Population Genetics and Evolution – III

Statistics of Genealogies: The Coalescent

Luca Peliti

Dipartimento di Fisica and Sezione INFN
Università di Napoli “Federico II”

Helsinki / June 2013

Outline

Introduction

The Coalescent

The Coalescent with selection

Introduction

Genealogies

- How far in the past must we go to reach the last common ancestor of n individuals? of the whole population?
- How many different genotypes can we expect to find by sampling n individuals?
- How do the times to the last common ancestor depend on the particular chosen sample? on the population size?
- How do they fluctuate as the population evolves in time?
- How are they affected by selection?

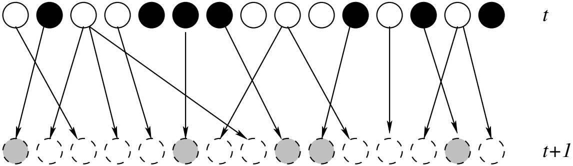
These questions can be addressed by using the concept of the *Coalescent*

The Coalescent

JFC Kingman

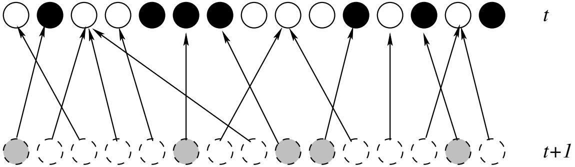
The Wright-Fisher model

Two ways of looking at the Wright-Fisher model:

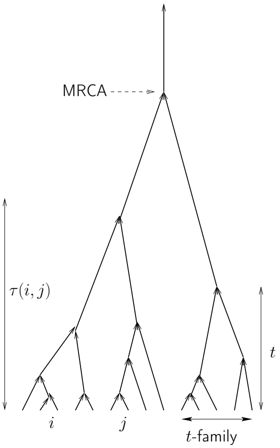


The Wright-Fisher model

Two ways of looking at the Wright-Fisher model:



Iterating the process



Iterating the process

Neutral Wright-Fisher process:

- Set $t = 0$ for the present, and count generations *backward* from the present
- Individual labels: $\{1, \dots, N\}$
- At each generation, define the application $p: \; i \mapsto p_t(i)$ from i to its parent
- $p_t(i)$ is extracted at random, independently for each i and each t
- Ancestor: $a_t(i) = \underbrace{p_t(p_{t-1}(\cdots p_2(p_1}_{t \text{ times}}(i))))$
- Lineage: $\mathrm{L}(i) = (a_0(i) = i, a_1(i), a_2(i), \dots)$
- Lineage coalescence: $a_t(i) = a_t(j)$, $i \neq j$
- Coalescence time: $\tau(i,j)$: $a_\tau(i) = a_\tau(j)$, $a_{\tau-1}(i) \neq a_{\tau-1}(j)$

Iterating the process

Disclaimer:

> *In this [lecture] gene genealogies will sometimes be referred to simply as genealogies. It should be understood that this refers to the genetic ancestry of a sample at some locus in the genome and not to the usual definition of a genealogy, being the family relationship of a set of individuals.*

J. WAKELEY, 2009

Iterating the process

Questions:

- How many generations to the MRCA?
- What is the distribution of $\tau(i, j)$?
- What are the consequences for quantities we can measure?

N.B.: When treating *diploids*, set $N = 2 \cdot$ population size
Discussion of the *effective* population size: later!

Coalescent statistics

Hypotheses:

1. Equal fitness for all types (neutral process)
2. No subdivisions in the population (geographical or otherwise)
3. Constant population size

Assumptions 1. and 2. lead to *exchangeability*: the number of offspring of any individual is statistically the same random variable as for any other individual

Coalescent statistics

- Probability that n individuals have all different parents:

$$\begin{aligned} w_n &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right) \\ &\simeq 1-\frac{n(n-1)}{2N} \qquad\qquad n \ll N \end{aligned}$$

