Rigorous results for a population model with selection

by Jason Schweinsberg
University of California at San Diego

The Model (Moran model with selection)

Consider a population with N individuals.

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

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

where $M(t)$ 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, its replacement is chosen at random from the population, with probability proportional to fitness.

Although model is simple, much is unknown about how the population behaves.

Questions of Interest

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

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

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

One mutation at a time

If $s_N = s > 0$ and

$$\mu_N \ll \frac{1}{N \log N},$$

there will be only one beneficial mutation in population at a time that has not already spread to the entire population.

Mutations happen at rate $N\mu_N$. Then the number of individuals with the mutation behaves like an asymmetric random walk.

With probability approximately s, a selective sweep occurs, and the beneficial mutation spreads to the entire population.

Exponential waiting time with rate $N\mu_N s$ until first selective sweep, another exponential waiting time until next one, etc.

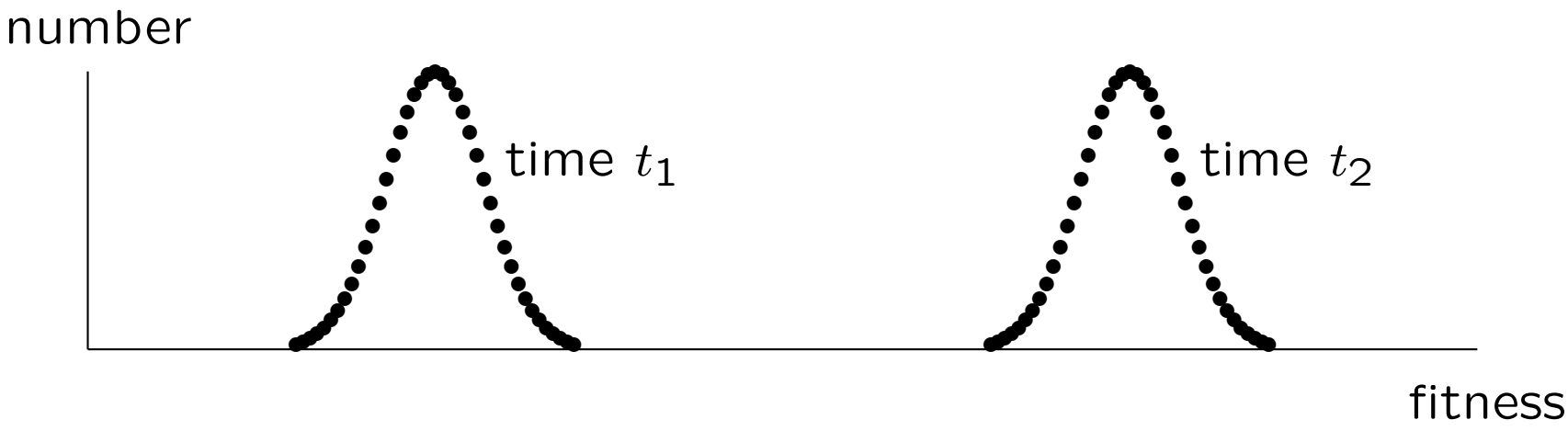
If mutations happen faster, so more than one beneficial mutation is in the population at a time, analysis is much more complicated.

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)
Rouzine, Brunet, and Wilke (2008)
Park, Simon, and Krug (2010)

They obtained precise estimates of 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."



Previous Non-Rigorous Work, continued

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

The two groups of authors worked in different parameter regimes.

Goal: For some range of μ_N and s_N, obtain rigorously:

- The speed of evolution.
- The Gaussian shape for the distribution of fitnesses.
- The genealogy of the population.

Previous Rigorous Work

Durrett and Mayberry (2011) consider the case with $s_N = s > 0$ and $\mu_N \sim N^{-\beta}$, where $0 < \beta < 1$. If T_j is the first time some individual has j mutations,

$$\frac{T_j - T_{j-1}}{\log N} \to_p t_j,$$

where $(t_j)_{j=1}^{\infty}$ is a sequence of constants depending on s and β. They also calculated the 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 $\delta > 0$,

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

for sufficiently large N.

