

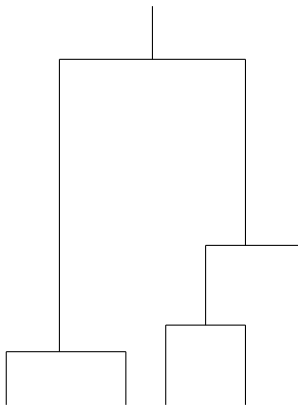
EWENS SAMPLING FORMULA AND RECOMBINATION

Majid Salamat

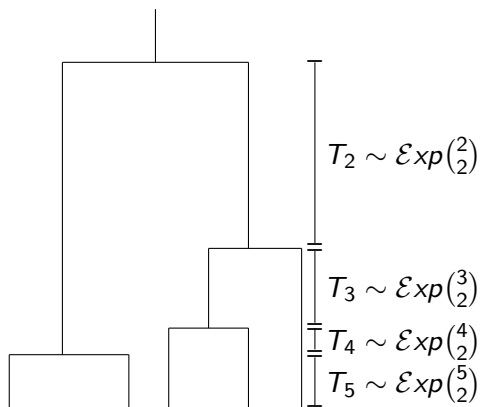
- 1 PRELIMINARIES
- 2 EWENS SAMPLING FORMULA WITHOUT RECOMBINATION
- 3 ESF WITH RECOMBINATION FOR A SAMPLE OF SIZE $n = 2, 3$
- 4 THE JOINT DISTRIBUTION OF T_L AND T_R
- 5 A POPULATION OF SIZE $n = 3$
- 6 GENERAL CASE

PRELIMINARIES

An example

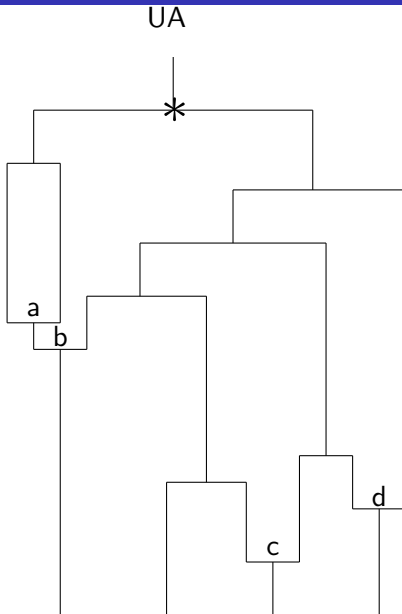


Kingman's coalescent



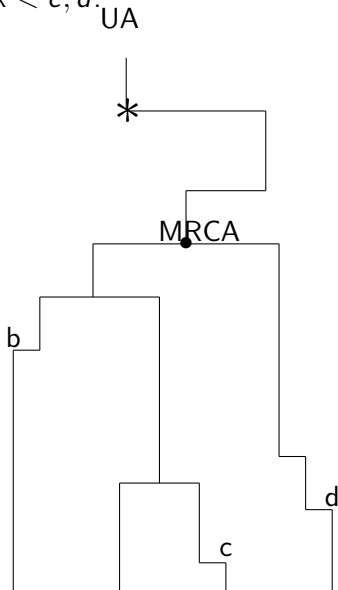
HOW TO MODEL THE RECOMBINATION IN THE PICTURE?

An example for ARG



An example for ARG

Assume that $x > b$ and $x < c, d$.



For each gene we have a different coalescent tree, because of the presence of the recombination. So the ARG contains all these trees.

EWENS SAMPLING FORMULA WITHOUT RECOMBINATION

ESF without recombination

- Infinitely-many-alleles mutation scheme, a configuration is considered as $c = (c_1, \dots, c_n)$

$c_i =$ number of alleles represented i times

and $|c| = c_1 + 2c_2 + \dots + nc_n = n$.

$q(c) := \mathbb{P}(\text{sample of size } |c| \text{ taken at stationarity has configuration } c)$,
with $q(e_1) = 1$.

Theorem

In a stationary sample of size n , the probability of sample configuration c is

$$q(c) = \mathbb{P}(C_1(n) = c_1, \dots, C_n(n) = c_n) = 1_{\{|c|=n\}} \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \binom{\theta}{j}^{c_j} \frac{1}{c_j!}$$

*where $x_{(j)} = x(x+1)\dots(x+j-1)$, $j = 1, 2, \dots$ and
 $|c| = c_1 + 2c_2 + \dots + nc_n$.*

ESF without recombination

- Infinitely-many-alleles mutation scheme, a configuration is considered as $c = (c_1, \dots, c_n)$

$c_i =$ number of alleles represented i times

and $|c| = c_1 + 2c_2 + \dots + nc_n = n$.

$q(c) := \mathbb{P}(\text{sample of size } |c| \text{ taken at stationarity has configuration } c)$,
with $q(e_1) = 1$.

Theorem

In a stationary sample of size n , the probability of sample configuration c is

$$q(c) = \mathbb{P}(C_1(n) = c_1, \dots, C_n(n) = c_n) = 1_{\{|c|=n\}} \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \binom{\theta}{j}^{c_j} \frac{1}{c_j!}$$

where $x_{(j)} = x(x+1)\dots(x+j-1)$, $j = 1, 2, \dots$ and
 $|c| = c_1 + 2c_2 + \dots + nc_n$.