- $\Pi_n(t)$: probability of n independent lineages at time t

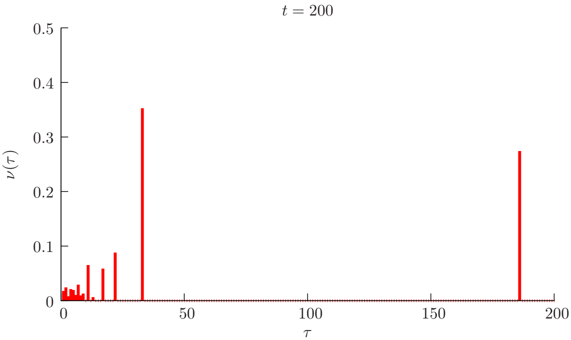
$$\Pi_n(t+1) = w_n \Pi_n(t) \simeq \left(1-\frac{n(n-1)}{2N}\right)\Pi_n(t)$$

- $\Pi_n(t) = \left(1-\frac{n(n-1)}{2N}\right)^t \simeq \mathrm{e}^{-n(n-1)t/(2N)}$
- In particular $\Pi_2(t) \simeq \mathrm{e}^{-t/N}$

Coalescent statistics

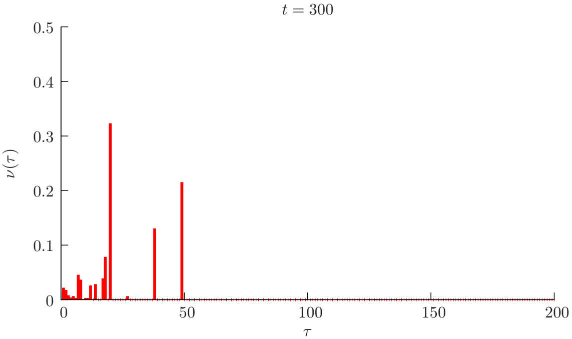
- Averages over the *process* are expressed by $\left[\ldots\right]_{\mathrm{av}}$
- Averages over the *population* are expressed by $\langle\ldots\rangle$
- Thus $[\tau(i,j)]_{\mathrm{av}} = N$
- Mutation rate u per genome and generation, infinite *site* model
- Expected # of mutations wrt the common ancestor: Nu
- Expected # of mutations between i and j: $2Nu = \theta$

Distribution of coalescent times



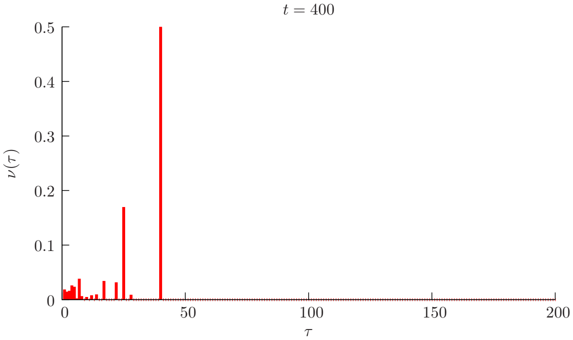
$N = 50$

Distribution of coalescent times



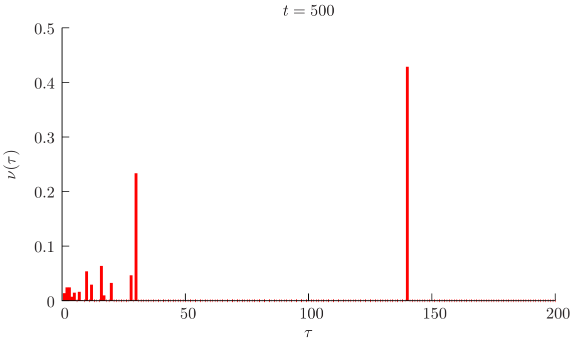
$N = 50$

Distribution of coalescent times



$N = 50$

Distribution of coalescent times



$N = 50$

Universality of the coalescent

- Reproduction model: Distribution of offspring size m: π_m

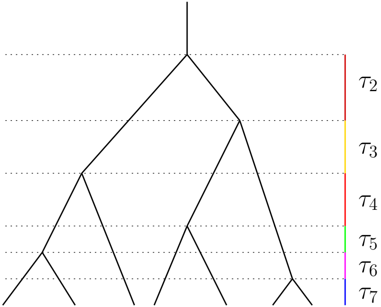
$$\begin{aligned} \text{WF model: } \pi_m &= \mathrm{e}^{-1}/m! \qquad \text{(Poisson)} \\ \text{Moran model: } \pi_0 &= \pi_2 = \frac{1}{N}\left(1-\frac{1}{N}\right),\ \pi_1 = 1-\frac{2}{N}\left(1-\frac{1}{N}\right) \end{aligned}$$