Kelly (2013, 2015) studied model of Yu, Etheridge, and Cuthbertson and showed that $M(t)/t$ is of the order

$$\frac{\log N}{(\log \log N)^2}.$$

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

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

Suppose $1/2 < c < 1$ and $0 < d < 1 - c$. Assumptions hold if:

$$\mu_N = e^{-(\log N)^c}, \qquad e^{-(\log N)^d} \le s_N \le \frac{1}{\sqrt{\log N}}.$$

Hereafter, we write μ and s for μ_N and s_N.

Beginning of the process

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

When t is small, $X_0(t) \approx N$ and $M(t) \approx 0$.

Approximate by multitype branching process: type j individual gives birth at rate $1 + js$, dies at rate 1, mutates to type $j + 1$ at rate μ. We get

$$E[X_1(t)] \approx \int_0^t \mu E[X_0(u)] \cdot e^{s(t-u)}\, du \approx \frac{N\mu(e^{st} - 1)}{s}.$$

By induction,

$$E[X_j(t)] \approx \frac{N\mu^j}{s^j j!}(e^{st} - 1)^j.$$

Validity of the approximation

The approximation

$$E[X_j(t)] \approx \frac{N\mu^j}{s^j j!}(e^{st}-1)^j$$

can only hold as long as $M(t) \approx 0$. This requires $X_1(t) \ll N$, for which we need

$$t \leq \frac{1}{s}\log\left(\frac{s}{\mu}\right) = a_N.$$

Second-moment arguments show give $X_j(t) \approx E[X_j(t)]$ when $t \leq a_N$ and $j \leq k_N$, where

$$k_N = \frac{\log N}{\log(s/\mu)}.$$

For $j \leq k_N$, we have $P(X_j(\varepsilon a_N) > 0) \to 1$.

For $j > k_N$, it is not true that $X_j(t) \approx E[X_j(t)]$.

Evolution of type j individuals

Let $\tau_j = \min\{t : X_{j-1}(t) \geq s/\mu\}$.

Stage 0: Before time τ_j, typically no type j individuals appear.

Stage 1: Between times τ_j and τ_{j+1}, type $j-1$ individuals acquire mutations, causing the type j population to emerge.

Stage 2: After τ_{j+1}, type j population is well-established. Further mutations from type $j-1$ to type j have a negligible effect. Type j population grows at a predictable rate. For $t \geq \tau_{j+1}$,

$$X_j(t) \approx \frac{s}{\mu} \exp\left(\int_{\tau_{j+1}}^{t} s(j - M(u))\, du \right).$$

It is important for the analysis that the type j population be well-established before type $j+1$ individuals appear. This is the reason for Assumption 2, which upper bounds the mutation rate.

Stage 1 in more detail

For $t \in [\tau_j, \tau_{j+1}]$,

$$X_{j-1}(t) \approx \frac{s}{\mu} \exp\left(\int_{\tau_j}^{t} s(j-1-M(u))\, du \right).$$

Let $Q(t) = \max\{j : X_j(t) > 0\} - M(t)$. Then

$$X_{j-1}(t) \approx \frac{s}{\mu} e^{sQ(\tau_j)(t-\tau_j)},$$

so

$$\begin{aligned} X_j(t) &\approx \int_{\tau_j}^{t} \mu \cdot \frac{s}{\mu} e^{sQ(\tau_j)(u-\tau_j)} \cdot e^{s(Q(\tau_j)+1)(t-u)}\, du \\ &\approx e^{s(Q(\tau_j)+1)(t-\tau_j)} \end{aligned}$$

if $t - \tau_j \gg 1/s$.

