New evolutionary models for the long range dependencies of loosely linked loci

Paul A. Jenkins

Departments of Statistics
University of Warwick

Joint work with Paul Fearnhead (Lancaster), Yun Song (Berkeley)

17 June 2015
CONFERENCE on PROBABILITY AND BIOLOGICAL EVOLUTION,
Centre International de Rencontres Mathématiques (CIRM)
Marseille-Luminy.

The basic problem (Computing likelihoods)

For a given population genetics model, what is the probability of observing a sample of DNA sequences randomly drawn from a population?

Haplotype 1 = AACTAGG......CCGTGACC......ACAGCTAT
Haplotype 2 = AACTAGG......CCGTAACC......ACAGCTAT
Haplotype 3 = AACTGGG......CCGTGACC......ACAGCTAT
Haplotype 4 = AACTGGG......CCGTAACC......ACAGTTAT
Haplotype 5 = AACTAGG......CCGTGACC......ACAGTTAT

Applications

- Estimating evolutionary parameters: $L(\theta, \rho) = \mathbb{P}(D \mid \theta, \rho)$
- Ancestral inference
- Disease gene mapping

Closed-form one-locus likelihood functions

- $\boldsymbol{n} = (n_1, \ldots, n_K)$, where n_i = number of samples with allele i.
- $q(\boldsymbol{n})$, probability of an ordered sample with configuration $\boldsymbol{n}$.
- $\theta = 4Nu$, mutation parameter.

Closed-form one-locus likelihood functions

- $\boldsymbol{n} = (n_1, \ldots, n_K)$, where n_i = number of samples with allele i.
- $q(\boldsymbol{n})$, probability of an ordered sample with configuration $\boldsymbol{n}$.
- $\theta = 4Nu$, mutation parameter.

Finite alleles, parent-independent mutation (PIM) model

- Mutation transition matrix satisfies $P_{ij} = P_j$.
- Wright's sampling formula (1949):

$$q_{\mathrm{WSF}}(\boldsymbol{n}) = \frac{\prod_{i=1}^{K} \theta P_i(\theta P_i + 1)\ldots(\theta P_i + n_i - 1)}{\theta(\theta+1)\ldots(\theta+n-1)}$$

Closed-form one-locus likelihood functions

- $\boldsymbol{n} = (n_1, \ldots, n_K)$, where n_i = number of samples with allele i.
- $q(\boldsymbol{n})$, probability of an ordered sample with configuration $\boldsymbol{n}$.
- $\theta = 4Nu$, mutation parameter.

Finite alleles, parent-independent mutation (PIM) model

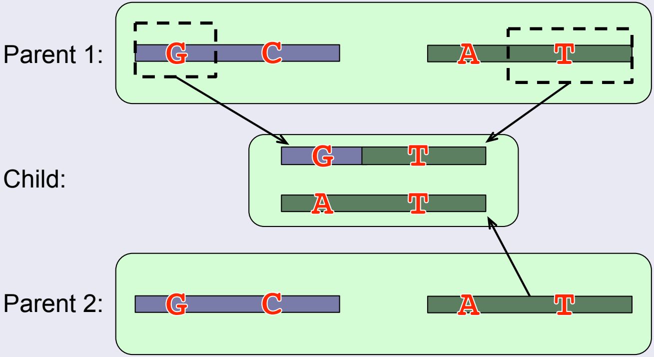
- Mutation transition matrix satisfies $P_{ij} = P_j$.
- Wright's sampling formula (1949):

$$q_{\text{WSF}}(\boldsymbol{n}) = \frac{\prod_{i=1}^{K} \theta P_i(\theta P_i + 1) \ldots (\theta P_i + n_i - 1)}{\theta(\theta + 1) \ldots (\theta + n - 1)}$$

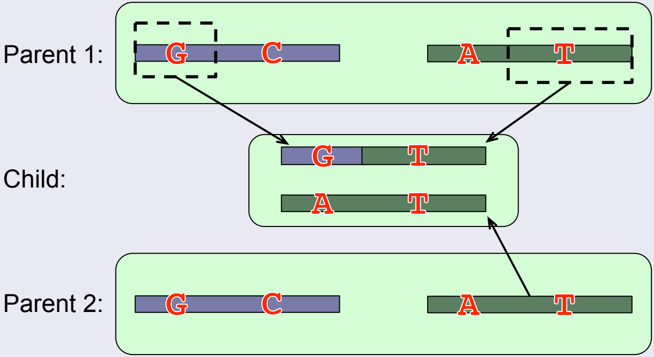
Infinite alleles model

Ewens sampling formula (1972): $$q_{\text{ESF}}(\boldsymbol{n}) = \frac{\theta^K \prod_{i=1}^{K} (n_i - 1)!}{\theta(\theta + 1) \ldots (\theta + n - 1)}$$

