A Coalescent Dual Process in a Moran model with Genic Selection

Bob Griffiths

University of Oxford

Joint research with Alison Etheridge.

A Moran model with selection. N individuals and type space $[d] = \{1, \ldots, d\}$. Birth-death events occur at birth-rate λ_j to type j individuals. Mutation from type i to type j individuals occurs at rate μp_{ij}, where $P = (p_{ij})$ is a transition matrix. Once a birth takes place an individual is chosen uniformly to die. $z = (z_1, \ldots, z_d)$ are numbers of individuals of types in $[d]$.

The rate of events $z \to z - e_i + e_j$ is

$$\varphi(z, z - e_i + e_j) = \lambda_j z_j \frac{z_i}{N} + \mu z_i p_{ij}, \; i, j \in [d].$$

Generator

$$\mathcal{L}f(z) = \sum_{i,j=1}^{d} \varphi(z, z - e_i + e_j)\Big(f(z - e_i + e_j) - f(z)\Big)$$

History.

Krone and Neuhauser (2007), Neuhauser and Krone (2007). The Ancestral Selection Graph. A branching coalescing ancestral graph. When the graph has j edges coalescence is at rate $\binom{j}{2}$ and branching at rate $\sigma j/2$. The true genealogy of a sample is extracted from the graph starting from the ultimate ancestor and following lines to the current time with the fittest individual being transmitted at a branch point.

Donnelly and Kurtz (1999). A construction of a Fleming-Viot process with selection and simultaneously the ancestral selection graph using a modified lookdown approach.

Stephens and Donnelly (2002). Ancestry in a general diploid selection model.

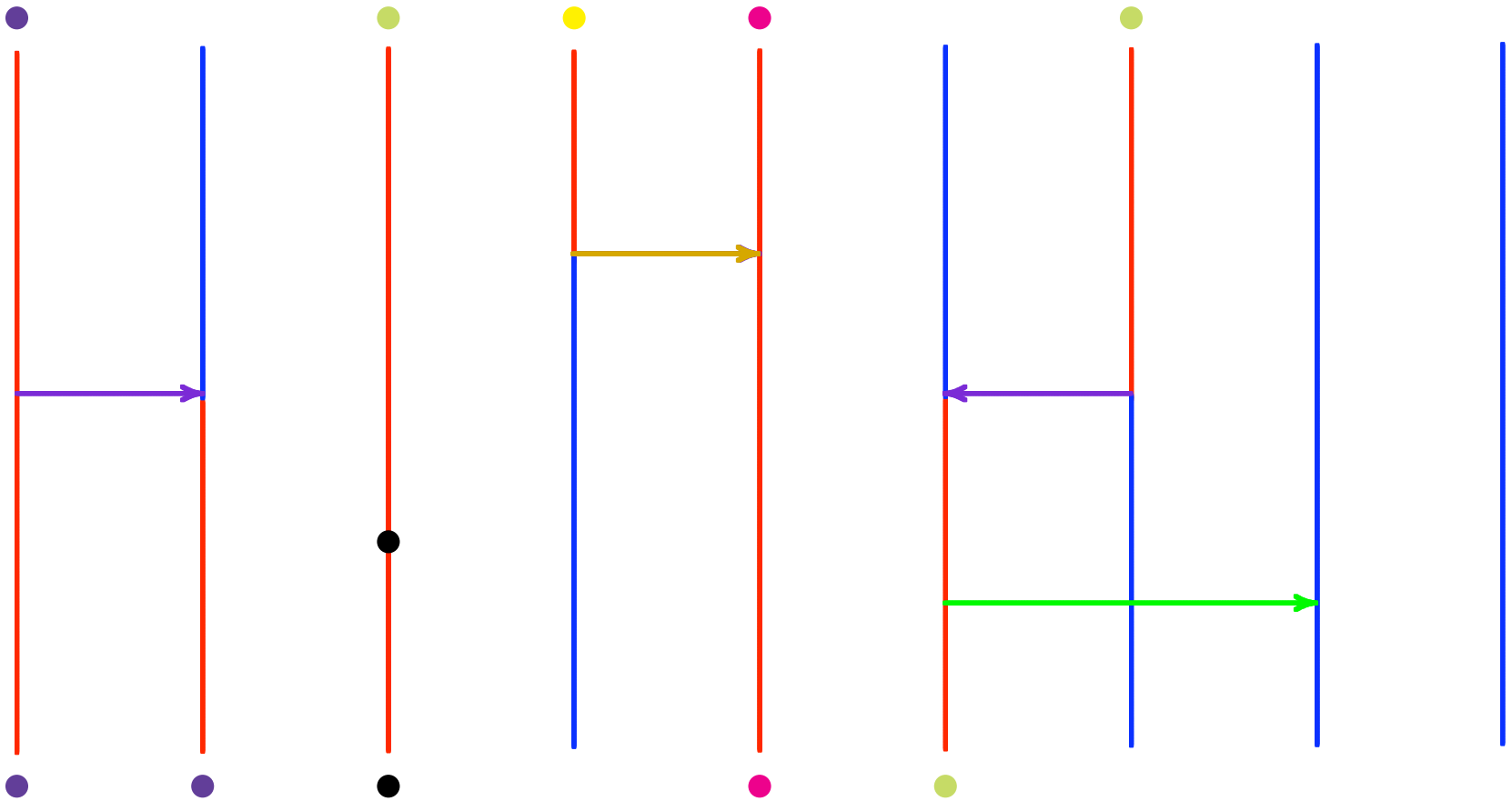
Fearnhead (2002). Ancestry in a genic selection model following back ancestral typed lines.

