How do geographic distances translate into genetic distances ?

Emmanuel Scherzer. Joint work with V. Miro Pina.

September 7, 2017

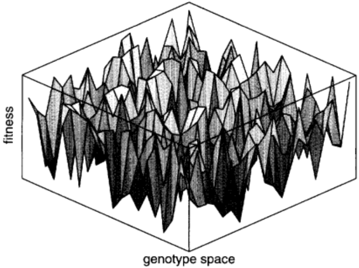
Part 1: Biological motivation

Speciation

- Speciation: when two subpopulations accumulate enough genetic differences, they become genetically incompatible.
  1. Pre-zygotic isolation: preferential mating.
  2. Post-zygotic isolation: hybrid depression.
- It is well established (e.g., Malécot) that geographic structure affects the genetic diversity of a population.
- We aim at modeling the genetic divergence of populations in a structured population.
- General question: Under which geographical conditions can a species remain genetically coherant ? or at the contrary, under which conditions can speciation occur ? how long does it take ?

How do populations diverge (I) ? Rugged fitness landscape

- Fitness landscape: each genotype gets assigned a fitness value.
- According to Wright (1931): fitness andscapes should have local adaptive peaks separated by adaptive valleys.
- Adaptive peaks are interpreted as different species
- Adaptive valleys are interpreted as unfit hybrids



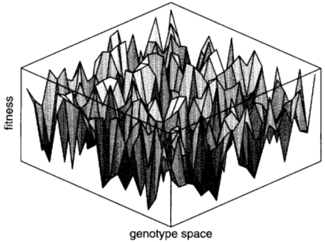
Rugged fitness landscape.

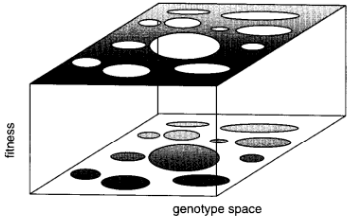
- Speciation occurs when a sub-population goes from one peak to the other.
- Need to pass through a valley.
- Intuitive idea of Wright : founder effect.
- In a small enough population, genetic drift is strong enough to counterbalance the effect of selection.
- Example: Diploid population. Genome only consists of a single locus with two alleles a and A with

$$w_{aa} = 1, w_{aA} = 1 - s, w_{AA} = 1$$

- When $ns = 20$ (say a population size of 200 and a fitness penalty of $s = 0.1$), the probability to cross the valley is approximately 10^{-8} to cross the valley.

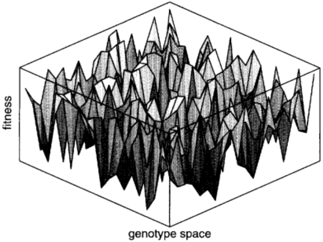
How do populations diverge (II) ? Holey landscape

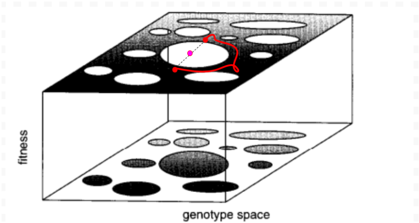




- Alternative topography: Local maxima could be be partitioned into connected sets (or evolutionary ridges)
- Holey landscape: Evolutionary ridges typically have complicated geometry
- Speciation: a population diffuses until it stands at the other side of a hole
- Maynard Smith (1970) : "if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates".

How do populations diverge (II) ? Holey landscape

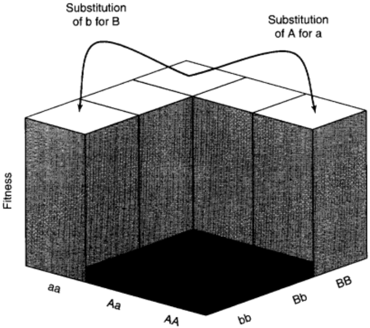




- Alternative topography: Local maxima could be be partitioned into connected sets (or evolutionary ridges)
- Holey landscape: Evolutionary ridges typically have complicated geometry
- Speciation: a population diffuses until it stands at the other side of a hole
- Maynard Smith (1970) : "if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates".

Holey landscape. Dobzhansky model (1967)

- two loci with two alleles aA and bB respectively.
- $w_{aa**} = w_{**BB} = 1$ but any other genotype gets assigned a fitness value $1 - s$.
- Starting from a population $aaBB$, the population can drift in two ways: either to $aabb$ or $AABB$.
- Finally, any recombination of types $aabb$ and $AABB$ produce an unfit individual.