Crossover Recombination



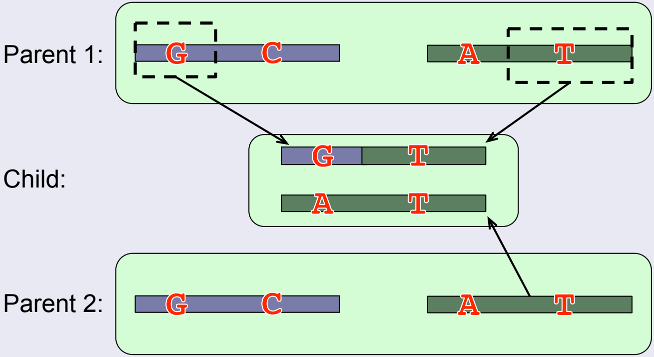
Crossover Recombination



Multi-locus models

- Ancestral recombination graph (ARG)
- Wright-Fisher diffusion with recombination

Crossover Recombination

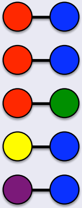


Multi-locus models with recombination

Obtaining an exact, analytic likelihood function under these models has so far remained a challenging open problem, even for just two loci.

Problem setup

A two-locus sample configuration, $\boldsymbol{c} = (c_{ij})$



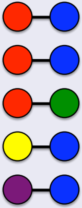
	2	1
	1	0
	1	0

Row sums: $\boldsymbol{c}_A = (c_{i\cdot}) = (3, 1, 1)$

Column sums: $\boldsymbol{c}_B = (c_{\cdot j}) = (4, 1)$

Problem setup

A two-locus sample configuration, $\boldsymbol{c} = (c_{ij})$

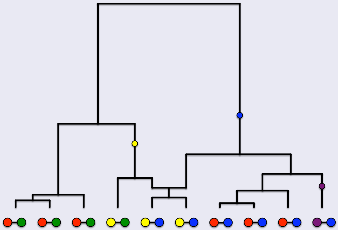


		
	2	1
	1	0
	1	0

Row sums: $\boldsymbol{c}_A = (c_{i\cdot}) = (3, 1, 1)$

Column sums: $\boldsymbol{c}_B = (c_{\cdot j}) = (4, 1)$

Goal: Compute the sampling distribution, $q(\boldsymbol{c})$.



Previous work

Key Idea: Asymptotic Series (Jenkins & Song, 2009, 2010, 2012)

Write
$$q(\boldsymbol{c};\rho) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \ldots,$$
where $q_0, q_1, \ldots$ are independent of the recombination parameter, ρ $(= 4Nr)$ (but implicitly depend on θ_A, θ_B). Now recursively solve for q_0, $q_1, \ldots$.

Previous work

Key Idea: Asymptotic Series

(Jenkins & Song, 2009, 2010, 2012)

Write

$$q(\boldsymbol{c};\rho) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \dots,$$

where $q_0, q_1, \dots$ are independent of the recombination parameter, ρ $(= 4Nr)$ (but implicitly depend on θ_A, θ_B). Now recursively solve for $q_0, q_1, \dots$.

$q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c})$ is the exact sampling distribution when the two loci are unlinked ($\rho = \infty$).

Previous work

Key Idea: Asymptotic Series

(Jenkins & Song, 2009, 2010, 2012)

Write

$$q(\boldsymbol{c};\rho) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \dots,$$

where $q_0, q_1, \dots$ are independent of the recombination parameter, ρ $(= 4Nr)$ (but implicitly depend on θ_A, θ_B). Now recursively solve for $q_0, q_1, \dots$.

$q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c})$ is the exact sampling distribution when the two loci are unlinked ($\rho = \infty$).

