The Genealogy of Populations Undergoing Selection

by Jason Schweinsberg
University of California at San Diego

Outline

1. Introduction to coalescent processes
2. The model and previous work
3. Main results
4. Further remarks and open problems

Mathematical Population Genetics

Study mathematical models of evolving populations.

By comparing predictions of models to observations, draw inferences about how populations evolve, causes of genetic variability.

Moran Model (Moran, 1958):

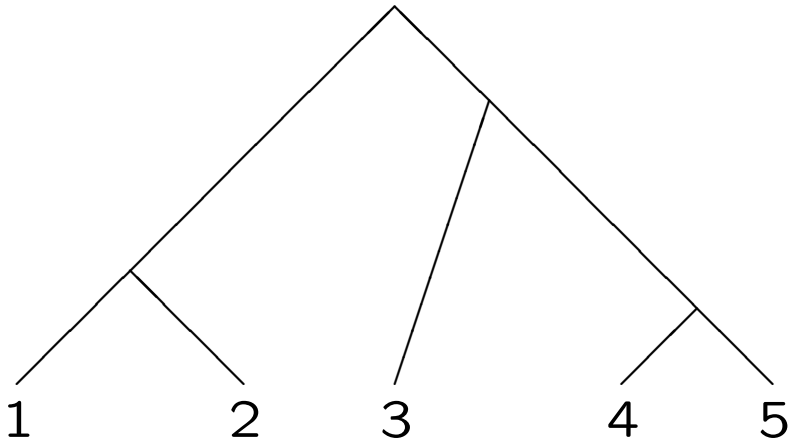
- The population has fixed size N.
- Each individual independently lives for Exponential(1) time.
- When an individual dies, a new individual is born. Its parent is chosen uniformly at random from the population.

Coalescent Processes

Sample n individuals at random from a population. Follow their ancestral lines backwards in time. The lineages coalesce, until they are all traced back to a common ancestor.

Represent by a stochastic process $(\Pi(t), t \geq 0)$ taking its values in the set of partitions of $\{1, \ldots, n\}$.

Kingman's Coalescent (Kingman, 1982): Only two lineages merge at a time. Each pair of lineages merges at rate one.

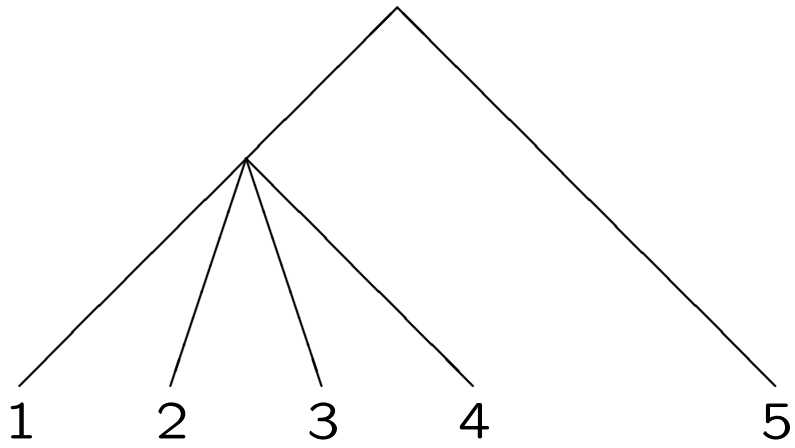


Kingman's coalescent describes genealogy of Moran model.

Bolthausen-Sznitman coalescent

When there are b lineages, each k-tuple ($2 \leq k \leq b$) of lineages merges at rate

$$\lambda_{b,k} = \int_0^1 p^{k-2}(1-p)^{b-k}\, dp.$$



Bolthausen-Sznitman (1998): Ruelle's probability cascades

Bovier-Kurkova (2007): Derrida's GREM

Poisson process construction

Consider a Poisson point process on $[0, \infty) \times (0, 1]$ with intensity

$$dt \times p^{-2}\, dp.$$

Begin with n lineages at time 0. If (t, p) is a point of this Poisson process, then at time t, there is a merger event in which each lineage independently participates with probability p.

This leads to

$$\lambda_{b,k} = \int_0^1 p^{k-2}(1-p)^{b-k}\, dp.$$

Rate of mergers impacting more than a fraction x of lineages is

$$\int_x^1 p^{-2}\, dp = \frac{1-x}{x}.$$

Branching Brownian motion with absorption

Begin with particles in $(0, \infty)$. Each particle independently moves according to one-dimensional Brownian motion with drift $-\mu$,

$$\mu = \sqrt{2 - \frac{2\pi^2}{(\log N + 3 \log\log N)^2}}.$$

Each particle splits into two at rate 1. Particles are killed if they reach the origin.

particles	$\to$	individuals in the population
positions of particles	$\to$	fitness of individuals
branching events	$\to$	births
absorption at 0	$\to$	deaths of unfit individuals
movement of particles	$\to$	changes in fitness over generations

Bolthausen-Sznitman coalescent gives genealogy of particles.

Heuristic argument: Brunet, Derrida, Mueller, Munier (2006)

Rigorous argument: Berestycki, Berestycki, Schweinsberg (2013)

Another Model

The population has fixed size N.