Rugged vs Holey landscape

- Experimental justifaction: Orr (1995) identified pairs of loci on the Drosophilia chromosome suggesting a Dobzhansky-type mechanism.
- Theorectical justification: In high-dimensional genotype space, fitness peaks are typically related by evolutionary ridges.
- Gavrilets and Gravener (1997) used a simple percolation model on the hypercube $\{0,1\}^n$.
- A genome is viable (resp., unviable) with probability p (resp., $1-p$).
- When $p > 1/n$, as $n \to \infty$, the size of the largest viable connected component (or evolutionary ridge) goes to ∞ at a speed $O(p2^n)$.
- The classical NK model exhibits similar behavior in high dimension (quasi-Holey landscape).

General framework to study speciation (Gavrilets 1997, 1998, 2002), Yamagushi, Iwasa (2015)

- Ignore deleterious mutations. In large populations, they are washed away by selection at the micro-evolutionary scale.
- Describe the dynamics on the evolutionary ridge as neutral (Any genotypes on the ridge can be accessed by single-mutation neutral steps)
- Evolutionary dynamics along an evolutionary ridge is assumed to be slow. Along the evolutionary ridge, random mutations are very likely to be deleterious.

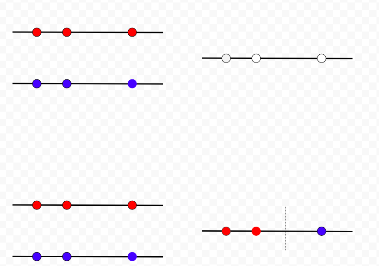
Part II: Individual based model, Main results

Individual based model

- Multi-locus Moran model with mutation and migration.
- Structured population: pop. is subdivided into N subpopulations. Island i is composed by n_i individuals.
- Each individual is identified with a chromosome of size 1.
- $l = \#$ of Loci responsible for speciation.
- loci are distributed uniformly along the chromosome.

An underlying individual based model

- Reproduction: haploid Moran model with recombination
  - → Each ind. reproduces at rate 1, chooses a random partner.
  - → Their offspring replaces a randomly chosen ind.
  - → Recombination: Offspring is a obtained by pasting together fragments of the parents chromosomes.
  - → Number of cross-overs follows Poisson(λ)

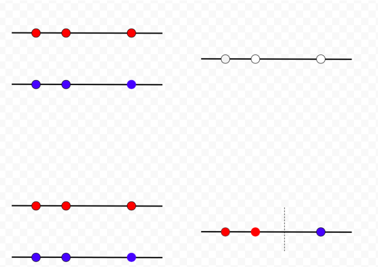


An underlying individual based model

- Reproduction: haploid Moran model with recombination

- Mutation at rate u per individual per locus (infinite allele model).

- Migration $i \to j$, at rate m_{ij}. A copy of one random individual in i migrates from i to j, and replaces an individual chosen uniformly at random in population j.

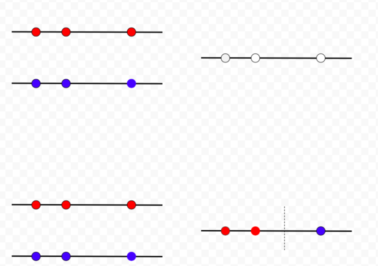


An underlying individual based model

- Reproduction: haploid Moran model with recombination

- Mutation at rate u per individual per locus (infinite allele model).

- Migration $i \to j$, at rate m_{ij}. A copy of one random individual in i migrates from i to j, and replaces an individual chosen uniformly at random in population j.



  - ▶ migration tends to reduce the genetic distances between subpopulation (homogenization effect)
  - ▶ mutation tends to increase distances

Scaling limit

We will consider the following regime

$$u\ ,\ m_{i,j} \qquad \underbrace{<<}_{\text{low mut.–migr.}} \qquad \frac{1}{n_i}\ ,\ \frac{1}{l} \qquad \underbrace{<<}_{\text{large pop.}} \qquad 1$$

- In the usual so-called weak limit regime (structured Kingman coalescent – Wright-Fisher diffusion with mutation-migration), it is assumed that

$$m_{i,j},\ u = O(\epsilon), \quad 1/\epsilon \text{ is a typical population size}$$

- In the weak limit regime,at a given locus, there is a non-trivial polymorphism at the intra-population level.
- Here, we assume that mutation events and migration events are rare so that intra-population diversity can be negligated at the limit.
- This will allow to approximate our IBM model by a PBM.
- Rationale: Along the evolutionary ridge, changes occur at the macro-evolutionary time-scale