Barbour Ethier and Griffiths (2000). A transition density expansion using an ancestral dual process.

Moran model graphical description with N lines
Forward and Backward Dual Processes

Forward. Arrows are shot between ordered pairs of lines at a rate of λ^*/N per pair. An arrow where the tail is of type j modifies the type of the line at the head with probability λ_j/λ^*, or leaves the type unaltered with probability $\lambda_j^*/\lambda^* = 1 - \lambda_j/\lambda^*$. Mutations occur along the lines, with a line type i changing to a type j at rate μp_{ij}.

Backward. The dual process follows a typed sample of lines back in time. Lines are lost by coalescence or mutation. Virtual lines which attempt to change lines at the head of an arrow but do not succeed increase the number of lines in the dual process.



Time runs down the diagram. The Dual process has red lines.
Real arrows are purple. Virtual arrows are orange.

Moran model with parent independent mutation and genic selection.

N individuals and type space $[d]$.

Mutation rates $p_{ij} = p_j, i, j \in [d]$.

Rates $z \to z - e_i + e_j$ are

$$\varphi(z, z - e_i + e_j) = z_i \Big[\lambda_j \frac{z_j}{N} + \mu p_j \Big], \; i, j \in [d].$$

The model is reversible with stationary distribution a weighted multinomial-Dirichlet distribution

$$\pi(z;\theta,N) \propto \lambda_1^{z_1}\cdots\lambda_d^{z_d}\binom{N}{z}\frac{\theta_{1(z_1)}\cdots\theta_{d(z_d)}}{|\theta|_{(N)}},$$

where

$$\theta_i = \frac{N\mu p_i}{\lambda_i},\ i\in[d].$$

The scale constant in this distribution is

$$u(\theta,\lambda,N) = \mathbb{E}\Big[\Big(\sum_{j=1}^{d}\lambda_j\xi_j\Big)^N\Big],$$

where $\xi = (\xi_1,\ldots,\xi_d)$ has a $\mathcal{D}(\theta)$ distribution.

Diffusion process limit with weak selection $(X_1(t)\ldots,X_d(t))$ are gene frequencies of d types, labelled $1,\ldots,d$ at time $t \geq 0$. Mutations $i \to j$ occur at rate $\frac{1}{2}\theta_j$, $i,j = 1,\ldots d$. $\theta \geq 0$, with possibly some rates zero.

Generator

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} x_i(\delta_{ij}-x_j)\frac{\partial^2}{\partial x_i \partial x_j} + \frac{1}{2}\sum_{i=1}^{d}(\theta_i - |\theta| x_i + x_i(\sigma_i - s(x)))\frac{\partial}{\partial x_i}$$

where $s(x) = \sum_{j=1}^{d}\sigma_j x_j$.