Each individual independently acquires mutations at times of a rate μ_N Poisson process. (μ_N = mutation rate)

Mutations are beneficial. An individual with j mutations (called "type j") at time t has fitness

$$\max\{1 + s_N(j - m_N(t)),\, 0\},$$

where $m_N(t)$ is the average number of mutations of the N individuals at time t. (s_N = selective benefit from a mutation)

Each individual independently lives for an exponential(1) time.

When an individual dies, parent of new individual is chosen with probability proportional to fitness.

Note: In branching Brownian motion with absorption, all individuals have the same birth rate, but individuals with low fitness are killed. Here, all individuals have the same death rate, but individuals with higher fitness have a higher birth rate.

Questions of Interest

1. Speed of evolution: how fast does $m_N(t)$ increase?

2. What is the distribution of fitnesses of individuals at a given time?

3. How can we describe the genealogy of the population?

Previous Non-Rigorous Work

Detailed non-rigorous work has been done on this model:

Rouzine, Wakeley, and Coffin (2003)
Desai and Fisher (2007)
Beerenwinkel et. al. (2007)
Brunet, Rouzine, and Wilke (2008)

They obtained precise estimates on the speed of evolution. They concluded that the distribution of fitnesses of individuals at a fixed time t is Gaussian, leading to a "Gaussian traveling wave."

Neher and Hallatschek (2013) and Desai, Walczak, and Fisher (2013) argued that the genealogy of the population is given by the Bolthausen-Sznitman coalescent.

Goal: For some range of μ_N and s_N, obtain rigorously the speed of evolution, the Gaussian shape for the distribution of fitnesses, and the genealogy.

Previous Rigorous Work

If $s_N = s > 0$ and $\mu_N \ll 1/(N \log N)$, only one beneficial mutation at a time. Exponental($N\mu_N s$) times between selective sweeps.

Durrett and Mayberry (2011) consider the case with $s_N = s > 0$ and $\mu_N \sim N^{-\beta}$, where $0 < \beta < 1$. They rigorously obtained the speed of evolution and distribution of fitnesses at a fixed time. Only finitely many types present in the population at once.

Yu, Etheridge, and Cutherbertson (2010) considered similar model with $s_N = s > 0$ and $\mu_N = \mu > 0$. They showed that for all $\delta > 0$,

$$\frac{E[m_N(t)]}{t} \geq (\log N)^{1-\delta}$$

for sufficiently large N.

Kelly (2013) studied model of Yu, Etheridge, and Cuthbertson and showed that

$$\limsup_{t\to\infty} \frac{E[m_N(t)]}{t} \leq \frac{C \log N}{(\log\log N)^2}.$$

Assumptions

1. $\lim_{N\to\infty} \dfrac{\log N}{\log(s_N/\mu_N)\log(1/s_N)} = \infty.$

2. $\lim_{N\to\infty} \dfrac{\log N}{(\log(s_N/\mu_N))^2} \log\left(\dfrac{\log N}{\log(s_N/\mu_N)}\right) = 0.$

3. $\lim_{N\to\infty} \dfrac{s_N \log N}{\log(s_N/\mu_N)} = 0.$

4. There exists $\varepsilon > 0$ such that $(\log N)/\log(s_N/\mu_N)$ is not within ε of an integer for any N.

Assumption 4, and the iterated logarithm in Assumption 2, are probably not necessary for the results.

Assumptions imply $s_N \to 0$ and $N^{-a} \ll \mu_N \ll s_N^b$ for all $a, b > 0$.

Assumptions hold when $s_N = e^{-(\log N)^a}$ and $\mu_N = e^{-(\log N)^b}$, where $0 < a < 1/2 < b < 1$ and $a + b < 1$.

Speed of Evolution

Heuristics due to Desai and Fisher (2007).

Let $X_j(t)$ be the number of individuals at time t with j mutations.

Let $\tau_{j+1} = \min\{t : X_j(t) \geq s_N/\mu_N\}$ (when type $j+1$ appears).

For $t \geq \tau_{j+1}$, stochastic fluctuations not important and

$$X_j(t) \approx \frac{s_N}{\mu_N} e^{\int_{\tau_j+1}^{t} s_N(j - m_N(u))\, du}.$$

Let $k_{j,N} = j - m_N(\tau_j)$. For $t \in [\tau_j, \tau_{j+1}]$,

$$\begin{aligned} E[X_j(t)] &\approx \int_{\tau_j}^{t} \mu_N E[X_{j-1}(u)] e^{s_N k_{j,N}(t-u)}\, du \\ &\approx \int_{\tau_j}^{t} \mu_N \cdot \frac{s_N}{\mu_N} e^{s_N(k_{j,N}-1)(u-\tau_j)} e^{s_N k_{j,N}(t-u)}\, du \\ &\approx e^{s_N k_{j,N}(t-\tau_j)}, \end{aligned}$$

so

$$\frac{s_N}{\mu_N} \approx e^{s_N k_{j,N}(\tau_{j+1}-\tau_j)}, \qquad \tau_{j+1} - \tau_j \approx \frac{1}{s_N k_{j,N}} \log\left(\frac{s_N}{\mu_N}\right).$$