In order to implement the regime

$$u \ , \ m_{i,j} \quad \underbrace{<<}_{\text{low mut.–migr.}} \quad \frac{1}{n_i} \ , \ \frac{1}{l} \quad \underbrace{<<}_{\text{large pop.}} \quad 1$$

We assume that the parameters of our model $(n_i, m_{i,j}, u, l)$ depend on two scaling factors (ϵ, γ) with

$$\begin{cases} n_i \equiv n_i^\epsilon & \text{with } \epsilon n_i^\epsilon \to N_i \\ l \equiv l^\epsilon & \text{with } l^\epsilon \to \infty \end{cases}$$

($1/\epsilon$ typical size of a population, l^ϵ typical number of loci involved in speciation) and

$$\begin{cases} m_{i,j} \equiv m_{i,j}^\gamma & \text{with } \frac{1}{\gamma} m_{i,j}^\gamma \to M_{i,j} \\ u \equiv u^{\gamma,\epsilon} & \text{with } \frac{1}{\epsilon\gamma} u^{\gamma,\epsilon} \ \to \ U_\infty \end{cases}$$

(γ typical rate of migration)

Then we let succesively γ and then ϵ go to 0 (so that $\epsilon >> \gamma$).

Note that and $u/m_{i,j} = O(\epsilon)$ (balance mutation/migration).

Distance between islands

- We aim at describing the genetic distance between islands.
- When $\epsilon >> \gamma$, sub-populations are typically monomorphic.
- When island i and j are monomorphic, define

$$d_t^{\epsilon,\gamma}(i,j) = \frac{1}{l}\#\text{segregating loci between island } i \text{ and } j \text{ at time } t.$$

- (otherwise take the average number of segregating sites between two randomly sampled individuals)

- The genetic distance between two populations evolve when one or several alleles fixate in the a population following a mutation or migration event.
- Since those events are rare, we accelerate time by $1/\gamma\epsilon$

Theorem 1 (Miro Pina, S.)

When island i and j are monomorphic, define

$$d_t^{\epsilon,\gamma}(i,j) \;=\; \frac{1}{l}\#\text{segregating loci between island } i \text{ and } j \text{ at time } t.$$

For every i, j, there is a deterministic process $(D_t(i,j); t \geq 0)$ s.t.:

$$\lim_{\epsilon\to 0}\lim_{\gamma\to 0}\left(d_{t/\gamma\epsilon}^{\epsilon,\gamma}(i,j);\ t\geq 0\right) \;=\; \left(D_t(i,j);\ t\geq 0\right) \text{ in distribution (in the weak topology).}$$

Moreover $\lim_{t\to\infty} D_t(i,j) = 1 - \mathbb{E}(e^{-2U_\infty \tau_{ij}})$, where

$$\tau_{ij} = \inf\{t \geq 0 \ :\ S^i(t) = S^j(t)\},$$

and where S^i and S^j are two independent random walks on $\{1, \cdots, N\}$ starting respectively from i and j and whose transition rate from k to l is given by

$$\tilde{M}_{kl} \;:=\; \frac{M_{lk}}{N_k} \quad \text{for every } k, l \in \{1, \cdots, N\}.$$

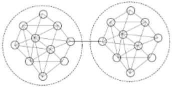
Example: Geographic Bottleneck

- Two complete graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with N vertices.
- $v_1 \in \mathcal{G}_1$, $v_2 \in \mathcal{G}_2$, $v_1 \sim v_2$.
- For $i \sim j$, $M_{i,j} = \frac{1}{N}$.
- $U_\infty = \frac{c}{N}$ for some $c > 0$.

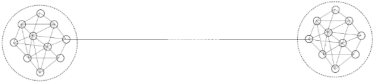
Proposition

Then for any two neighbours $i, j \in \mathcal{G}$

$$1 - \mathbb{E}\left(\exp(-2U_\infty \tau_{ij})\right) = \begin{cases} \frac{c}{1+c} + o(1) & \text{if } i,j \in \mathcal{G}_1, \text{ or if } i,j \in \mathcal{G}_2 \\ 1 - \frac{1}{N} + o(\frac{1}{N}) & \text{if } i = v_1 \text{ and } j = v_2 \end{cases}$$



(a) Geographic distances



(b) Genetic distances

Part III: Idea of the proof

A population based model

- Since u , $m_{i,j}$ $<<$ $\frac{1}{n_i}$, $\frac{1}{s}$, intra-subpopulation diversity can be neglected .
- As $\gamma \to 0$ (ϵ fixed): Mutiscale Moran model. Slow dynamics at the inter-population level. Fast dynamics at the intra-population level.