Speed of evolution

Recall $X_j(t) \approx e^{s(Q(\tau_j)+1)(t-\tau_j)}$. Set equal to s/μ to get

$$\tau_{j+1} - \tau_j \approx \frac{1}{sQ(\tau_j)} \log\left(\frac{s}{\mu}\right).$$

We will see that $Q(\tau_j) \approx 2\log N/[\log(s/\mu)]$ for large j, so rate at which mutations take hold:

$$\frac{1}{\tau_{j+1} - \tau_j} \approx \frac{s}{\log(s/\mu)} Q(\tau_j) \approx \frac{2s \log N}{[\log(s/\mu)]^2}$$

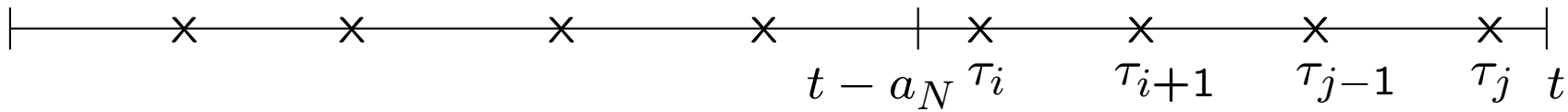
Desai and Fisher (2007): $\dfrac{2s\log(Ns)}{[\log(s/\mu)]^2}$

Rouzine, Brunet, and Wilke (2008): $\dfrac{2s\log(N\sqrt{s\mu})}{[\log((s/\mu)\log(N\sqrt{s\mu}))]^2}$

They used similar heuristics, but assumed the population was in equilibrium. Expressions are equivalent under our assumptions.

Estimation of $Q(t)$

$X_j(t)$ surpasses $X_{j-1}(t)$ around time $\tau_j + a_N$. $[a_N = (1/s)\log(s/\mu)]$
$M(t)$ is approximately the value of i such that $\tau_i \approx t - a_N$.
$Q(t)$ is approximately the number of τ_j between $t - a_N$ and t.



Rate at which the τ_j appear:

$$\frac{1}{\tau_{j+1} - \tau_j} = \frac{s}{\log(s/\mu)} Q(\tau_j).$$

After scaling,

$$\frac{\log(s/\mu)}{\log N} Q(a_N t) \approx q(t) = \begin{cases} e^t & \text{if } 0 \le t < 1 \\ \int_{t-1}^{t} q(u)\, du & \text{if } t \ge 1. \end{cases}$$

Let $U(t)$ be expected number of renewals by time t when times between renewals are independent Uniform(0,1).

Then $U'(t) = q(t)$ for all $t \ge 0$, so $\lim_{t\to\infty} q(t) = 2$.

Precise Result

Theorem (Schweinsberg, 2017): Suppose assumptions 1, 2, and 3 hold. Let K be a compact subset of $(0,1) \cup (1,\infty)$. Then

$$\sup_{t \in K} \left| \frac{\log(s/\mu)}{\log N} Q(a_N t) - q(t) \right| \to_p 0.$$

Also, let

$$m(t) = \begin{cases} 0 & \text{if } 0 \le t < 1 \\ 1 + \int_0^{t-1} q(u)\, du & \text{if } t \ge 1. \end{cases}$$

Then

$$\sup_{t \in K} \left| \frac{\log(s/\mu)}{\log N} M(a_N t) - m(t) \right| \to_p 0,$$

where $\to_p$ denotes convergence in probability as $N \to \infty$.

Remark: The convergence is not guaranteed at $t = 1$ because q and m are not continuous at 1. Near time a_N, the mean number of mutations rapidly increases from 0 to $k_N = (\log N)/\log(s/\mu)$.

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, 2017): Let $\varepsilon > 0$. Let $\ell \in \mathbb{Z}$. There exists $t(\varepsilon)$ such that for each fixed $t > t(\varepsilon)$,

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

Resembles Gaussian with variance $\sigma_N^2 = 2\log N/[\log(s/\mu)]^2$.
However, $\sigma_N^2 \to 0$, so one type dominates.