Limit from the Moran model with

$$\lambda_j = \frac{N}{2}\Big(1+\frac{\sigma_j}{N}\Big),\ \mu p_j \to \frac{\theta_j}{2},\ N \to \infty$$

The stationary density when $\theta > 0$ is a weighted Dirichlet density proportional to

$$e^{\sum_{j=1}^{d} \sigma_j x_j} \frac{\Gamma(|\theta|)}{\Gamma(\theta_1) \cdots \Gamma(\theta_d)} x_1^{\theta_1 - 1} \cdots x_d^{\theta_d - 1}$$

for $x_1, \ldots x_d > 0$ and $\sum_{i=1}^{d} x_i = 1$.

Transition distribution in the diffusion

$$P(t, x, \cdot) = \sum_{l \in \mathbb{Z}_+^d} h_{xl}^{\infty}(t) \Pi_{l+\theta}(\cdot)$$

$\{h_{xl}^{\infty}(t), t \geq 0\}$ are transition functions of the dual limit process with an entrance boundary at infinity, with type frequencies having relative frequency $x = (x_1, \ldots, x_d)$ and $\Pi_{l+\theta}$ is the weighted Dirichlet $(\theta + l)$ distribution.

Barbour, A.D., Ethier, S.N., and Griffiths, R.C. (2000). A transition function expansion for a diffusion model with selection. Ann. Appl. Prob. 10, 123-162.

The Q matrix for the multitype birth and death process is given by

$$\begin{aligned} q(\beta, \beta + e_j) &= \frac{1}{2} \frac{|\sigma_j||\beta|(\beta_j + \theta_j)}{|\beta| + |\theta|} \times \frac{v(\theta + \beta + e_j)}{v(\theta + \beta)} \\ q(\beta, \beta - e_j) &= \frac{1}{2} \beta_j (|\theta| + |\beta| - 1) \times \frac{v(\beta + \theta - e_j)}{v(\theta + \beta)} \\ q(\beta, \beta) &= -\frac{1}{2} \Big[\sum_{j=1}^{d} \beta_j |\sigma_j| + |\beta|(|\beta| + |\theta| - 1) \Big] \end{aligned}$$

where

$$v(\gamma) = \mathbb{E}\Big[\exp\Big\{ \sum_{j=1}^{d} \sigma_j \xi_j \Big\} \Big]$$

with ξ having a Dirichlet (γ) distribution. $\sigma_j \leq 0$, $j \in [d]$.

Neutral Wright-Fisher diffusion $(X_1(t)\ldots,X_d(t))$ are gene frequencies of d types, labelled $1,\ldots,d$ at time $t\geq 0$.

Mutations $i\longrightarrow j$ occur at rate $\frac{1}{2}\theta_j$, $i,j=1,\ldots d$.
$\theta\geq 0$, with possibly some rates zero.

Generator

$$\mathcal{L}=\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}x_i(\delta_{ij}-x_j)\frac{\partial^2}{\partial x_i\partial x_j}+\frac{1}{2}\sum_{i=1}^{d}(\theta_i-|\theta|x_i)\frac{\partial}{\partial x_i}$$

If $\theta>0$ the stationary distribution is Dirichlet

$$\mathcal{D}(x,\theta)=\frac{\Gamma(|\theta|)}{\Gamma(\theta_1)\cdots\Gamma(\theta_d)}x_1^{\theta_1-1}\cdots x_d^{\theta_d-1}$$

