Genealogies with recombination in spatial population genetics

Luminy, 19/06/15

We'll be interested in a population with a spatial structure and in how space influences the genetic evolution/diversity of the population. As we have seen in earlier talks, there are several evolutionary forces that can influence this evolution, and it is important to understand the signature left by each to be able to detect how important each of them is.

Here we'll concentrate on the neutral case, in which no alleles gives an advantage to its owners, not at least because this is a sound null model to compare to, but also because it enables us to highlight the main features of the spatial structure.

In what follows we'll be talking about a population of plants essentially (no migration, though we could include it). We'll try to describe the correlations between the genetic diversities of 2 linked loci with recombin by looking at how correlated their ancestries are, and next we'll use the same kind of ideas to infer the parameters summarizing the evolution of the population in the simplest case of purely local reproduct

I The model
II Genealogies when large-scale extinction/recolonization events
III Perspectives on inference.

(1) The model

We'll work in a spatial continuum, because that's natural for many populations but all we'll see will be valid for discrete subpopulations as well.

Key points : 2 linked loci A/B
population homogeneously filling $\mathbb{R}^2$ (uniform density with "infinitely many individuals everywhere")

L = parameter scaling space (will tend to ∞).

Evolution given by 2 ⊥ PPP :

- Π_S ("small events") on $\mathbb{R}\times\mathbb{R}^2$ with intensity $dt\otimes dx$. We fix $u_S \in]0,1]$, $R_S>0$

  If $(t,x)\in\Pi_S$, reproduction event occurs in $B(x,R_S)$ at time t

  → choose 2 individuals ⊥ and uniformly at random within $B(x,R_S)$ → alleles (A,B) and (a,b)

  → $\forall y\in B(x,R_S)$, kill a fraction u_S of the pop at site y

  ⋯ and replace by $(1-r_L)u_S$ non recombinants (A,B) and (a,b) in equal proportions

  plus $r_L u_S$ recombinants (A,b), (a,B) in equal proportions.

  r_L is the fraction of recombinant offspring ~~and we suppose that~~, will depend on L

- Π_B ("big events") on $\mathbb{R}\times\mathbb{R}^2$ with intensity $\rho_L^{-1}dt\otimes dx$. We fix $\alpha>0$, $R_B>0$ and $u_B\in]0,1[$

  If $(t,x)\in\Pi_B$, reproduction event occurs in $B(L^\alpha x, L^\alpha R_B)$.

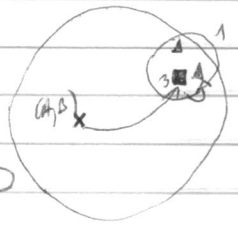
  → choose 1 individual $\overrightarrow{(A,B)}$ unif. at random within $B(xL^\alpha, L^\alpha R_B)$

  → $\forall y\in B(L^\alpha x, L^\alpha R_B)$, kill a fraction u_B of indiv. at y and replace by fraction u_B of individuals (A,B).

No recombination during big events, but could be generalized.
Many other ways to generalize this model !!

We assume that $\rho_L \xrightarrow[L\to\infty]{} +\infty$ with $\frac{\rho_L}{L^{2\alpha}} \xrightarrow[L\to\infty]{} c \in [0,\infty)$.

↳ Regime where large events do have an impact.



II

Suppose we sample 2 individuals at distance L^β for some $\beta > 0$
↳ (A,B) and (a,b)

Q: how correlated are their genealogies at each locus ?

More precisely, let us consider 1 individual and 1 locus first. (Say A).

If we look backward in time, starting from sampling time, there will be a first reproduction event during which our individual was among the offspring of a parent chosen there. Therefore, if we defin the ancestral line/lineage of our individual as the function of time that traces back the location of the ancestors of the individu at locus A, at the time of this first event in the past the ancest line jumps to the location of the parent, which uniformly distributed over the area of the event. We carry on going back into the past and find another reproduction event in which this ancestor was among the offspring created and so the ancestral line jumps to a new location uniformly distributed over the area of this event, and so on. With our construction, the ancestral line is a jump process (finite rate) which makes

jumps of size $O(1)$ at rate $O(1)$ and jumps of size $O(L^\alpha)$ at rate $O(\rho_L^{-1})$. All the same for the ancestral lines corresponding to the other individual, and both lines are a priori correlated. Maybe at some point the two will emanate from the same parent during the same reproduction event, in which case they coalesce.

τ_{Aa}^L = coal. time of the ancestral lines of A & a

τ_{Bb}^L = B & b

Effect of recombination: initially the lines of A and B start at the same location (they correspond to the same individual), but they get separated into 2 uniformly sampled locations if the ancestor belongs to the fraction $\sim r_L$ of recombinants. Then they are close together and so may coalesce again very soon

$\Rightarrow$ a priori τ_{Aa}^L and τ_{Bb}^L are correlated, but how strong is this correlation?