- $[m]_{\mathrm{av}} = \sum_m m\,\pi_m = 1$
- Probability of coalescence for n lineages:

$$1-w_n = \binom{n}{2}\frac{1}{N}\sum_m m(m-1)\,\pi_m = \frac{n(n-1)}{2N}\left(\left[m^2\right]_{\mathrm{av}}-1\right)$$

- Define $[m(m-1)]_{\mathrm{av}} = \left[m^2\right]_{\mathrm{av}} - 1 = \kappa$
- Thus $w_n = 1 - \frac{n(n-1)}{2}\frac{\kappa}{N}$
- If $\left[m^2\right]_{\mathrm{av}} < \infty$, all results hold, up to a time rescaling
- Choose time units so that $w_n = 1 - \frac{n(n-1)}{2}$

Probability of a genealogy



$$P(\tau_2, \ldots, \tau_7) = \exp\left\{-\frac{1}{2}\left[7 \cdot 6 \cdot \tau_7 + 6 \cdot 5 \cdot \tau_6 + \cdots + 2 \cdot 1 \cdot \tau_2\right]\right\}$$

Each τ_k is independent, with distribution $\mathcal{P}_k(\tau) = \binom{k}{2}\, \mathrm{e}^{-\binom{k}{2}\tau}$

Distribution of the total length

- Define $T_{\mathrm{total}} = \sum_{k=2}^{n} T_k$, $T_k = k \cdot \tau_k$
- Then each T_k is an exponentially distributed random variable, of average $[T_k]_{\mathrm{av}} = 2/(k-1)$

Distribution of the total length

$$\mathcal{P}_{\mathrm{total}}(T) = \mathrm{Prob}(T_{\mathrm{total}} = T) = \int_0^\infty \prod_{k=2}^{n} \left(\mathrm{d}\,T_k \, \frac{(k-1)\,\mathrm{e}^{-(k-1)T_k/2}}{2} \right)$$

$$\times\, \delta\left(\sum_{k=2}^{N} T_k - T \right)$$

$$= \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \frac{\mathrm{d}\lambda}{2\pi\mathrm{i}} \int_0^\infty \prod_{k=2}^{n} \left(\mathrm{d}\,T_k \, \frac{k-1}{2}\, \mathrm{e}^{-(k-1)T_k/2} \right)$$

$$\times \exp\left[-\lambda \left(\sum_{k=2}^{N} T_k - T \right) \right]$$

$$= \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \frac{\mathrm{d}\lambda}{2\pi\mathrm{i}}\, \mathrm{e}^{\lambda T} \prod_{k=2}^{n} \left(\frac{k-1}{2\lambda + (k-1)} \right)$$

Distribution of the total length

Summing over the residues

$$\begin{aligned}\mathcal{P}_{\text{total}}(T) &= \sum_{k=2}^{n} \frac{k-1}{2}\,\mathrm{e}^{-(k-1)T/2} \prod_{j\,(\neq k)} \frac{j-1}{j-k} \\ &= \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2}\,\mathrm{e}^{-(k-1)T/2} \\ &= \frac{n-1}{2}\mathrm{e}^{-T/2}\left(1-\mathrm{e}^{-T/2}\right)^{n-2}\end{aligned}$$

TAVARÉ, 1984; WIUF AND HEIN, 1999

Distribution of the age of the MRCA

- Define T_{MRCA} as the age of the MRCA of n samples
- Then $T_{\mathrm{MRCA}} = \sum_{k=2}^{n} \tau_k$
- Each τ_k is exponentially distributed, with average $[\tau_k]_{\mathrm{av}} = \left[\binom{k}{2}\right]^{-1}$