for $x_1,\ldots x_d>0$ and $\sum_{i=1}^{d}x_i=1$

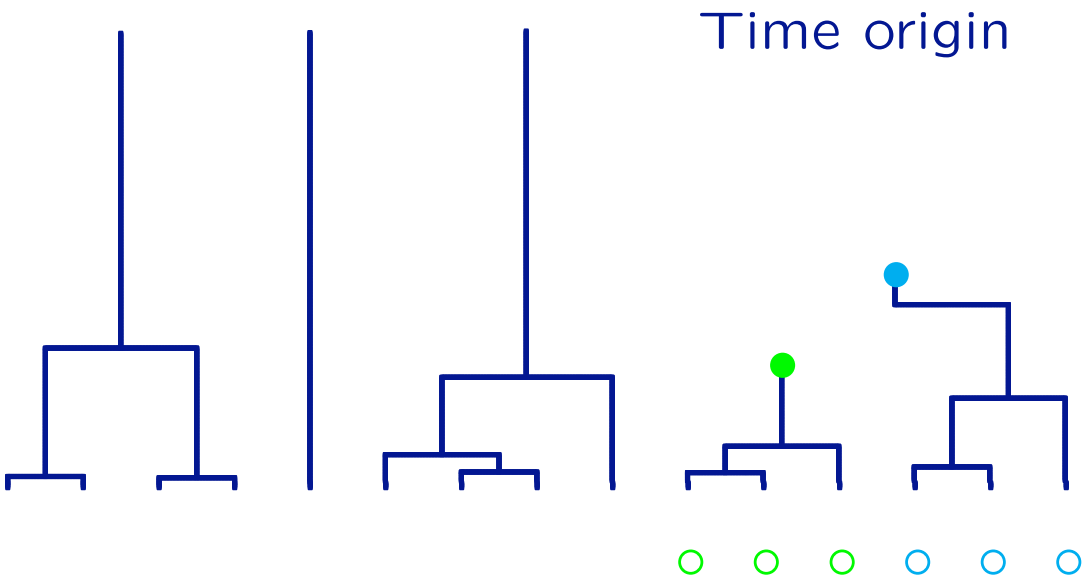
Transition distribution arising from an infinite-leaf coalescent

$$P(x,\cdot;t) = \sum_{k=0}^{\infty} q_k^{|\theta|}(t) \sum_{|l|=k} \text{mult}(l;k,x)\mathcal{D}(\cdot,\theta+l)$$

$q_k^{|\theta|}(t)$ is the distribution of $L^{|\theta|}(t)$, the number of non-mutant founder lineages at time t back. $L^{|\theta|}(t)$ is a death process back in time, starting from infinity, where lineages are reduced by coalescence or mutation from $k \to k-1$ at rate $k(k-1)/2+k|\theta|/2$.

Families are either from founder lineages or new mutations, giving the Dirichlet mixture.

Founder lineages and new mutations



Dirichlet family sizes

$$\sum_{|l|=k} \text{mult}(l; k, x)\mathcal{D}(\cdot, \theta + l)$$

$|l|$ non-mutant founder lineages are divided into $l = (l_1, \ldots, l_d)$ numbers of types $1, \ldots, d$ with probability $\text{mult}(l; k, x)$.
Let $U = (U_1, \ldots, U_k)$ be their relative family sizes in the leaves of the tree, and $V = (V_1, \ldots, V_d)$ be the frequencies of families derived from new mutations on the tree edges in $(0, t)$.
$U \oplus V = (U_1, \ldots, U_k, V_1, \ldots, V_d)$ is $\mathcal{D}(u \oplus v, (1, \ldots, 1) \oplus \theta)$.
$\mathcal{D}(y, \theta + l)$ is obtained by adding Dirichlet parameters corresponding to types $1, \ldots, d$.

Ethier, S.N. and Griffiths, R.C. (1993). The transition function of a Fleming-Viot process. Ann. Prob. 21, 1571-1590.

Dual Genealogical Process Derivation for two types

$$\mathcal{L} = \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} + \frac{1}{2}(\theta_1 - |\theta|x)\frac{\partial}{\partial x}$$

$|\theta| = \theta_1 + \theta_2$.

$x_1 = x, x_2 = 1 - x$, frequency of type 1 and 2 genes.