- Infinite alleles: $q_0(\boldsymbol{c}) = q^A_{\text{ESF}}(\boldsymbol{c}_A) q^B_{\text{ESF}}(\boldsymbol{c}_B)$
- Finite alleles, parent-independent mutation: $q_0(\boldsymbol{c}) = q^A_{\text{WSF}}(\boldsymbol{c}_A) q^B_{\text{WSF}}(\boldsymbol{c}_B)$

Previous work

Key Idea: Asymptotic Series

(Jenkins & Song, 2009, 2010, 2012)

Write

$$q(\boldsymbol{c};\rho) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \dots,$$

where $q_0, q_1, \dots$ are independent of the recombination parameter, $\rho\ (= 4Nr)$ (but implicitly depend on θ_A, θ_B). Now recursively solve for q_0, q_1, ...

$q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c})$ is the exact sampling distribution when the two loci are unlinked ($\rho = \infty$).

- Infinite alleles: $q_0(\boldsymbol{c}) = q^A_{\text{ESF}}(\boldsymbol{c}_A) q^B_{\text{ESF}}(\boldsymbol{c}_B)$
- Finite alleles, parent-independent mutation: $q_0(\boldsymbol{c}) = q^A_{\text{WSF}}(\boldsymbol{c}_A) q^B_{\text{WSF}}(\boldsymbol{c}_B)$
- Key property: $q_0(\boldsymbol{c})$ is expressible in terms of the relevant one-locus sampling distributions.

Higher order terms

(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.

Higher order terms

(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.

Higher order terms

(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.

Higher order terms

(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

Higher order terms

(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

Higher order terms

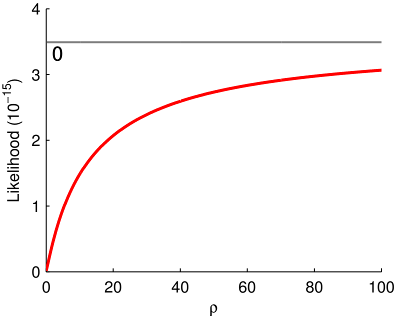
(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

Before Padé summation

Example

$\boldsymbol{c} = \binom{10\ 7}{2\ \ 1}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).

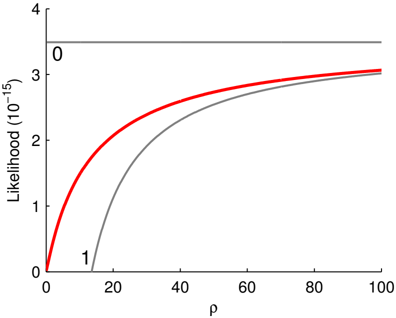


Higher order terms

(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

Before Padé summation



Example

$\boldsymbol{c} = \begin{pmatrix} 10 & 7 \\ 2 & 1 \end{pmatrix}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).

Higher order terms

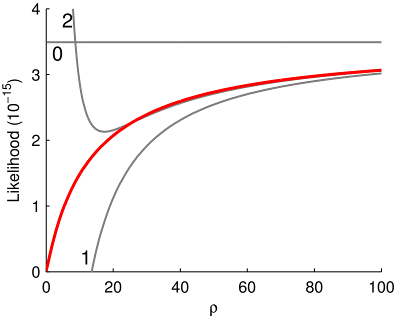
(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

Before Padé summation

Example

$\boldsymbol{c} = \begin{pmatrix} 10 & 7 \\ 2 & 1 \end{pmatrix}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Higher order terms

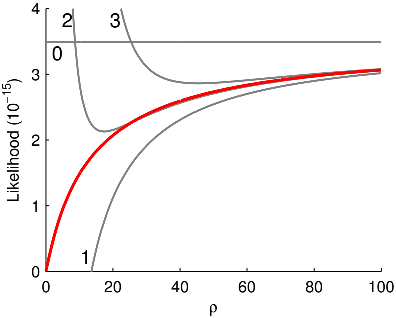
(Jenkins & Song, 2012)

- We have developed a systematic and automatable method to compute higher order terms: $q_1(\boldsymbol{c}), q_2(\boldsymbol{c}), q_3(\boldsymbol{c}), \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

Before Padé summation

Example

$\boldsymbol{c} = \binom{10\ 7}{2\ \ 1}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Higher order terms

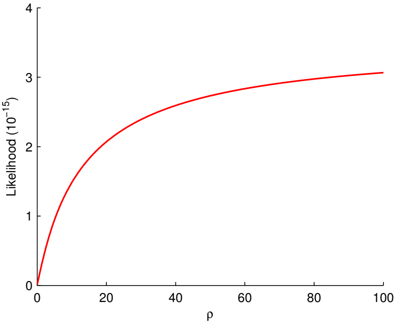
(Jenkins & Song, 2012)