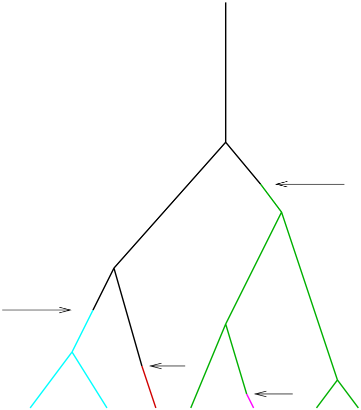
Distribution of the age of the MRCA

Using the same method one obtains

$$\begin{aligned}\mathcal{P}_{\mathrm{MRCA}}(T) &= \mathrm{Prob}(T_{\mathrm{MRCA}} = T)\\ &= \sum_{k=2}^{n} \binom{k}{2} \mathrm{e}^{-\binom{k}{2}T} \prod_{j\,(\neq k)} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{k}{2}}\\ &= \sum_{k=2}^{n} \binom{k}{2} (-1)^k (2k-1) \frac{n(n-1)\cdots(n-k+1)}{n(n+1)\cdots(n+k-1)}\, \mathrm{e}^{-\binom{k}{2}T}\end{aligned}$$

TAVARÉ, 1984; TAKAHATA AND NEI, 1985

Coalescence and mutations



The probability of a mutation occurring is uniform per unit length of the genealogy

Coalescence and mutations

- Assume mutation rate u per genome and generation, infinite *allele* model
- Two individuals carry the same allele if they encounter no mutation before their last common ancestor
- The probability of *not* having a mutation in a generation in a lineage is $1 - u$
- The probability that *neither* lineage exhibits a mutation is $(1-u)^{2\tau(i,j)} \simeq \exp\left(-2u\tau(i,j)\right)$
- Thus the probability that two individuals have the same allele is

$$
\begin{aligned}
p_{\text{same}} &= \frac{1}{N}\int_0^\infty \mathrm{d}\tau\ \mathrm{e}^{-2u\tau-\tau/N} \\
&= \frac{1}{1+2uN} = \frac{1}{1+\theta}
\end{aligned}
$$

Ewens' sampling formula

- Infinite-allele model
- Take n samples from a large population with $\theta = 2Nu$
- Samples belong to the same group if they exhibit the same allele
- What is the probability that there are b_1 groups with 1 element, b_2 groups with 2 elements,... b_k with k elements,... ?

Ewens' sampling formula

$$n = \sum_{k=1}^{n} k\, b_k \qquad \text{\# of samples}$$

$$\boxed{P(b_1, \ldots, b_n) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \frac{1}{1^{b_1} \cdot 2^{b_2} \cdots n^{b_n}} \frac{\theta^{\sum_k b_k}}{b_1! b_2! \cdots b_n!}}$$

The Chinese Restaurant Process

The Chinese Restaurant Process

At each step, when there are n customers:

- The customer sits at a new empty table with probability $\theta/(\theta + n)$, or
- The customer picks up one of the customers at random and sits at the same table

The Chinese Restaurant Process

- At each step, we get a factor $1/(\theta + n)$ $(n = 0, 1, \ldots)$
- Each new table gets a factor θ
- In going from k to $k+1$, each table gets a factor k
- Thus the probability that the (labeled) customers sit at ℓ tables, $i = 1, \ldots, \ell$ of size k_i, $\sum_{i=1}^{\ell} k_i = n$ is given by

$$P^{\text{lab}}(k_1, \ldots, k_\ell) = \frac{\theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{\ell} (k_i - 1)!$$

- There are $n!/(k_1! \cdots k_\ell!)$ distributions of the customers compatible with $(k_1, \ldots, k_\ell)$, thus

$$\begin{aligned} P(k_1, \ldots, k_\ell) &= \frac{n!}{k_1! \cdots k_\ell!} \frac{\theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{\ell} (k_i - 1)! \\ &= \frac{n!\, \theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{\ell} \frac{1}{k_i} \end{aligned}$$

The Chinese Restaurant Process

- Labelling the tables has introduced an overcounting: only the sizes of the tables matter! Thus defining

$$b_j = \sum_{i=1}^{\ell} \delta_{k_i,j}$$

we obtain

$$P(b_1,\ldots,b_n) = \frac{n!\,\theta^{\ell}}{\theta(\theta+1)\cdots(\theta+n-1)}\frac{1}{1^{b_1}\cdots n^{b_n}}\underbrace{\frac{1}{b_1!\cdots b_n!}}_{\text{Table permutations}}$$