- ESF WITH RECOMBINATION FOR A SAMPLE OF SIZE
 $n = 2, 3$

- 2-locus infinite-alleles model.
- loci are denoted by L and R .
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L is the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R ,
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R ,
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

- 2-locus infinite-alleles model.
- loci are denoted by L and R.
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L is the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R.
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R.
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

- 2-locus infinite-alleles model.
- loci are denoted by L and R .
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L : the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R ,
 - n_R : the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R ,
 - n_{LR} : the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

- 2-locus infinite-alleles model.
- loci are denoted by L and R.
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R,
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R,
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

- 2-locus infinite-alleles model.
- loci are denoted by L and R.
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R,
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R,
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

- 2-locus infinite-alleles model.
- loci are denoted by L and R.
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R,
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R,
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

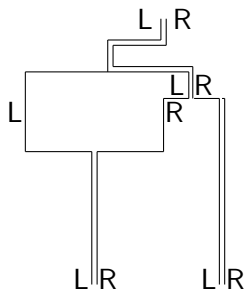
- 2-locus infinite-alleles model.
- loci are denoted by L and R .
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R ,
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R ,
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

The Model

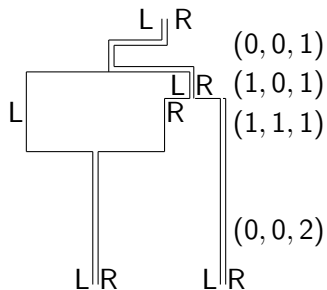
- 2-locus infinite-alleles model.
- loci are denoted by L and R .
- T_L (resp. T_R):= time that locus L (resp. R) finds its MRCA.
- Sample configuration (n_L, n_R, n_{LR}) ,
 - n_L the number of genes with ancestral alleles at locus L and non-ancestral alleles at locus R ,
 - n_R is the number of genes with non-ancestral alleles at locus L and ancestral alleles at locus R ,
 - n_{LR} is the number of genes with ancestral alleles at both loci.
- Recombination rate ρ per individual and the rate of coalescent per pair is 1.
- Remark: Only those recombination events that happen to those individuals having ancestral alleles at both loci will be taken into account.

A POPULATION OF SIZE 2

A population of size 2



A population of size 2



A population of size 2

- X_t : Continuous-time Markov chain which starts at state 1. The waiting time in state i is exponential with parameter λ_i . Then the process goes to other states according to its transition matrix P and stops whenever it reaches the state $(0, 0, 1)$.

$$\lambda_1 = 1 + 2\rho, \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 1, \lambda_5 = 3 + \rho, \lambda_9 = 6.$$

- The states are

$$1 = (0, 0, 2), 2 = (0, 0, 1), 3 = (0, 1, 1), 4 = (1, 0, 1), 5 = (1, 1, 1), \\ 6 = (1, 1, 0), 7 = (1, 2, 0), 8 = (2, 1, 0), 9 = (2, 2, 0).$$

A population of size 2

- X_t : Continuous-time Markov chain which starts at state 1. The waiting time in state i is exponential with parameter λ_i . Then the process goes to other states according to its transition matrix P and stops whenever it reaches the state $(0, 0, 1)$.

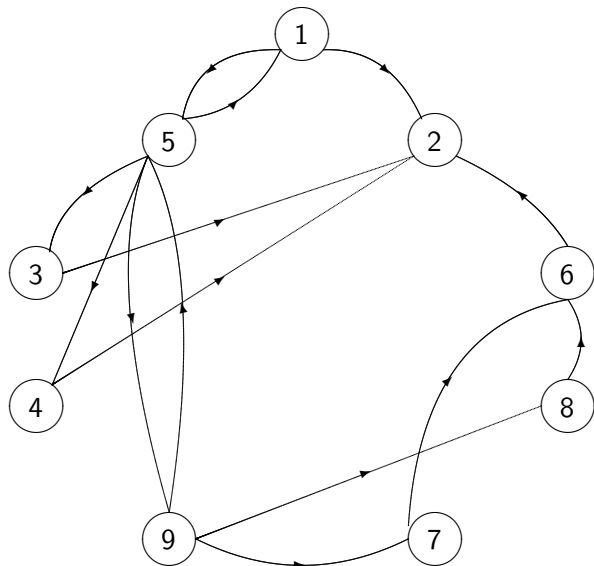
$$\lambda_1 = 1 + 2\rho, \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 1, \lambda_5 = 3 + \rho, \lambda_9 = 6.$$

- The states are

$$1 = (0, 0, 2), 2 = (0, 0, 1), 3 = (0, 1, 1), 4 = (1, 0, 1), 5 = (1, 1, 1), \\ 6 = (1, 1, 0), 7 = (1, 2, 0), 8 = (2, 1, 0), 9 = (2, 2, 0).$$

THE JOINT DISTRIBUTION OF T_L AND T_R

The joint distribution of T_L and T_R



The joint distribution of T_L and T_R

$$P = \begin{pmatrix} 0 & \frac{1}{1+2\rho} & 0 & 0 & \frac{2\rho}{1+2\rho} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3+\rho} & 0 & \frac{1}{3+\rho} & \frac{1}{3+\rho} & 0 & 0 & 0 & 0 & \frac{\rho}{3+\rho} \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{4}{6} & 0 & \frac{1}{6} & \frac{1}{6} & 0 \end{pmatrix}$$

The joint distribution of T_L and T_R

- $\nu_i :=$ the number of visits to state i and $V := (\nu_1, \nu_2, \dots, \nu_9)$ the vector of number of visits.
- So we have

Table: Number of visits to each states.

case	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	ν_7	ν_8	ν_9	path
1	≥ 1	1	0	0	$\nu_1 + \nu_9 - 1$	0	0	0	≥ 0	12
2	≥ 1	1	0	1	$\nu_1 + \nu_9$	0	0	0	≥ 0	54-42
3	≥ 1	1	1	0	$\nu_1 + \nu_9$	0	0	0	≥ 0	53-32
4	≥ 1	1	0	0	$\nu_1 + \nu_9 - 1$	1	0	1	≥ 1	98-862
5	≥ 1	1	0	0	$\nu_1 + \nu_9 - 1$	1	1	0	≥ 1	97-762