This allows to approximate the IBM by the following population based model (PBM).

A population based model

When $\gamma \to 0$ (scaling parameter for mutation and migration) and ϵ remains fixed, each island is represented by a single chromosome indexed from $\{1, \cdots, N\}$. Two types of transition:

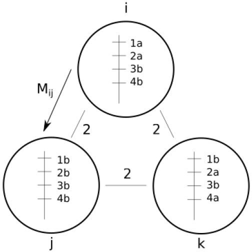
→ *Mutation* For every island i, locus k, fix a mutation at rate U_∞.

→ *Migration*

1. Start with 1 migrant individual in a monomorphic resident population of size n_j^ϵ. Define $\mathcal{F}_j^\epsilon$ to be the random set of loci at which the migrant allele fixates.
2. At rate $\frac{1}{\epsilon} M_{ij}$, fixate the migrant alleles (island i) in resident population (island j) at a random set of loci, where the random set of loci is distributed as $\mathcal{F}_j^\epsilon$.

Genetic partition

- As $\gamma \to 0$, the IBM converges to the PBM (indexed by the inverse population size ϵ).
- In the PBM, at every locus $k \in \{1, \cdots, l\}$, types induce a partition of the meta-population denoted by $\Pi_k^\epsilon(t)$:



$$\Pi_1^\epsilon(t) = \{i\}\{j,k\}$$

$$\Pi_4^\epsilon(t) = \{i,j\}\{k\}$$

The genetic partition vector $\Pi^\epsilon(t) = (\Pi_m^\epsilon(t);\ m \in \{1, \cdots, l\})$ describes the genetic composition of the population at time t.

Some properties of the genetic partition vector

- For every $k \in \{1, \cdots, l\}$, $(\Pi_k^\epsilon(t);\ t \geq 0)$ is a Markov process on the set of partitions.

(mutation) island i is singled out at rate U_∞ (i takes on a new type).

(migration) with rate

$$M_{i,j} \times \frac{1}{\epsilon n_j^\epsilon}$$

displace j in the block containing i (j inherits the type of i)

- Stationarity: For every $m \leq n$, Π_m^ϵ is identical in law to Π_n^ϵ.
- Non trivial correlation between loci: a single migration event has an impact on several loci simultaneously.
- Cornerstone of the approach: ergodic theorem along the sequence when $\epsilon \to 0$.
- For all $\Pi \in (\mathcal{P}_N)^l$, $X(\Pi) \ = \ \frac{1}{l} \sum_{k \leq l} \delta_{\Pi_k}$, is the empirical measure associated to the "sample" $\Pi_1, \cdots, \Pi_l$. In the following,

$$\xi_t^\epsilon \ = \ X(\Pi^\epsilon(t))$$

Ergodic theorem along the chromosome

Theorem 2 (Miro Pina, S.)

Assume $\exists\ P^0 \in \mathcal{M}_N$ s.t. $X(\Pi^\epsilon(0)) \underset{\epsilon\to 0}{\longrightarrow} P^0$. Then

$$(\xi^\epsilon_t;\ t \geq 0) \underset{\epsilon\to 0}{\Longrightarrow} (P_t;\ t \geq 0) \text{ in distribution in the weak topology,}$$

where P is a deterministic probability measure on the space of partitions. More precisely, P solves the forward Kolmogorov equation associated to a one-locus Moran model, i.e.,

$$\frac{d}{ds} P_s \ = \ {}^t G P_s$$

with initial condition $P_0 = P^0$, where G is the generator describing the dynamics of the partition at an arbitrary locus on the chromosome.

Proof of Thm 1 based on Thm 2

- Define $d_t^\epsilon = \frac{1}{I^\epsilon}$#segregating loci between i and j at time t the genetic distance in the PBM. Then

$$\begin{aligned} d_t^\epsilon(i,j) &= \frac{1}{I}\sum_{k=1}^{I} 1_{i \not\sim_{\Pi_k(t)} j} \\ &= \xi_t^\epsilon(\{\pi \in \mathcal{P}_N : \ i \not\sim_\pi j)\}) \end{aligned}$$

- By Theorem 2, $d_t^\epsilon(i,j) \to P_t(\{\pi \in \mathcal{P}_N : \ i \not\sim_\pi j)\})$.
- Finally,

$$P_t\left(\{\pi \in \mathcal{P}_N : \ i \not\sim_\pi j)\}\right) = \ 1 \ - \ \mathbb{E}\left(\exp(-2U_\infty \tau_{ij})\right)$$

using a standard duality principle.

Thank you !