Observables

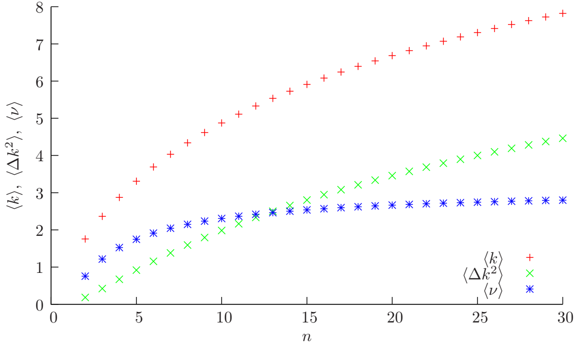
- Distribution of the number k of segregating alleles:

$$
\begin{aligned}
p_k(n+1) &= \frac{n}{\theta+n}p_k(n)+\frac{\theta}{\theta+n}p_{k-1}(n) \\
[k(n+1)]_{\mathrm{av}} &= [k(n)]_{\mathrm{av}}+\frac{\theta}{\theta+n}=\theta\sum_{j=1}^{n-1}\frac{1}{\theta+j} \\
\left[\Delta k^2(n+1)\right]_{\mathrm{av}} &= \left[k^2(n)\right]_{\mathrm{av}}-[k(n)]_{\mathrm{av}}^2=\left[\Delta k^2(n)\right]_{\mathrm{av}}+\frac{n\theta}{(\theta+n)^2}
\end{aligned}
$$

- Distribution of the number ν of singletons:

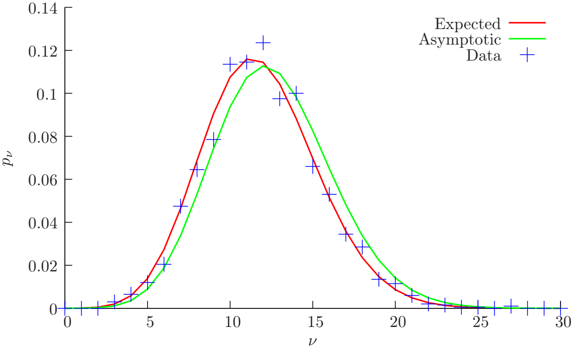
$$
\begin{aligned}
p_\nu(n+1) &= \frac{\theta}{\theta+n}p_{\nu-1}(n)+\frac{\nu}{\theta+n}p_{\nu+1}(n)+\frac{n-\nu}{\theta+n}p_\nu(n) \\
[\nu(n)]_{\mathrm{av}} &= \frac{n\theta}{\theta+n-1}
\end{aligned}
$$

Observables



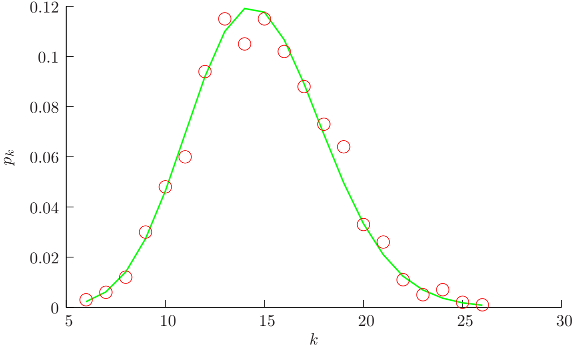
Average $[k]_{\mathrm{av}}$, variance $\left[\Delta k^2\right]_{\mathrm{av}}$ of segregating alleles and average $[\nu]_{\mathrm{av}}$ of singletons vs. n for $\theta = 3.1$

Observables



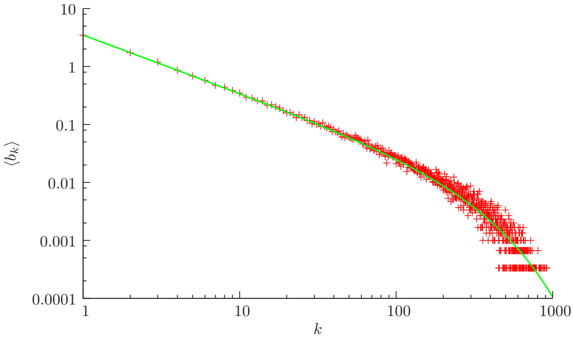
Distribution p_ν of the number of singletons for $n = 200$ and $\theta = 12.6$, together with the asymptotic distribution for $n \to \infty$ and simulation data over 1000 samples