Define

$$g_k(x) = \frac{|\theta|_{(|k|)}}{\theta_{1(k_1)}\theta_{2(k_2)}} x_1^{k_1} x_2^{k_2}$$

If X is stationary with distribution π which is Beta (θ_1, θ_2)

$$\mathbb{E}^{\pi}[g_k(X)] = 1$$

Dual Generator acting on k

$$\mathcal{L}g_k(x) = \frac{1}{2}(|\theta| - 1 + |k|)\left[k_1 g_{k-e_1}(x) + k_2 g_{k-e_2}(x)\right] - |k| g_k(x)]$$

Dual process is a two dimensional death process $\{L^{|\theta|}(t), t \geq 1\}$ with rates

$$k \to k - e_i \quad \text{of} \quad \frac{1}{2}\frac{k_i}{|k|}|k|(|k| + |\theta| - 1)$$

$|L^{|\theta|}(t)|$ is a death process with rates

$$|k| \to |k| - 1 \quad \text{of} \quad \frac{1}{2}|k|(|k| + |\theta| - 1)$$

Hypergeometric sampling of types which do not 'die'

$$P(L^{|\theta|}(t) = l \mid L^{|\theta|}(0) = m) = q_{|m||l|}(t)\frac{\binom{m_1}{l_1}\binom{m_2}{l_2}}{\binom{|m|}{|l|}}$$

Dual representation

$$\mathbb{E}_{X(0)}\left[g_{L^{|\theta|}(0)}(X(t))\right] = \mathbb{E}_{L^{|\theta|}(0)}\left[g_{L^{|\theta|}(t)}(X(0))\right]$$

Transition density limit

Dual equation gives

$$\mathbb{E}_x\Big[\binom{|m|}{m_1} X_1(t)^{m_1} X_2(t)^{m_2}\Big]$$
$$= \binom{|m|}{m_1} \frac{\theta_{1(m_1)}\theta_{2(m_2)}}{|\theta|_{(m_1+m_2)}} \times \mathbb{E}_{L^{|\theta|}(0)}\Big[g_{L^{|\theta|}(t)}(X(0))\Big]$$

Sample inversion. $m_1, m_2 \to \infty$ with $m_1/|m| \to y_1, m_2/|m| \to y_2$ to obtain the transition density of the diffusion

$$f(x,y;t) = \sum_{l\in\mathbb{Z}_+^2} q_{|l|}^{|\theta|}(t)\binom{|l|}{l_1} x_1^{l_1} x_2^{l_2} B(\theta_1+l_1, \theta_2+l_2)^{-1} y_1^{l_1+\theta_1-1} y_2^{l_2+\theta_2-1}$$

Dual process representation in the Moran model

$$f_\alpha(z) = \prod_{j=1}^{d} z_{j[\alpha_j]}$$

$$g_\alpha(z) = \frac{f_\alpha(z)}{m_\alpha} = \frac{\mathcal{H}(\alpha \mid z)}{\mathcal{H}(\alpha)}$$

where $a_{[k]} = a(a-1)\cdots(a-k+1)$, and $m_k = \mathbb{E}^\pi\Big[f_k(Z)\Big]$.

$\mathcal{H}(\alpha \mid z)$ and $\mathcal{H}(\alpha)$ are the hypergeometric sampling distribution, given population frequencies z, and the unconditional sampling distribution mixed over z.

Important point

$$\mathbb{E}^\pi\Big[g_\alpha(Z)\Big] = 1$$

The generator equation acting on $g_k(z)$ can be written

$$\mathcal{L}g_\alpha(z) = \sum_\beta q(\alpha,\beta)g_\beta(z)$$

A Markov chain $\{L(t), t \geq 0\}$ in $\mathbb{Z}^d$ with rate matrix $Q = (q(\alpha,\beta))$ is dual to $\{Z(t), t \geq 0\}$ and the dual representation is

$$\mathbb{E}_{Z(0)}\Big[g_{L(0)}(Z(t))\Big] = \mathbb{E}_{L(0)}\Big[g_{L(t)}(Z(0))\Big],$$

where expectation on the left is with respect to $Z(t)$ and on the right with respect to $L(t)$.

Notation:
$\{L(t), t \geq 0\}$ has transition functions $\{h_{\alpha\beta}(t), \alpha, \beta \in \mathbb{Z}_+^d\}$.
$\{Z(t), t \geq 0\}$ has transition functions $\{s_{yz}(t), |y| = |z| = N\}$.