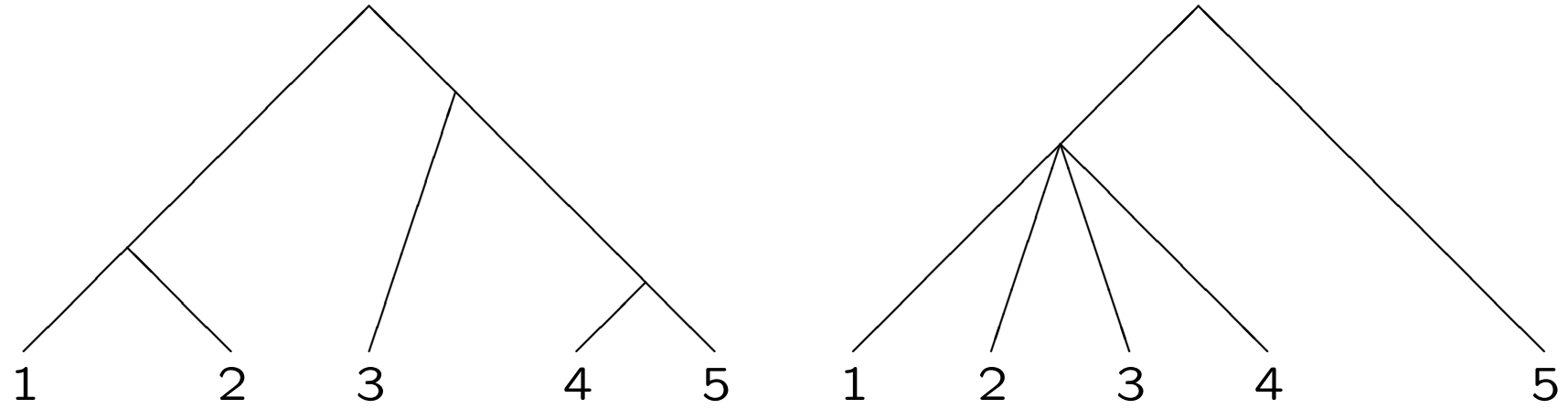
Coalescent Processes

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.

Coalescents with multiple mergers (Pitman, 1999; Sagitov, 1999): Many lineages can merge at once.



Bolthausen-Sznitman coalescent

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

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

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.

Rate of mergers impacting more than a fraction $x/(1+x)$ of lineages is

$$\int_{x/(1+x)}^1 p^{-2}\, dp = x^{-1}.$$

Bolthausen-Sznitman coalescent describes the genealogy when, if the population has size K, new "families" of size at least Kx appear at a rate proportional to x^{-1}.

Genealogy of the Population

Recall the approximation $X_j(t) \approx e^{sq_j(t-\tau_j)}$, $q_j = Q(\tau_j) + 1$.

Usually, there will be many small type j families.

Consider the possibility of an unusually early mutation:

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

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

- A successful mutation at time

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

  has approximately $We^{-B}e^{sq_j(t-\tau_j)}$ descendants at time t.

- The probability that there will be such a mutation with $We^{-B} \geq x$ is approximately $q_j^{-1}x^{-1}$.

Tracing back ancestral lines

Sample n individuals at time $a_N T$.

The individuals will most likely have the same type (type j) and come from different type j ancestors at time $\tau_{j+1} \approx a_N(T-1)$. No coalescence during this period.

Type j individuals at time τ_{j+1} get traced back to type $j-1$ ancestors at time τ_j, to type $j-2$ ancestors at time τ_{j-1}, etc.

At each step, small chance of multiple merger due to unusually early mutation, merger rates match Bolthausen-Sznitman coalescent.

Similar heuristics appear in Desai, Walczak, and Fisher (2013).

Main coalescent result

Theorem (Schweinsberg, 2017): Fix $t > 0$ and $T > t+2$. Sample n individuals at time $a_N T$. For $0 \leq u \leq t+1$, 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.