Observables



Distribution p_k of the number of segregating alleles for $n = 300$ and $\theta = 3.1$, together with simulation data averaged over 1000 samples

Frequency spectrum



Average number $[b_k]_{\mathrm{av}}$ of groups of size k with $n = 1000$ and $\theta = 3.5$. The average is taken over 3000 realizations of the process.

The line corresponds to $[b_k]_{\mathrm{av}} = [b_1]_{\mathrm{av}}\, \mathrm{e}^{-\theta k/n}/k$, with $[b_1]_{\mathrm{av}} = n\theta/(\theta + n - 1)$

Effective population size N_e

The *effective population size* N_e can be different from the *census population* N:

- In sexual populations, because only some males actually reproduce(*leks*)
- Generally due to fluctuating population size:

$$\frac{1}{N_\mathrm{e}} \simeq \left[\frac{1}{N}\right]_\mathrm{av} > \frac{1}{[N]_\mathrm{av}}$$

- If fitness is nonuniform N_e is reduced wrt N:

$$N_\mathrm{e} = \frac{N}{1 + \mathrm{var}(\#\mathrm{offspring})}$$

Effective population size N_{e}

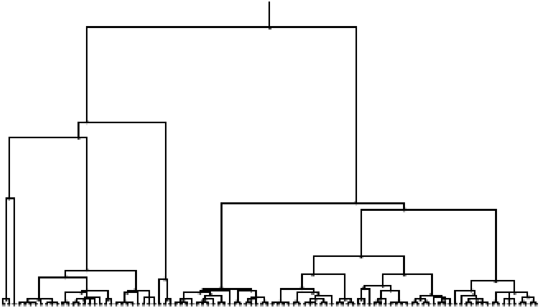
In practice, N_{e} is chosen to fit the data:

- For several human genes, $T_{\mathrm{MRCA}} \simeq 400\,000\ \mathrm{yrs}$
- One generation $\simeq 20\ \mathrm{yrs}$
- Assuming neutrality, $N_{\mathrm{e}} \simeq 10\,000$ (diploidy!)
- "Out-of-Africa" bottleneck?

The Coalescent with selection

The Coalescent in the presence of selection

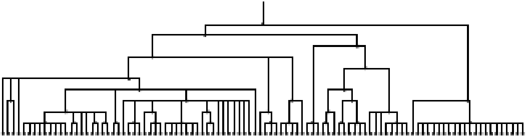
BRUNET, DERRIDA *et al.*, 2006–2012



Neutral genealogy: $N = 100$, $T_{\mathrm{MRCA}} = 125$

The Coalescent in the presence of selection

BRUNET, DERRIDA *et al.*, 2006–2012



Genealogy with selection: $N = 100$, $T_{\mathrm{MRCA}} = 10$

Coalescent times

A general coalescence model (Λ-coalescent):

- One starts with N points: in each interval of duration $\mathrm{d}t$ there is a probability $\pi_k \,\mathrm{d}t$ for every subset of k points to coalesce into one
- Then for some measure Λ one has

$$\pi_k = \int_0^1 x^k \,\Lambda(\mathrm{d}x)$$

- Rate $\lambda_{b,k}$ at which k ($2 \leq k \leq p$) points out of p coalesce into one is given by

$$\lambda_{p,k} = \int_0^1 x^{k-2}(1-x)^{p-k}\lambda(\mathrm{d}x) = \sum_{n=0}^{p-k} \frac{(p-k)!}{n!(p-k-n)!}(-1)^n \pi_{n+k}$$