Transition functions from the dual.

A property of $g_z(y)$ is that if $|y| = |z| = N$ then

$$g_z(y) = \delta_{yz}/\pi(z)$$

Inversion formula from

$$\mathbb{E}_y\Big[g_z(Z(t))\Big] = \mathbb{E}_z\Big[g_{L(t)}(y)\Big].$$

$$s_{yz}(t) = \pi(z) \sum_{\alpha \in \mathbb{Z}_+^d, |\alpha| \leq N} h_{z\alpha}(t) \frac{\mathcal{H}(\alpha \mid y)}{\mathcal{H}(\alpha)}$$

If $Z(0)$ has a stationary distribution π, the transition expansion can be written as

$$\begin{aligned}
&\mathbb{P}(Z(0) = y, Z(t) = z) \\
&\quad = \sum_{|\alpha| \le N} \mathbb{P}(Z(t) = z)\mathbb{P}(L(t) = \alpha \mid L(0) = z) \\
&\qquad\qquad \times \mathbb{P}(Z(0) = y \mid L(t) = \alpha).
\end{aligned}$$

The joint configuration $Z(0) = y, Z(t) = z$ is obtained by looking backward in time along the typed sample lines to the initial configuration of α founder lines at time 0.
The population configuration at time 0 is the posterior distribution of types in a stationary population, given a sample configuration of $L(t) = \alpha$ in $|\alpha|$ lines.

A key step uses reversibility in the PIM Model to find $s_{yz}(t)$.

$$
\begin{aligned}
s_{yz}(t) &= s_{zy}(t)\pi(y)^{-1}\pi(z) \\
&= \pi(z) \sum_{\alpha\in\mathbb{Z}_+^d, |\alpha|\leq N} h_{y\alpha}(t) g_\alpha(z) \\
&= \sum_{\alpha\in\mathbb{Z}_+^d, |\alpha|\leq N} h_{y\alpha}(t)\pi(z-\alpha;\theta+\alpha, N-|\alpha|)
\end{aligned}
$$

Theorem. For $y, z \in \mathbb{Z}_+^d$, $|y| = |z| = N$,

$$s_{yz}(t) = \sum_{\alpha \in \mathbb{Z}_+^d, |\alpha| \leq N} h_{y\alpha}(t) \pi(z - \alpha; \theta + \alpha, N - |\alpha|),$$