- We have developed a simple, systematic and automatable method to compute higher order terms: $q_1, q_2, q_3, \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

After Padé summation

Example

$\mathbf{c} = \begin{pmatrix} 10 & 7 \\ 2 & 1 \end{pmatrix}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Higher order terms

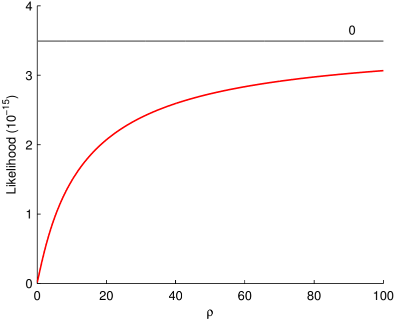
(Jenkins & Song, 2012)

- We have developed a simple, systematic and automatable method to compute higher order terms: $q_1, q_2, q_3, \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

After Padé summation

Example

$\boldsymbol{c} = \binom{10\ 7}{2\ \ 1}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Higher order terms

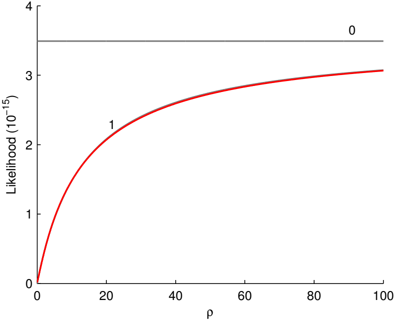
(Jenkins & Song, 2012)

- We have developed a simple, systematic and automatable method to compute higher order terms: $q_1, q_2, q_3, \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

After Padé summation

Example

$\boldsymbol{c} = \binom{10\ 7}{2\ \ 1}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Higher order terms

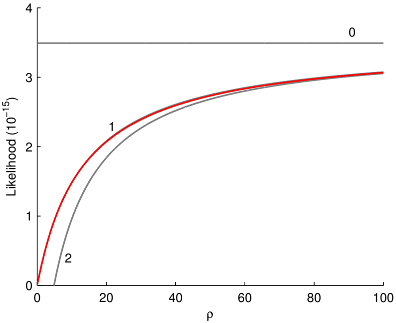
(Jenkins & Song, 2012)

- We have developed a simple, systematic and automatable method to compute higher order terms: $q_1, q_2, q_3, \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

After Padé summation

Example

$\boldsymbol{c} = \begin{pmatrix} 10 & 7 \\ 2 & 1 \end{pmatrix}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Higher order terms

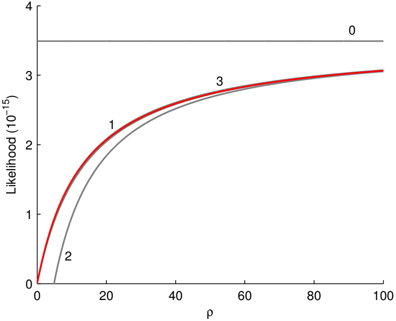
(Jenkins & Song, 2012)

- We have developed a simple, systematic and automatable method to compute higher order terms: $q_1, q_2, q_3, \ldots$.
- A technique known as Padé summation guarantees that our asymptotic series converges exactly to the truth, for any ρ.
- The method generalizes to handle missing alleles.
- The method generalizes to incorporate selection at one locus.

After Padé summation

Example

$\boldsymbol{c} = \binom{10\ 7}{2\ \ 1}$, $\theta_A = \theta_B = 0.01$ (symmetric mutation).



Intriguing observation

Reminder: Asymptotic expansion

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \dots,$$

Intriguing observation

Reminder: Asymptotic expansion

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \dots,$$

Reminder: $q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c}) = q^A(\boldsymbol{c}_A)q^B(\boldsymbol{c}_B)$ is a simple linear combination of products of one-locus sampling distributions, and universal—independent of the assumed mutation model.

Intriguing observation

Reminder: Asymptotic expansion

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \dots,$$

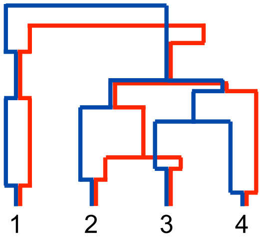
Reminder: $q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c}) = q^A(\boldsymbol{c}_A)q^B(\boldsymbol{c}_B)$ is a simple linear combination of products of one-locus sampling distributions, and universal—independent of the assumed mutation model.

Observation: The same is true of $q_1(\boldsymbol{c})$.