The joint distribution of T_L and T_R

- $\nu_i :=$ the number of visits to state i and $V := (\nu_1, \nu_2, \dots, \nu_9)$ the vector of number of visits.
- So we have

Table: Number of visits to each states.

case	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	ν_7	ν_8	ν_9	path
1	≥ 1	1	0	0	$\nu_1 + \nu_9 - 1$	0	0	0	≥ 0	12
2	≥ 1	1	0	1	$\nu_1 + \nu_9$	0	0	0	≥ 0	54-42
3	≥ 1	1	1	0	$\nu_1 + \nu_9$	0	0	0	≥ 0	53-32
4	≥ 1	1	0	0	$\nu_1 + \nu_9 - 1$	1	0	1	≥ 1	98-862
5	≥ 1	1	0	0	$\nu_1 + \nu_9 - 1$	1	1	0	≥ 1	97-762

The joint distribution of T_L and T_R

- $Q(\nu) := \mathbb{P}(V = \nu (= (\nu_1, \nu_2, \dots, \nu_9)))$
- $Q_i(\nu)$ be $Q(\nu)$ given case i .

Theorem

$Q_i(\nu)$'s are given as follows,

$$Q_1(\nu) = \frac{1}{1+2\rho} H(\nu_1 - 1, \nu_9),$$

$$Q_2(\nu) = Q_3(\nu) = H(\nu_1, \nu_9),$$

$$Q_4(\nu) = Q_5(\nu) = \frac{\rho}{6} H(\nu_1, \nu_9 - 1),$$

where

$$H(x, y) = \left(\frac{2\rho}{(1+2\rho)(3+\rho)} \right)^x \left(\frac{2\rho}{3(3+\rho)} \right)^y \binom{x+y-1}{y}.$$

The joint distribution of T_L and T_R

- $Q(\nu) := \mathbb{P}(V = \nu (= (\nu_1, \nu_2, \dots, \nu_9)))$
- $Q_i(\nu)$ be $Q(\nu)$ given case i .

Theorem

$Q_i(\nu)$'s are given as follows,

$$Q_1(\nu) = \frac{1}{1+2\rho} H(\nu_1 - 1, \nu_9),$$

$$Q_2(\nu) = Q_3(\nu) = H(\nu_1, \nu_9),$$

$$Q_4(\nu) = Q_5(\nu) = \frac{\rho}{6} H(\nu_1, \nu_9 - 1),$$

where

$$H(x, y) = \left(\frac{2\rho}{(1+2\rho)(3+\rho)} \right)^x \left(\frac{2\rho}{3(3+\rho)} \right)^y \binom{x+y-1}{y}.$$

The joint distribution of T_L and T_R

- $Q(\nu) := \mathbb{P}(V = \nu (= (\nu_1, \nu_2, \dots, \nu_9)))$
- $Q_i(\nu)$ be $Q(\nu)$ given case i .

Theorem

$Q_i(\nu)$'s are given as follows,

$$Q_1(\nu) = \frac{1}{1 + 2\rho} H(\nu_1 - 1, \nu_9),$$

$$Q_2(\nu) = Q_3(\nu) = H(\nu_1, \nu_9),$$

$$Q_4(\nu) = Q_5(\nu) = \frac{\rho}{6} H(\nu_1, \nu_9 - 1),$$

where

$$H(x, y) = \left(\frac{2\rho}{(1 + 2\rho)(3 + \rho)} \right)^x \left(\frac{2\rho}{3(3 + \rho)} \right)^y \binom{x + y - 1}{y}.$$

- SKETCH OF THE PROOF

Sketch of the proof

- Assume $P_{i,j}$ be the probability of going from state i to state j . We can easily find the following relations.
- For case 1

$$\begin{aligned} Q_1(\nu) &= \mathbb{P}(V = (\nu_1, 1, 0, 0, \nu_1 + \nu_9 - 1, 0, 0, 0, \nu_9)) \\ &= (P_{15}P_{51})^{\nu_1-1} (P_{59}P_{95})^{\nu_9} \binom{\nu_1 + \nu_9 - 2}{\nu_9} P_{12} \\ &= \frac{1}{1 + 2\rho} H(\nu_1 - 1, \nu_9). \end{aligned}$$

Sketch of the proof

- Assume $P_{i,j}$ be the probability of going from state i to state j . We can easily find the following relations.
- For case 1

$$\begin{aligned} Q_1(\nu) &= \mathbb{P}(V = (\nu_1, 1, 0, 0, \nu_1 + \nu_9 - 1, 0, 0, 0, \nu_9)) \\ &= (P_{15}P_{51})^{\nu_1-1} (P_{59}P_{95})^{\nu_9} \binom{\nu_1 + \nu_9 - 2}{\nu_9} P_{12} \\ &= \frac{1}{1 + 2\rho} H(\nu_1 - 1, \nu_9). \end{aligned}$$

The joint distribution of T_L and T_R

T_i := the total time spent in state i .

Given ν_i , T_i is a sum of ν_i i.i.d copies of exponential random variables $t_i \sim \text{Exp}(\lambda_i)$.