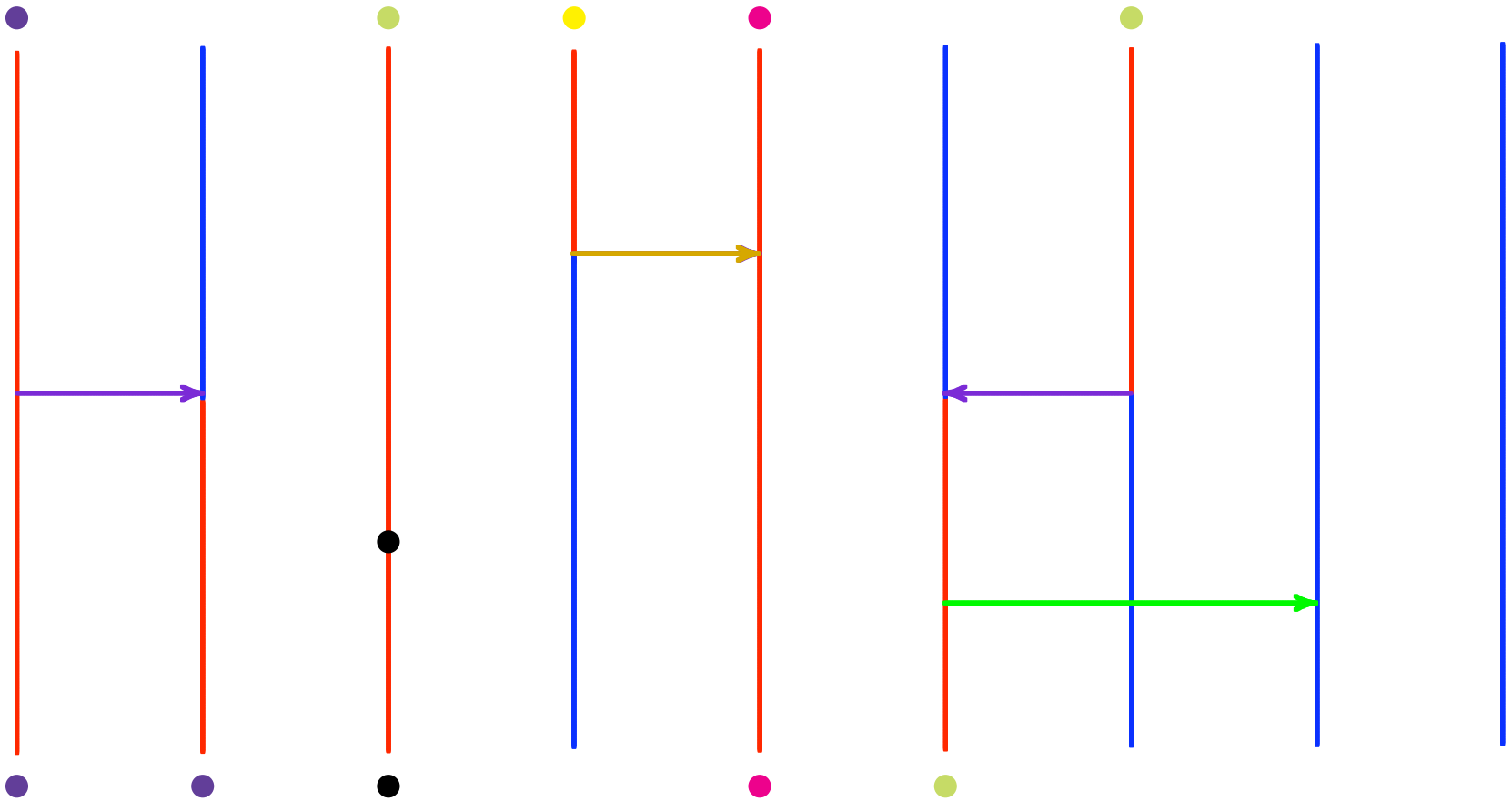
where $\{s_{yz}(t), t \geq 0\}$ are the transition functions in the Moran model with genic selection $\{Z(t), t \geq 0\}$ and $\{h_{y\alpha}(t), t \geq 0\}$ are transition functions of a d-dimensional birth and death process $\{L(t), t \geq 0\}$ with rates $Q = (q(\alpha, \beta))$.
$\pi(z; \theta, N)$ is the λ-weighted Multinomial-Dirichlet distribution.

Etheridge, A.M. and Griffiths, R.C. (2009). A Coalescent Dual Process in a Moran model with Genic Selection. Theor. Popul. Biol.

The rate matrix Q

$$
\begin{aligned}
q(\beta, \beta + e_j) &= \lambda^* \Big(1 - \frac{|\beta|}{N}\Big)\Big(1 - \frac{\lambda_j}{\lambda^*}\Big)|\beta| \frac{\beta_j + 1}{|\beta| + 1} \frac{\mathcal{H}(\beta + e_j)}{\mathcal{H}(\beta)} \\
q(\beta, \beta - e_j) &= \Big[\lambda^* \frac{\lambda_j}{\lambda^*} \frac{(\beta_j - 1)}{N} + \frac{1}{N} \mu p_j\Big] |\beta| \frac{\mathcal{H}(\beta - e_j)}{\mathcal{H}(\beta)} \\
q(\beta, \beta) &= -\Big[\frac{N - |\beta|}{N} \sum_{j=1}^{d} \beta_j \lambda_j^* + \frac{|\beta| - 1}{N} \sum_{j=1}^{d} \beta_j \lambda_j + |\beta| \mu\Big].
\end{aligned}
$$

$\mathcal{H}(\beta)$ is the sampling distribution in a sample of $|\beta|$ genes.



The Strong Dual process has red lines.
Real arrows are purple. Virtual arrows are orange.

Random drift Wright-Fisher diffusion process.