- $r_p(\ell)\,\mathrm{d}t$: probability of having ℓ lineages at time $t + \mathrm{d}t$ if there are p lineages at time t:

$$r_p(\ell) = \frac{p!}{(\ell-1)!(p-\ell+1)!}\lambda_{p,p-\ell+1}$$

Coalescent times

- T_p: coalescence time for p lineages
- Assume steady state:

$$[T_p]_{\mathrm{av}} = \mathrm{d}t + [T_p]_{\mathrm{av}} \left(1 - \mathrm{d}t \sum_{k<p} r_p(k) \right) + \mathrm{d}t \sum_{k<p} r_p(k) \, [T_k]_{\mathrm{av}}$$

Thus

$$\begin{aligned} [T_2]_{\mathrm{av}} &= \frac{1}{\pi_2} \\ \frac{[T_3]_{\mathrm{av}}}{[T_2]_{\mathrm{av}}} &= \frac{4\pi_2 - 3\pi_3}{3\pi_2 - 2\pi_3} \\ \frac{[T_4]_{\mathrm{av}}}{[T_2]_{\mathrm{av}}} &= \frac{27\pi_2^2 - 56\pi_2\pi_3 + 28\pi_3^2 + 12\pi_2\pi_4 + 10\pi_3\pi_4}{(3\pi_2 - 2\pi_3)(6\pi_2 - 8\pi_3 + 3\pi_4)} \\ &\vdots \end{aligned}$$

Coalescent times

In particular:

- The Kingman coalescent:

$$\pi_2 \neq 0 \qquad \pi_k = 0, \quad \forall k > 2$$

  yields

$$[T_2]_{\mathrm{av}} = \frac{1}{\pi_2}, \quad \frac{[T_3]_{\mathrm{av}}}{[T_2]_{\mathrm{av}}} = \frac{4}{3}, \quad \frac{[T_4]_{\mathrm{av}}}{[T_2]_{\mathrm{av}}} = \frac{3}{2} \quad \ldots$$

- The *Bolthausen-Sznitman coalescent*:

$$\pi_k = \frac{\pi_2}{k-1}$$

  yields

$$[T_2]_{\mathrm{av}} = \frac{1}{\pi_2}, \quad \frac{[T_3]_{\mathrm{av}}}{[T_2]_{\mathrm{av}}} = \frac{5}{4}, \quad \frac{[T_4]_{\mathrm{av}}}{[T_2]_{\mathrm{av}}} = \frac{25}{18} \quad \ldots$$

A solvable model

Brunet, Derrida et al., 2006–2012

- N individuals, discrete generations
- Individual i at generation t has “fitness” $x_i(t)$
- *Reproduction*: Probability that one offspring of individual i has “fitness” between x and $x + \mathrm{d}x$:

  $$P(x)\,\mathrm{d}x = \mathrm{e}^{-(x - x_i(t))}\,\mathrm{d}x$$

  Infinite # of offspring: but only finite # on the right of any given point
- *Selection*: At generation $t + 1$ one keeps only the N rightmost individuals

A solvable model

BRUNET, DERRIDA ET AL., 2006–2012

- Now

$$\sum_{i=1}^{N} \mathrm{e}^{-(x-x_i(t))} = \mathrm{e}^{-(x_i-X_t)} \qquad \text{with} \qquad \mathrm{e}^{X_t} = \sum_{i=1}^{N} \mathrm{e}^{x_i(t)}$$

- Thus generation $(t+1)$ is given by the N rightmost points of a Poisson process with density $\mathrm{e}^{-(x-X_t)}$
- Thus we have

$$x_i(t+1) = X_t + Y_{t+1} + y_i(t+1)$$

with

$$\begin{aligned} P(Y)\,\mathrm{d}Y &= \frac{1}{N!}\exp\left[-(N+1)Y - \mathrm{e}^{-Y}\right]\,\mathrm{d}Y \\ P(y)\,\mathrm{d}y &= \theta_{\mathrm{H}}(y)\,\mathrm{e}^{-y}\,\mathrm{d}y \end{aligned}$$

A solvable model

BRUNET, DERRIDA ET AL., 2006–2012

Results:

- Probability that the parent of *i* has “fitness” *x*:

$$p_i(x) = \frac{\mathrm{e}^{-(x-x_i(t))}}{\sum_j \mathrm{e}^{-(x-x_k(t))}} = \frac{\mathrm{e}^{y_i(t)}}{\sum_j \mathrm{e}^{y_{j(t)}}}$$

- Rate of *k*-coalescences:

$$\pi_k = \left[\sum_i p_i^k\right]_{\mathrm{av}} \simeq \frac{1}{(k-1)\log N}$$

Bolthausen-Sznitman!