Speed of evolution is approximately

$$\frac{1}{\tau_{j+1} - \tau_j} \approx \frac{s_N k_{j,N}}{\log(s_N/\mu_N)}.$$

For a population of size N in equilibrium,

$$k_{j,N} \approx k_N = \frac{2 \log N}{\log(s_N/\mu_N)},$$

so the speed of evolution is

$$v_N = \frac{2 s_N \log N}{(\log(s_N/\mu_N))^2}.$$

Define the natural time scale

$$a_N = \frac{1}{s_N} \log\left(\frac{s_N}{\mu_N}\right).$$

Theorem (Schweinsberg, 2014+): Let $\varepsilon > 0$. There exists $t(\varepsilon)$ such that for each fixed $t > t(\varepsilon)$,

$$\lim_{N \to \infty} P(|m_N(a_N t) - v_N a_N t| > \varepsilon v_N a_N t) = 0.$$

Distribution of Fitnesses

If $Z \sim N(\mu, \sigma^2)$, and let f be the density of Z. Then

$$\log\left(\frac{f(\mu+\ell)}{f(\mu)}\right) = -\frac{\ell^2}{2\sigma^2}.$$

Let $\gamma_j = \tau_j + \left(1 + \frac{1}{2k_N}\right)a_N$, which is time when type j peaks.
Let $j(t)$ be the value of j for which γ_j is closest to $a_N t$.

Theorem (Schweinsberg, 2014+): Let $\varepsilon > 0$. Let $\ell \in \mathbb{Z}$. There exists $t(\varepsilon)$ such that for each fixed $t > t(\varepsilon)$,

$$\lim_{N\to\infty} P\left(\left|\log\left(\frac{X_{j(t)+\ell}(\gamma_{j(t)})}{X_{j(t)}(\gamma_{j(t)})}\right) + \frac{\ell^2(\log(s_N/\mu_N))^2}{4\log N}\right| > \varepsilon\frac{(\log(s_N/\mu_N))^2}{\log N}\right) = 0.$$

Variance of "Gaussian" tends to zero, so one type dominates.

Genealogy of the Population

Heuristics due to Desai, Walczak, and Fisher (2013).

Recall that for $t \in [\tau_j, \tau_{j+1}]$, we have $E[X_j(t)] \approx e^{s_N k_N (t-\tau_j)}$.

Consider the possibility of an unusually early mutation:

- Mutations at time u happen at rate $\mu_N X_{j-1}(u)$.
- A mutation has probability approximately $s_N k_N$ of spreading, then number of descendants at time t is approximately

$$\frac{W}{s_N k_N} e^{s_N k_N (t-u)}, \qquad W \sim \text{exponential(1)}.$$

- A successful mutation at time

$$\tau_j + \frac{1}{s_N k_N} \log\left(\frac{1}{s_N k_N}\right) + \frac{B}{s_N k_N}$$

  has approximately $W e^{-B} e^{s_N k_N (t-\tau_j)}$ descendants at time t.

- This mutation will be the ancestor of a fraction at least x of the population if $W \geq x e^B/(1-x)$, and the probability of such a mutation is approximately $(1-x)/(k_N x)$.

Limit Theorem for the Genealogy

Theorem (Schweinsberg, 2014+): Fix $t > 0$ and $T > t + 2$. Sample n individuals at time $a_N T$. For $0 \leq u \leq t$, let $\Pi_N(u)$ be the partition of $\{1, \ldots, n\}$ such that i and j are in the same block if and only if the ith and jth sampled individuals have the same ancestor at time $a_N(T - u)$. Then

$$\lim_{N \to \infty} P(\Pi_N(1) = \{\{1\}, \ldots, \{n\}\}) = 1.$$

The finite-dimensional distributions of $(\Pi_N(1 + u), 0 \leq u \leq t)$ converge as $N \to \infty$ to those of Bolthausen-Sznitman coalescent.

Comments about proofs

1. For $t \in [\tau_j, \tau_{j+1}]$, approximate $X_j(t)$ by a supercritical branching process with immigration.

2. For $t \geq \tau_{j+1}$, control fluctuations in $X_j(t)$ using second moment arguments, similar to Durrett and Mayberry (2011).

3. **Challenge**: We want to approximate

$$X_j(t) \approx \frac{s_N}{\mu_N} e^{\int_{\tau_j+1}^{t} s_N(j - m_N(u))\, du}$$

but $m_N(u)$ is random and depends on $X_j(u)$.
Solution: Show that the approximation works when $m_N(u)$ stays in a tube, and that $m_N(u)$ stays in a tube as long as the approximation works.

4. **Challenge**: Heuristics rely on the population being in equilibrium, don't rigorously understand the stationary distribution.
Solution: Follow process as it moves towards equilibrium.

Future Work

The proof and the heuristic argument break down when μ_N is too large and Assumption 2 fails.

Fluctuations in $X_j(t)$ after time τ_{j+1} can no longer be ignored. We have to consider stochastic effects for more than just fittest type of individuals.

Open question: How many of the main results still hold?