Th 1 (EVI2) Suppose that $\lim_{L\to\infty} \frac{\log(1+\frac{\log\rho_L}{r_L\rho_L})}{2\log L} \le \beta - \alpha$

Then $\forall t \ge \beta$, $\lim_{L\to\infty} \left[\tau_{Aa}^L \wedge \tau_{Bb}^L > \rho_L L^{2(t-\alpha)}\right] = \left(\frac{\beta-\alpha}{t-\alpha}\right)^2$

$= \mathbb{P}(\tau_{Aa}^L > \rho_L L^{2(t-\alpha)})^2$

Ex: $r_L \equiv r \in (0,1)$

Th 2 (EV 12) Suppose there exists $\gamma > \beta$ such that

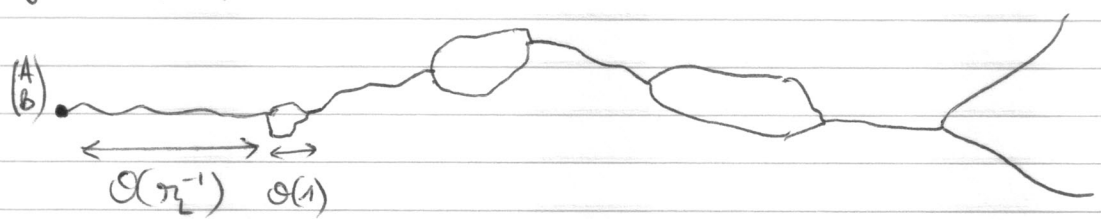
$$\lim_{L\to\infty} \frac{\log\left(1 + \frac{\log \rho_L}{r_L \rho_L}\right)}{2 \log L} = \gamma - \alpha$$

(i) $\forall t \in [\beta, \gamma]$, $\mathbb{P}\left(\tau^L_{Aa} \wedge \tau^L_{Bb} > \rho_L L^{2(t-\alpha)}\right) \to \frac{\beta - \alpha}{t - \alpha}$

(ii) $\forall t > \gamma$, $\mathbb{P}\left(\tau^L_{Aa} \wedge \tau^L_{Bb} > \rho_L L^{2(t-\alpha)}\right) \to \frac{\beta - \alpha}{(\gamma-\alpha)} \left(\frac{\gamma-\alpha}{t-\alpha}\right)^2$

In the first case, the 2 genealogies are asymptotically independent.
In the second case, there is a first period of complete correlation, and conditionally on seeing no coalescence by time $\rho_L L^{2(t-\alpha)}$, the genealogies become (in the limit) independent.

Idea of the proof:



The lines get well-separated only when one of them is affected by a big event when it is not in the same ancestor as the other.
Can happen only during an excursion of length at least $O(\rho_L)$, which takes $O(\log \rho_L)$ excursions to occur.

Total time to wait: $(\log \rho_L)\, r_L^{-1} + \rho_L = \rho_L\left(1 + \frac{\log \rho_L}{r_L \rho_L}\right)$

then they "quickly" become decorrelated.

Now with 2 individuals, their ancestral lines make jumps of size $\mathcal{O}(L^\alpha)$ at rate $\mathcal{O}(\rho_L^{-1})$ so we'd rather consider $\left(\frac{X_{\rho_L t}}{L^\alpha}\right)_{t\geq 0}$ to obtain something that looks like a finite variance jump process.

If start at distance $L^{\beta-\alpha}$ in these new units, they need a time at least $\mathcal{O}(L^{2(\beta-\alpha)})$ to meet and have a chance to coalesce.

In original units: if $\rho_L\left(1+\frac{\lg \rho_L}{r_L \rho_L}\right) \ll \rho_L L^{2(\beta-\alpha)}$ decorrelation occurs before any coalescence A/B

and if $\rho_L\left(1+\frac{\lg \rho_L}{r_L \rho_L}\right) \approx \rho_L L^{2(\gamma-\alpha)}$ with $\gamma>\beta$, then either

a coalescence occurs before the lines of the same individuals decorrelate ($\rightarrow$ period of full correlation)

or we manage to reach the regime of decorrelation.

III Perspectives on inference (and tests)

$\rightarrow$ We have shown that if we sample the 2 individuals at distance $\gg D_L = \underbrace{L^\alpha}_{\text{radius of the big events}}\sqrt{1+\frac{\lg \rho_L}{r_L \rho_L}}$ ancestries are decorrelated and this distance is somewhat a threshold.

A natural question would be: can we use this result to test the presence of rare but large extinction-recol. events? And can we learn more about them?

$\hookrightarrow$ Requires to sample our pairs of individuals at very large distance larger than the radius of the largest events, even in the case when there are only small events.

→ Sampling at large distances makes us loose a lot of the precise local mecanisms (R_S, R_B, u_S, u_B).

If sample at small distance, very small chance to see an "effective recombination" occur and separate the ancestral lines corresponding to 2 loci of the same individual. But if we consider sufficiently many loci, this chance becomes macroscopic.

→ only small events

In [BEKV13], we show that if we consider a typical large region of the genome at which the 2 individuals have strongly correlated genealogies, then the distribution is approximately geometric with a parameter that depends only on

σ^2 = variance of the motion of a single line in 1 unit of time

and $\mathcal{N}$ = neighborhood size (inverse of the local coalescence rate)

↔ gives some hope to be able to infer the local summary statistics σ^2, $\mathcal{N}$ though still a lot to do in this direction.