$T_i \sim \Gamma(\nu_i, \lambda_i)$ with the following probability density function

$$p(t; \nu_i, \lambda_i) = \begin{cases} \frac{\lambda_i^{\nu_i} t^{\nu_i-1} \exp(-\lambda_i t)}{(\nu_i-1)!} & t \geq 0 \\ 0 & t < 0, \end{cases}$$

where

$$\lambda_1 = 1 + 2\rho, \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 1, \lambda_5 = 3 + \rho, \lambda_9 = 6.$$

The joint distribution of T_L and T_R

T_i := the total time spent in state i .

Given ν_i , T_i is a sum of ν_i i.i.d copies of exponential random variables $t_i \sim \text{Exp}(\lambda_i)$.

$T_i \sim \Gamma(\nu_i, \lambda_i)$ with the following probability density function

$$p(t; \nu_i, \lambda_i) = \begin{cases} \frac{\lambda_i^{\nu_i} t^{\nu_i-1} \exp(-\lambda_i t)}{(\nu_i-1)!} & t \geq 0 \\ 0 & t < 0, \end{cases}$$

where

$$\lambda_1 = 1 + 2\rho, \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 1, \lambda_5 = 3 + \rho, \lambda_9 = 6.$$

The joint distribution of T_L and T_R

T_i := the total time spent in state i .

Given ν_i , T_i is a sum of ν_i i.i.d copies of exponential random variables $t_i \sim \text{Exp}(\lambda_i)$.

$T_i \sim \Gamma(\nu_i, \lambda_i)$ with the following probability density function

$$p(t; \nu_i, \lambda_i) = \begin{cases} \frac{\lambda_i^{\nu_i} t^{\nu_i-1} \exp(-\lambda_i t)}{(\nu_i-1)!} & t \geq 0 \\ 0 & t < 0, \end{cases}$$

where

$$\lambda_1 = 1 + 2\rho, \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 1, \lambda_5 = 3 + \rho, \lambda_9 = 6.$$

The joint distribution of T_L and T_R

T_i := the total time spent in state i .

Given ν_i , T_i is a sum of ν_i i.i.d copies of exponential random variables $t_i \sim \text{Exp}(\lambda_i)$.

$T_i \sim \Gamma(\nu_i, \lambda_i)$ with the following probability density function

$$p(t; \nu_i, \lambda_i) = \begin{cases} \frac{\lambda_i^{\nu_i} t^{\nu_i-1} \exp(-\lambda_i t)}{(\nu_i-1)!} & t \geq 0 \\ 0 & t < 0, \end{cases}$$

where

$$\lambda_1 = 1 + 2\rho, \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 1, \lambda_5 = 3 + \rho, \lambda_9 = 6.$$

The joint distribution of T_L and T_R

The joint probability density function of T_L and T_R given V which is given as follows

$$\Phi_{T_L, T_R|V}(t, s|V = \nu) = \begin{cases} \delta(t - s)f_{T|V}(t|V = \nu) & \text{case 1} \\ 1_{\{s < t\}}f_{T|V}(s|V = \nu)f(t - s) & \text{case 2} \\ 1_{\{t < s\}}f_{T|V}(t|V = \nu)f(s - t) & \text{case 3} \\ 1_{\{s < t\}}f_{T|V}(s|V = \nu)f(t - s) & \text{case 4} \\ 1_{\{t < s\}}f_{T|V}(t|V = \nu)f(s - t) & \text{case 5} \end{cases}$$

where $f \sim \text{Exp}(1)$.

The joint distribution of T_L and T_R

The joint probability density function of T_L and T_R given V which is given as follows

$$\Phi_{T_L, T_R|V}(t, s|V = \nu) = \begin{cases} \delta(t - s)f_{T|V}(t|V = \nu) & \text{case 1} \\ 1_{\{s < t\}}f_{T|V}(s|V = \nu)f(t - s) & \text{case 2} \\ 1_{\{t < s\}}f_{T|V}(t|V = \nu)f(s - t) & \text{case 3} \\ 1_{\{s < t\}}f_{T|V}(s|V = \nu)f(t - s) & \text{case 4} \\ 1_{\{t < s\}}f_{T|V}(t|V = \nu)f(s - t) & \text{case 5} \end{cases}$$

where $f \sim \text{Exp}(1)$.

The joint distribution of T_L and T_R

- So we have to find the distribution of $T := T_1 + T_5 + T_9$.
- If $\nu_5 = \nu_9 = 0$, the only possibility is when $\nu_1 = 1$. Therefore

$$f_{T|V}(t|V = \nu) = (1 + 2\rho)\exp(-(1 + 2\rho)t),$$

- If $\nu_5 \geq 1, \nu_9 = 0$, we have $\nu_5 = \nu_1 - 1$. So the distribution of T given V is given as follows

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1-1} \exp(-\lambda_5 t)}{(\nu_1 - 1)!(\nu_1 - 2)!} \sum_{k=0}^{\nu_1-1} \sum_{r=0}^{\infty} \binom{\nu_1 - 1}{k} \frac{(-1)^k (2 - \rho)^r}{r!(\nu_1 + k + r)} t^{2\nu_1 + r - 2},$$

- If $\nu_5 \neq 0, \nu_9 \neq 0$, the distribution of T given V is given by

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1 + \nu_9 - 1} \lambda_9^{\nu_9} \exp(-\lambda_1 t)}{(\nu_1 - 1)!(\nu_1 + \nu_9 - 2)!(\nu_9 - 1)!} t^{2(\nu_1 + \nu_9 - 1)} \sum_{r=0}^{\infty} \sum_{k=0}^{\nu_1 + \nu_9 - 2} \sum_{j=0}^{\infty} \sum_{i=0}^{\nu_9 - 1} \binom{\nu_1 + \nu_9 - 2}{k} \binom{\nu_9 - 1}{i} \frac{(-1)^{k+i} (\rho - 2)^r (\rho - 3)^j}{r!j!(\nu_9 + k + r)(2\nu_9 + \nu_1 + r + i + j - 1)} t^{r+j}.$$

