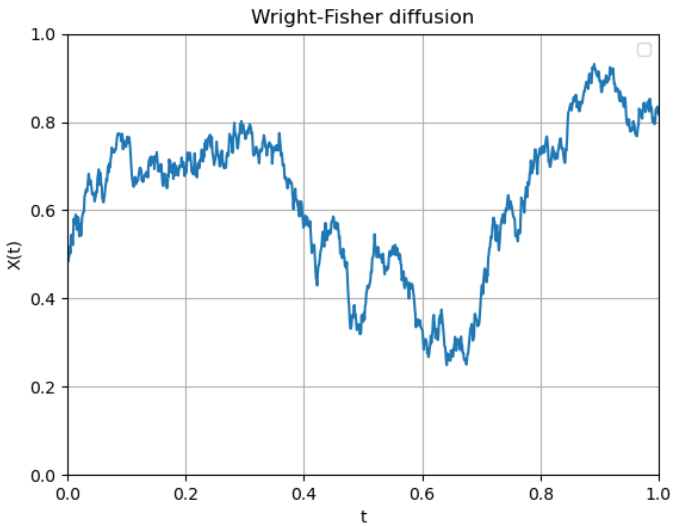
Wright-Fisher diffusion

If $N(t) = N$ was constant, in the diffusion limit $N = N(0) \to \infty$, the reparameterized process $\tilde{X}(t) = X(Nt)$ converges

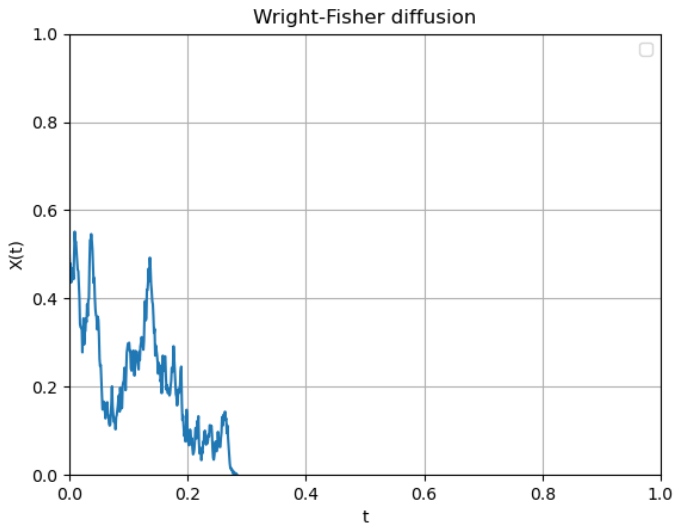
Definition (Wright-Fisher diffusion)

$$d\tilde{X}(t) = \sqrt{(\mu + \delta)\tilde{X}(t)(1 - \tilde{X}(t))}dB_t$$

$\mu + \delta = 1$



$\mu + \delta = 10$



Some remarks

- to obtain the limit of $X(t) = X^{A_i}(t)$, one just looks at

$$\mathbb{E}[X(t+h) - X(t)] \approx 0$$

$$\mathrm{Var}[X(t+h) - X(t)] \approx \frac{\mu + \delta}{N(t)} X(t)(1 - X(t))$$

- to obtain joint frequencies $X^{A_i A_j}$, one proves that the **linkage disequilibrium**

$$D^{A_i A_j}(t) = X^{A_i A_j}(t) - X^{A_i}(t) X^{A_j}(t) \approx 1/N(t)$$

remains small

- slow-fast analysis in the spirit of Ethier-Nagylaki (1989)

$$\mathbb{E}[D(t+h) - D(t)] \approx -\frac{h}{4} dt \; ; \; \mathrm{Var}[D(t+h) - D(t)] = o(h)$$

Model for genetic sampling:

- at time 0 and 1, sample n_{gen} individuals, $\tilde{X}^{A_i}$ = observed frequency
- **summary statistics**: $\mathcal{D}_{gen} = (\tilde{\Delta}^{A_i}, \tilde{D}^{A_i A_j})$

$$\tilde{\Delta}^{A_i} = \tilde{X}^{A_i}(1) - \tilde{X}^{A_i}(0) \text{ frequency variation}$$
$$\tilde{D}^{A_i A_j} = \tilde{X}^{A_i A_j}(0) - \tilde{X}^{A_i}(0)\tilde{X}^{A_j}(0) \text{ linkage disequilibrium}$$

Proposition

$$\tilde{\Delta}^{A_i} \approx \mathcal{N}\left(0, \left(\frac{1}{n_{gen}} + \frac{\mu+\delta}{N(0)}\right) X^{A_1}(0)(1 - X^{A_1}(0))\right)$$

are independent between loci thus a maximum likelihood estimator gives an estimator $N_e^{dem} = \hat{N}_e(\mathcal{D}_{gen})$

Model for demographic sampling:

- capture occasions at times $t_1 = 0, t_2, \ldots, t_6 = 1$
- at time t_i, an individual is captured with proba p_i (= nuisance parameter)
- every capture history ω is seen N_ω times
- summary statistics $\mathcal{D}_{dem} = (N_\omega)_{\omega \in \{0,1\}^6}$

Remark

From $\mathcal{D}_{dem}$ one obtains $\hat{\mu}, \hat{\delta}, \hat{N}(0), \hat{p}_i$ thus $\hat{N}_e^{dem} = \frac{\hat{\mu}+\hat{\delta}}{\hat{N}}$

demographic sampling has an explicit likelihood, but details of estimation are still to be worked out, $N(t_i)$ are latent variables

Our roadmap:

Conjecture

The demographic and genetic estimator are approximately independant $\hat{N^e}_{dem} \perp \hat{N^e}_{gen}$

- find the best combined estimator $\hat{N}^e_{opt} = \lambda \hat{N}^e_{dem} + (1-\lambda)\hat{N^e}_{gen}$
- need to estimate the variances, either from computation or data (bootstrap)

Merci !