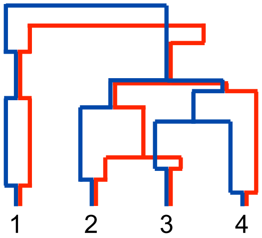
$$q_1(\boldsymbol{c}) = \binom{c}{2} q^A(\boldsymbol{c}_A)q^B(\boldsymbol{c}_B) + \sum_{i,j} \binom{c_{ij}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i)q^B(\boldsymbol{c}_B - \boldsymbol{e}_j)$$
$$- q^B(\boldsymbol{c}_B)\sum_i \binom{c_{i\cdot}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i) - q^A(\boldsymbol{c}_A)\sum_j \binom{c_{\cdot j}}{2} q^B(\boldsymbol{c}_B - \boldsymbol{e}_j).$$

[$\boldsymbol{e}_i = (0\dots,0,1,0,\dots,0)^T$, a unit vector with a 1 in the ith position.]



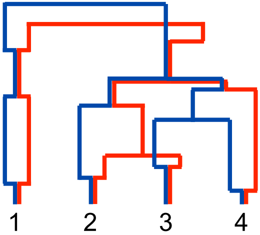
The standard coalescent with recombination

For large recombination rates, ARGs are typically very complicated, containing many recombination events.



Counterintuitive

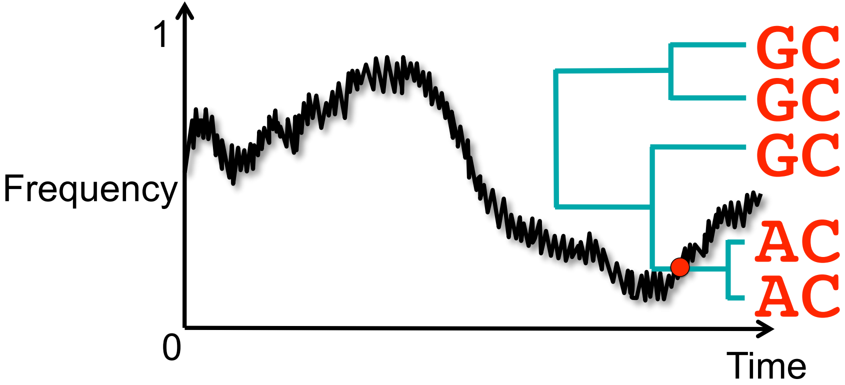
However, we in fact expect the dynamics to be easier to study for large recombination rates, since the loci under consideration would then be less dependent.



Conjecture

There exists a simpler stochastic process that describes the important dynamics of the ARG for large recombination rates, with $q_1(\boldsymbol{c})$ capturing its sampling distribution.

Duality



Conjecture

Furthermore, we should be able to make a similar statement about the Wright-Fisher diffusion, via duality.

A new diffusion model

Goal: Derive a diffusion model which is

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with
- a closed-form sampling distribution, which

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with
- a closed-form sampling distribution, which
- agrees with the “truth” [up to $O(\rho^{-2})$]: $q_0(\boldsymbol{c}) + q_1(\boldsymbol{c})/\rho$.

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with
- a closed-form sampling distribution, which
- agrees with the "truth" [up to $O(\rho^{-2})$]: $q_0(\boldsymbol{c}) + q_1(\boldsymbol{c})/\rho$.

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with
- a closed-form sampling distribution, which
- agrees with the "truth" [up to $O(\rho^{-2})$]: $q_0(\boldsymbol{c}) + q_1(\boldsymbol{c})/\rho$.

Outline of approach

1. Start with a two-locus Moran model.

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with
- a closed-form sampling distribution, which
- agrees with the "truth" [up to $O(\rho^{-2})$]: $q_0(\boldsymbol{c}) + q_1(\boldsymbol{c})/\rho$.

Outline of approach

1. Start with a two-locus Moran model.
2. Change coordinates from haplotype frequencies to marginal allele frequencies and coefficients of linkage disequilibrium (cf. Ohta & Kimura, 1969).

A new diffusion model

Goal: Derive a diffusion model which is

- simple to describe, with
- a closed-form sampling distribution, which
- agrees with the "truth" [up to $O(\rho^{-2})$]: $q_0(\boldsymbol{c}) + q_1(\boldsymbol{c})/\rho$.