The joint distribution of T_L and T_R

- So we have to find the distribution of $T := T_1 + T_5 + T_9$.
- If $\nu_5 = \nu_9 = 0$, the only possibility is when $\nu_1 = 1$. Therefore

$$f_{T|V}(t|V = \nu) = (1 + 2\rho)\exp(-(1 + 2\rho)t),$$

- If $\nu_5 \geq 1, \nu_9 = 0$, we have $\nu_5 = \nu_1 - 1$. So the distribution of T given V is given as follows

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1-1} \exp(-\lambda_5 t)}{(\nu_1 - 1)!(\nu_1 - 2)!} \sum_{k=0}^{\nu_1-1} \sum_{r=0}^{\infty} \binom{\nu_1 - 1}{k} \frac{(-1)^k (2 - \rho)^r}{r!(\nu_1 + k + r)} t^{2\nu_1 + r - 2},$$

- If $\nu_5 \neq 0, \nu_9 \neq 0$, the distribution of T given V is given by

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1 + \nu_9 - 1} \lambda_9^{\nu_9} \exp(-\lambda_1 t)}{(\nu_1 - 1)!(\nu_1 + \nu_9 - 2)!(\nu_9 - 1)!} t^{2(\nu_1 + \nu_9 - 1)} \sum_{r=0}^{\infty} \sum_{k=0}^{\nu_1 + \nu_9 - 2} \sum_{j=0}^{\infty} \sum_{i=0}^{\nu_9 - 1} \binom{\nu_1 + \nu_9 - 2}{k} \binom{\nu_9 - 1}{i} \frac{(-1)^{k+i} (\rho - 2)^r (\rho - 3)^j}{r!j!(\nu_9 + k + r)(2\nu_9 + \nu_1 + r + i + j - 1)} t^{r+j}.$$

The joint distribution of T_L and T_R

- So we have to find the distribution of $T := T_1 + T_5 + T_9$.
- If $\nu_5 = \nu_9 = 0$, the only possibility is when $\nu_1 = 1$. Therefore

$$f_{T|V}(t|V = \nu) = (1 + 2\rho)\exp(-(1 + 2\rho)t),$$

- If $\nu_5 \geq 1, \nu_9 = 0$, we have $\nu_5 = \nu_1 - 1$. So the distribution of T given V is given as follows

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1-1} \exp(-\lambda_5 t)}{(\nu_1 - 1)!(\nu_1 - 2)!} \sum_{k=0}^{\nu_1-1} \sum_{r=0}^{\infty} \binom{\nu_1 - 1}{k} \frac{(-1)^k (2 - \rho)^r}{r!(\nu_1 + k + r)} t^{2\nu_1+r-2},$$

- If $\nu_5 \neq 0, \nu_9 \neq 0$, the distribution of T given V is given by

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1+\nu_9-1} \lambda_9^{\nu_9} \exp(-\lambda_1 t)}{(\nu_1 - 1)!(\nu_1 + \nu_9 - 2)!(\nu_9 - 1)!} t^{2(\nu_1+\nu_9-1)} \sum_{r=0}^{\infty} \sum_{k=0}^{\nu_1+\nu_9-2} \sum_{j=0}^{\infty} \sum_{i=0}^{\nu_9-1} \binom{\nu_1 + \nu_9 - 2}{k} \binom{\nu_9 - 1}{i} \frac{(-1)^{k+i} (\rho - 2)^r (\rho - 3)^j}{r!j!(\nu_9 + k + r)(2\nu_9 + \nu_1 + r + i + j - 1)} t^{r+j}.$$

The joint distribution of T_L and T_R

- So we have to find the distribution of $T := T_1 + T_5 + T_9$.
- If $\nu_5 = \nu_9 = 0$, the only possibility is when $\nu_1 = 1$. Therefore

$$f_{T|V}(t|V = \nu) = (1 + 2\rho)\exp(-(1 + 2\rho)t),$$

- If $\nu_5 \geq 1, \nu_9 = 0$, we have $\nu_5 = \nu_1 - 1$. So the distribution of T given V is given as follows

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1-1} \exp(-\lambda_5 t)}{(\nu_1 - 1)!(\nu_1 - 2)!} \sum_{k=0}^{\nu_1-1} \sum_{r=0}^{\infty} \binom{\nu_1 - 1}{k} \frac{(-1)^k (2 - \rho)^r}{r!(\nu_1 + k + r)} t^{2\nu_1+r-2},$$

- If $\nu_5 \neq 0, \nu_9 \neq 0$, the distribution of T given V is given by

$$f_{T|V}(t|V = \nu) = \frac{\lambda_1^{\nu_1} \lambda_5^{\nu_1+\nu_9-1} \lambda_9^{\nu_9} \exp(-\lambda_1 t)}{(\nu_1 - 1)!(\nu_1 + \nu_9 - 2)!(\nu_9 - 1)!} t^{2(\nu_1+\nu_9-1)} \sum_{r=0}^{\infty} \sum_{k=0}^{\nu_1+\nu_9-2} \sum_{j=0}^{\infty} \sum_{i=0}^{\nu_9-1} \binom{\nu_1 + \nu_9 - 2}{k} \binom{\nu_9 - 1}{i} \frac{(-1)^{k+i} (\rho - 2)^r (\rho - 3)^j}{r!j!(\nu_9 + k + r)(2\nu_9 + \nu_1 + r + i + j - 1)} t^{r+j}.$$