$X(t) = (X_0(t) \ldots, X_n(t))$ are gene frequencies of $n+1$ types, labelled $0, 1, \ldots, n$ at time $t \geq 0$. $X_0(t) + \cdots + X_n(t) = 1$.

A random drift process with no mutation has a generator

$$\mathcal{L} = \frac{1}{2} \sum_{i,j=0}^{n} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j}$$

Harmonic measure Problem: Find the probability-density $h_0(x, y)$ that allele type 0 is lost first and the frequency distribution of alleles $y = (y_1, \ldots, y_n)$ at the time of loss, given initial frequencies $x = (x_0, \ldots, x_n)$.

Answer:

$$h_0(x,y) = \sum_{k=n+1}^{\infty} (1-x_0)^{k-1} x_0 \sum_{\{l \in \mathbb{N}_+^n : |l| = k-1\}} \cdot \binom{k-1}{l} \left(\frac{x_1}{1-x_0}\right)^{l_1} \cdots \left(\frac{x_n}{1-x_0}\right)^{l_n} \cdot \frac{\Gamma(|l|)}{\Gamma(l_1)\cdots\Gamma(l_n)} y_1^{l_1-1} \cdots y_n^{l_n-1}$$

Ethier, S.N. and Griffiths, R.C. (1991). Harmonic measure for random genetic drift.

Probability $p_0(x)$ that allele type 0 is lost first.
Integrate over y in $h_0(x, y)$.

$$p_0(x) = \sum_{k=n+1}^{\infty} \text{geom}(k; x_0) \cdot \sum_{\{l \in \mathbb{N}_+^n : |l| = k-1\}} \text{mult}(l; k-1, (1-x_0)^{-1} x_+)$$

Note that $l_i \geq 1$ in the multinomial sum.

Littler, R. A. (1975). Loss of variability at one locus in a finite population. Math. Biosciences. 25, 151-163.
In the case $n = 2$

$$p_0(x) = x_1 x_2 \Big(\frac{1}{1-x_1} + \frac{1}{1-x_2} \Big)$$

Harmonic measure calculation, Moran model with $n+1$ types.

Theorem. In a neutral Moran model with $\lambda_j = N/2$, $j = 0, \ldots, n$ the joint probability density that type 0 is lost first, at time τ, with frequencies at τ^+ of $z = (z_j)_{j\in[n]} > 0$ given $Z(0) = y > 0$ is

$$\sum_{\{l: l_j>0, j\in[n]\}} f_{|y||l|}(\tau)\mathcal{H}_0(l \mid y)\pi_n(z-l; l, N-|l|),$$

where $f_{|y|,|l|}(\tau)$ is the density of the transition time from $|y| \to |l|$ blocks in the Kingman coalescent with death rates $\binom{k}{2}$, $k = 2, 3, \ldots$; $\mathcal{H}_0(l \mid y)$ is the distribution of types $l = (l_j)_{j\in[n]}$ visited in a series of draws without replacement from a population of $y_0, \ldots, y_n$ individuals at the instant a type 0 individual is first drawn; and $\pi_n(z-l; l, N-|l|)$ is the Multinomial Dirichlet distribution.

Theorem. In a neutral Moran model with $\lambda_j = N/2$, $j = 0, \ldots, n$ the probability that type 0 is lost first is

$$P_0 = \sum_{\{l: l_j > 0, j \in [n]\}} \mathcal{H}_0(l \mid y)\pi_n(z - l; l, N - |l|),$$

and the mean time to loss, given type 0 is lost first is

$$\sum_{\{l: l_j > 0, j \in [n]\}} \mu_{|y||l|}\mathcal{H}_0(l \mid y)\pi_n(z - l; l, N - |l|)/P_0,$$

where

$$\mu_{ab} = \sum_{k=b+1}^{a} \frac{2}{k(k-1)}.$$

In the diffusion limit as $N \to \infty$, with initial frequencies p

$$\mathcal{H}_0(l \mid y)\pi_n(z - l; l, N - |l|) \to p_0(1 - p_0)^{|l|}\mathcal{M}(l, p_+/(1 - p_0))$$