- Speed of adaptation:

$$v = \langle X_t - X_{t-1}\rangle = \langle Y_t\rangle + \left\langle \log \sum_{i=1}^{N} \mathrm{e}^{y_i(t)} \right\rangle \sim \log\log N$$

A solvable model

BRUNET, DERRIDA ET AL., 2006–2012

Conditioning on the speed:

- Introduce a weighting parameter β:

$$[T_k]_\beta = \lim_{t\to\infty} \frac{1}{t} \sum_{t'=1}^{t} \frac{\left[\mathrm{e}^{-\beta X_t} \langle T_k(t') \rangle\right]_{\mathrm{av}}}{\left[\mathrm{e}^{-\beta X_t}\right]_{\mathrm{av}}}$$

- Coalescence rates:

$$\begin{aligned}\pi_k &= \frac{\left[\sum_i \mathrm{e}^{k y_i(t)} \left(\sum_j \mathrm{e}^{y_j(t)}\right)^{-\beta-k}\right]_{\mathrm{av}}}{\left[\left(\sum_j \mathrm{e}^{y_j(t)}\right)^{-\beta}\right]_{\mathrm{av}}} \\ &\simeq \frac{1}{\log N} \frac{(k-2)!\Gamma(\beta+1)}{\Gamma(\beta+k)}\end{aligned}$$

Interpolates between Bolthausen-Sznitman ($\beta = 0$) and Kingman ($\beta \to \infty$)

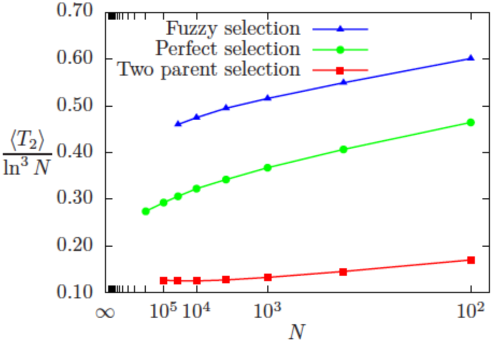
More generic models

BRUNET ET AL., 2006–2012

- Each individual has two potential offspring
- The fitness of each offspring is shifted by z wrt to the parent's one, with pdf $\rho(z)$ (flat in the simulations)
- Selection modes:
  - *Perfect selection*: The best N are retained
  - *Fuzzy selection*: Random choice among the $3N/2$ best
  - *Two-parent selection*: Each individual chooses two parents, but only the better one is kept

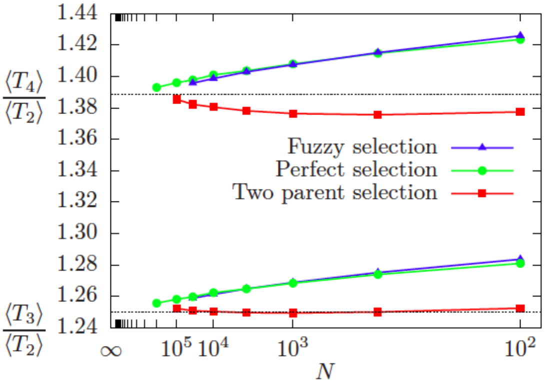
More generic models

BRUNET ET AL., 2006–2012



More generic models

BRUNET ET AL., 2006–2012



More generic models

BRUNET ET AL., 2006–2012

Coalescence time scale: $[T_2]_{\mathrm{av}} \sim \log^3 N$
Phenomenological theory

- The population looks like an advancing Kolmogorov-Fisher wave in “fitness” space
- Most of the time its motion is deterministic
- At intervals $\sim \log^3 N$ exceptionally “adapted” individuals arise
- These individual “sweep” a finite fraction of the population in a short time (multiple coalescence!)
- The distribution of the “sweep” sizes corresponds to the Bolthausen-Sznitman coalescent