Theorem

The joint distribution of T_L and T_R is given as follows

$$\begin{aligned}\mathbb{P}(ds, dt) &= \sum_{\nu_1=1}^{\infty} \sum_{\nu_9=0}^{\infty} Q_1(V) \delta(t-s) f_{T|V}(t|V=\nu) dt \\ &+ \sum_{\nu_1=1}^{\infty} \sum_{\nu_9=0}^{\infty} Q_2(V) (\exp(-(t-s)) f_{T|V}(s|V=\nu) 1_{\{s<t\}} + \exp(-(s-t)) f_{T|V}(t|V=\nu) 1_{\{t<s\}}) dt ds \\ &+ \sum_{\nu_1=1}^{\infty} \sum_{\nu_9=1}^{\infty} Q_4(V) (\exp(-(t-s)) f_{T|V}(s|V=\nu) 1_{\{s<t\}} + \exp(-(s-t)) f_{T|V}(t|V=\nu) 1_{\{t<s\}}) dt ds.\end{aligned}$$

The joint distribution of T_L and T_R

- L_i (resp. R_i):= locus L (resp. R) for the i -th individual
- $L_i = L_j$ (resp. $R_i = R_j$) := i -th and the j -th individuals are identical at locus L (resp. R).
-

$$\begin{aligned}\mathbb{P}(L_i = L_j, R_i = R_j) &= \mathbb{E}(\mathbb{P}(L_i = L_j | T_L) \mathbb{P}(R_i = R_j | T_R)) \\ &= \mathbb{E}(\exp(-2\theta T) \mathbf{1}_{\{T_L = T = T_R\}}) + \mathbb{E}(\exp(-\theta(T_L + T_R)) \mathbf{1}_{\{T_L < T_R\}}) \\ &\quad + \mathbb{E}(\exp(-\theta(T_L + T_R)) \mathbf{1}_{\{T_R < T_L\}}).\end{aligned}$$

The joint distribution of T_L and T_R

- L_i (resp. R_i):= locus L (resp. R) for the i -th individual
- $L_i = L_j$ (resp. $R_i = R_j$) := i -th and the j -th individuals are identical at locus L (resp. R).



$$\begin{aligned}\mathbb{P}(L_i = L_j, R_i = R_j) &= \mathbb{E}(\mathbb{P}(L_i = L_j | T_L) \mathbb{P}(R_i = R_j | T_R)) \\ &= \mathbb{E}(\exp(-2\theta T) \mathbf{1}_{\{T_L = T = T_R\}}) + \mathbb{E}(\exp(-\theta(T_L + T_R)) \mathbf{1}_{\{T_L < T_R\}}) \\ &\quad + \mathbb{E}(\exp(-\theta(T_L + T_R)) \mathbf{1}_{\{T_R < T_L\}}).\end{aligned}$$

The joint distribution of T_L and T_R

- L_i (resp. R_i):= locus L (resp. R) for the i -th individual
- $L_i = L_j$ (resp. $R_i = R_j$) := i -th and the j -th individuals are identical at locus L (resp. R).
-

$$\begin{aligned}\mathbb{P}(L_i = L_j, R_i = R_j) &= \mathbb{E}(\mathbb{P}(L_i = L_j | T_L) \mathbb{P}(R_i = R_j | T_R)) \\ &= \mathbb{E}(\exp(-2\theta T) \mathbf{1}_{\{T_L = T = T_R\}}) + \mathbb{E}(\exp(-\theta(T_L + T_R)) \mathbf{1}_{\{T_L < T_R\}}) \\ &\quad + \mathbb{E}(\exp(-\theta(T_L + T_R)) \mathbf{1}_{\{T_R < T_L\}}).\end{aligned}$$

Theorem

$$\mathbb{P}(L_1 = L_2, R_1 = R_2) = \frac{2\rho^2 + (\theta^2 + 6\theta + 13)\rho + 2\theta^3 + 11\theta^2 + 18\theta + 9}{(1 + \theta)[2(1 + \theta)\rho^2 + (6\theta^2 + 19\theta + 13)\rho + (1 + 2\theta)(3 + 2\theta)(3 + \theta)]},$$

$$\mathbb{P}(L_1 = L_2, R_1 \neq R_2) = \frac{2\theta\rho^2 + (5\theta^2 + 13\theta)\rho + 2\theta^3 + 9\theta^2 + 9\theta}{(1 + \theta)[2(1 + \theta)\rho^2 + (6\theta^2 + 19\theta + 13)\rho + (1 + 2\theta)(3 + 2\theta)(3 + \theta)]},$$

$$\mathbb{P}(L_1 \neq L_2, R_1 \neq R_2) = \frac{2\theta^2\rho^2 + (6\theta^3 + 14\theta^2)\rho + 4\theta^4 + 18\theta^3 + 18\theta^2}{(1 + \theta)[2(1 + \theta)\rho^2 + (6\theta^2 + 19\theta + 13)\rho + (1 + 2\theta)(3 + 2\theta)(3 + \theta)]}.$$

Remarks

- $\rho = 0$, $\mathbb{P}(L_1 = L_2, R_1 = R_2) = \frac{1}{1+2\theta}$, and the probability that these two individuals are not of the same type is $\frac{2\theta}{1+2\theta}$.
- If $\rho \rightarrow \infty$ then $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow \left(\frac{1}{1+\theta}\right)^2$.
- When $\frac{\theta}{2}$, the mutation rate per individual per locus is zero, $\mathbb{P}(L_i = L_j, R_i = R_j) = 1$.
- When $\frac{\theta}{2} \rightarrow \infty$, $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow 0$.

Remarks

- $\rho = 0$, $\mathbb{P}(L_1 = L_2, R_1 = R_2) = \frac{1}{1+2\theta}$, and the probability that these two individuals are not of the same type is $\frac{2\theta}{1+2\theta}$.
- If $\rho \rightarrow \infty$ then $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow \left(\frac{1}{1+\theta}\right)^2$.
- When $\frac{\theta}{2}$, the mutation rate per individual per locus is zero, $\mathbb{P}(L_i = L_j, R_i = R_j) = 1$.
- When $\frac{\theta}{2} \rightarrow \infty$, $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow 0$.

Remarks

- $\rho = 0$, $\mathbb{P}(L_1 = L_2, R_1 = R_2) = \frac{1}{1+2\theta}$, and the probability that these two individuals are not of the same type is $\frac{2\theta}{1+2\theta}$.
- If $\rho \rightarrow \infty$ then $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow \left(\frac{1}{1+\theta}\right)^2$.
- When $\frac{\theta}{2}$, the mutation rate per individual per locus is zero, $\mathbb{P}(L_i = L_j, R_i = R_j) = 1$.
- When $\frac{\theta}{2} \rightarrow \infty$, $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow 0$.

Remarks

- $\rho = 0$, $\mathbb{P}(L_1 = L_2, R_1 = R_2) = \frac{1}{1+2\theta}$, and the probability that these two individuals are not of the same type is $\frac{2\theta}{1+2\theta}$.
- If $\rho \rightarrow \infty$ then $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow \left(\frac{1}{1+\theta}\right)^2$.
- When $\frac{\theta}{2}$, the mutation rate per individual per locus is zero, $\mathbb{P}(L_i = L_j, R_i = R_j) = 1$.
- When $\frac{\theta}{2} \rightarrow \infty$, $\mathbb{P}(L_i = L_j, R_i = R_j) \rightarrow 0$.

A POPULATION OF SIZE $n = 3$

A population of size $n = 3$

- D : := the number of different types for the individuals in the population, so $D \in \{1, 2, 3\}$.
- T_1 : := time to the first coalescent and T_2 : := be the time to the second coalescent, so $T_1 \sim \text{Exp}(3)$ and $T_2 \sim \text{Exp}(1)$.



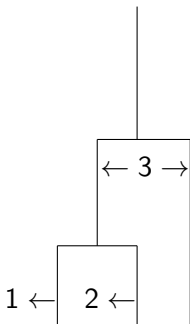
A population of size $n = 3$

- D := the number of different types for the individuals in the population, so $D \in \{1, 2, 3\}$.
- T_1 := time to the first coalescent and T_2 := be the time to the second coalescent, so $T_1 \sim \text{Exp}(3)$ and $T_2 \sim \text{Exp}(1)$.



A population of size $n = 3$

- D := the number of different types for the individuals in the population, so $D \in \{1, 2, 3\}$.
- T_1 := time to the first coalescent and T_2 := be the time to the second coalescent, so $T_1 \sim \text{Exp}(3)$ and $T_2 \sim \text{Exp}(1)$.



A population of size $n = 3$

For $D = 3$.

- At least one mutation on the branch indexed by 1 (or 2) and at least one mutation on the branches indexed by 3.
- At least one mutation on the branches labeled by 1 and 2 and no mutation on 3.

$$\begin{aligned}\mathbb{P}(D = 3|\mathcal{T}) &= 1 - \exp(-\theta T_1) - 2 \exp(\theta(T_1 + T_2)) \\ &\quad + 2 \exp\left(-\frac{3\theta T_1}{2} - \theta T_2\right).\end{aligned}$$

A population of size $n = 3$

For $D = 3$.

- At least one mutation on the branch indexed by 1 (or 2) and at least one mutation on the branches indexed by 3.
- At least one mutation on the branches labeled by 1 and 2 and no mutation on 3.

$$\begin{aligned}\mathbb{P}(D = 3|\mathcal{T}) &= 1 - \exp(-\theta T_1) - 2 \exp(\theta(T_1 + T_2)) \\ &\quad + 2 \exp\left(-\frac{3\theta T_1}{2} - \theta T_2\right).\end{aligned}$$

A population of size $n = 3$

- For $D = 2$. So for this case we can impose at least one mutation either on the branch called 1, or on the branch called 2, or 3. So

$$\mathbb{P}(D = 2|\mathcal{T}) = \exp(-\theta T_1) - \exp(-\theta(T_1 + T_2)) - 2 \exp(-\frac{3\theta}{2} T_1 - \theta T_2).$$

- For $D = 1$. This case can happen only if we have no mutation neither on 1 nor on 2 nor on 3. So

$$\mathbb{P}(D = 1|\mathcal{T}) = \exp(-\frac{\theta}{2}(2T_2 + 3T_1)).$$

A population of size $n = 3$

- For $D = 2$. So for this case we can impose at least one mutation either on the branch called 1, or on the branch called 2, or 3. So

$$\mathbb{P}(D = 2|\mathcal{T}) = \exp(-\theta T_1) - \exp(-\theta(T_1 + T_2)) - 2 \exp(-\frac{3\theta}{2} T_1 - \theta T_2).$$

- For $D = 1$. This case can happen only if we have no mutation neither on 1 nor on 2 nor on 3. So

$$\mathbb{P}(D = 1|\mathcal{T}) = \exp(-\frac{\theta}{2}(2T_2 + 3T_1)).$$

A population of size $n = 3$

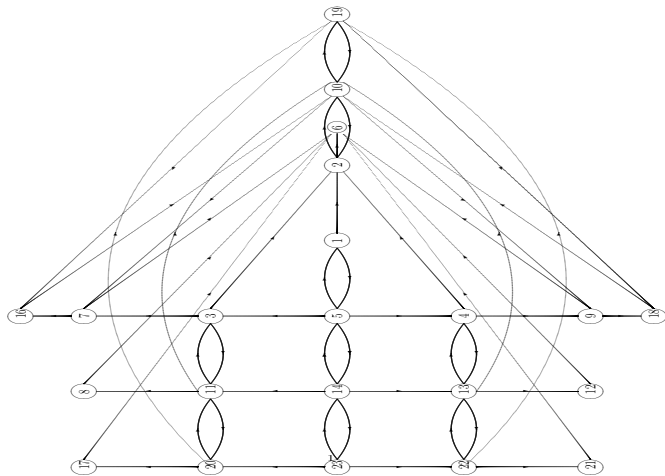
Assume D_L (resp. D_R) is the same as D for locus L (resp. R) and \mathcal{T}^L (resp. \mathcal{T}^R) is the same as \mathcal{T} for locus L (resp. R). So for $i, j \in \{1, 2, 3\}$ we have

$$\begin{aligned}\mathbb{P}(D_L = i, D_R = j) &= \mathbb{E}(\mathbb{P}(D_L = i, D_R = j | ARG)) \\ &= \mathbb{E}(\mathbb{P}(D_L = i | \mathcal{T}^L) \mathbb{P}(D_R = j | \mathcal{T}^R)).\end{aligned}$$

A population of size $n = 3$

1 = (0, 0, 3), 2 = (0, 0, 2), 3 = (0, 1, 2), 4 = (1, 0, 2), 5 = (1, 1, 2), 6 = (0, 0, 1)
7 = (0, 1, 1), 8 = (0, 2, 1), 9 = (1, 0, 1), 10 = (1, 1, 1), 11 = (1, 2, 1)
12 = (2, 0, 1), 13 = (2, 1, 1), 14 = (2, 2, 1), 15 = (1, 1, 0), 16 = (1, 2, 0)
17 = (1, 3, 0), 18 = (2, 1, 0), 19 = (2, 2, 0), 20 = (2, 3, 0), 21 = (3, 1, 0)
22 = (3, 2, 0), 23 = (3, 3, 0).

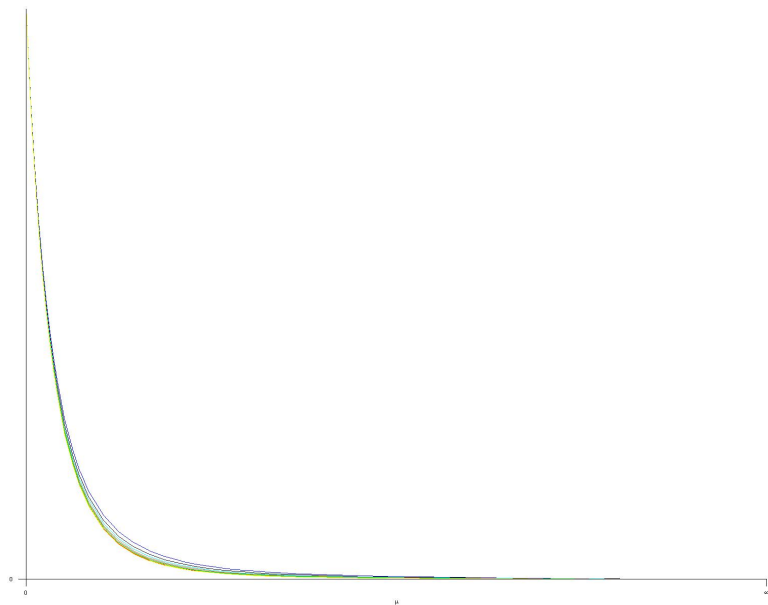
A population of size $n = 3$



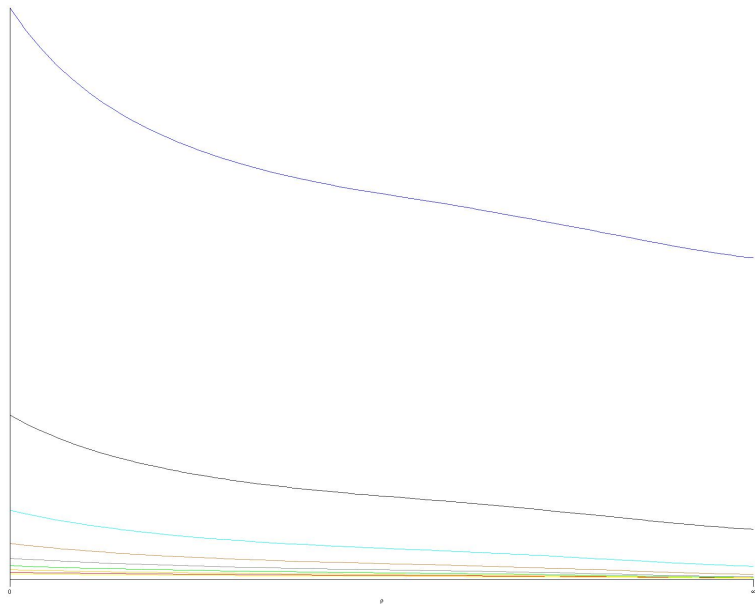
GENERAL CASE

Assume we have a population of size n . As before we consider the 2-locus infinite-alleles model, $n^2 + \sum_{k=1}^n k^2$ different states

The probability that all individuals are of the same type for fixed ρ



The probability that all individuals are of the same type for fixed θ



Thanks for your attention.