Outline of approach

1. Start with a two-locus Moran model.
2. Change coordinates from haplotype frequencies to marginal allele frequencies and coefficients of linkage disequilibrium (cf. Ohta & Kimura, 1969).
3. Suppose that $\rho_\beta = 4N^\beta r$ is fixed as $N \to \infty$, where $0 < \beta < 1$—instead of the usual $\beta = 1$.

A new diffusion model

Goal: Derive a diffusion model which is

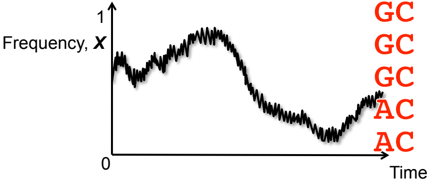
- simple to describe, with
- a closed-form sampling distribution, which
- agrees with the "truth" [up to $O(\rho^{-2})$]: $q_0(\boldsymbol{c}) + q_1(\boldsymbol{c})/\rho$.

Outline of approach

1. Start with a two-locus Moran model.
2. Change coordinates from haplotype frequencies to marginal allele frequencies and coefficients of linkage disequilibrium (cf. Ohta & Kimura, 1969).
3. Suppose that $\rho_\beta = 4N^\beta r$ is fixed as $N \to \infty$, where $0 < \beta < 1$—instead of the usual $\beta = 1$.
4. Take the diffusion limit of the fluctuations of the coordinates about the deterministic limit.

The Wright-Fisher diffusion

$$d\boldsymbol{X} = \mu(\boldsymbol{X})dt + \sigma(\boldsymbol{X})d\boldsymbol{W},$$
$$\boldsymbol{X} = (X_{ij}), \qquad i, j, \in \{A, C, G, T\}.$$

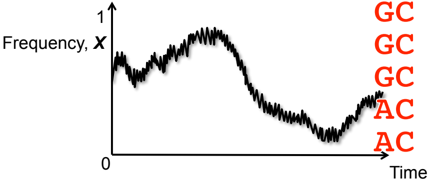


The (two-locus) Wright-Fisher diffusion

- State space: $\Delta = \left\{ \boldsymbol{x} = (x_{ij}) \in [0,1]^{K \times L} \mid \sum_{i,j} x_{ij} = 1 \right\}.$

The Wright-Fisher diffusion

$$d\boldsymbol{X} = \mu(\boldsymbol{X})dt + \sigma(\boldsymbol{X})d\boldsymbol{W},$$
$$\boldsymbol{X} = (X_{ij}), \qquad i, j, \in \{A, C, G, T\}.$$



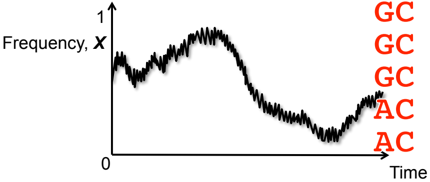
The (two-locus) Wright-Fisher diffusion

- State space: $\Delta = \left\{ \boldsymbol{x} = (x_{ij}) \in [0,1]^{K \times L} \mid \sum_{i,j} x_{ij} = 1 \right\}.$
- Drift coefficient

$$\mu_{ij}(\boldsymbol{x}) = -\frac{\rho}{2}(x_{ij} - x_{i \cdot} x_{\cdot j}) + (\text{mutation terms}; \theta_A, \theta_B)$$

The Wright-Fisher diffusion

$$dX = \mu(X)dt + \sigma(X)dW,$$
$$X = (X_{ij}), \qquad i, j, \in \{A, C, G, T\}.$$



The (two-locus) Wright-Fisher diffusion

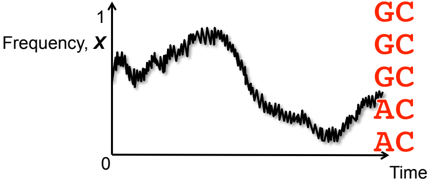
- State space: $\Delta = \left\{ \mathbf{x} = (x_{ij}) \in [0,1]^{K \times L} \mid \sum_{i,j} x_{ij} = 1 \right\}.$
- Drift coefficient

$$\mu_{ij}(\mathbf{x}) = -\frac{\rho}{2}(x_{ij} - x_{i\cdot}x_{\cdot j}) + (\text{mutation terms}; \theta_A, \theta_B)$$

- Diffusion coefficient: $\sigma^2_{ij,kl}(\mathbf{x}) = x_{ij}(\delta_{ij,kl} - x_{kl}).$

The Wright-Fisher diffusion

$$d\boldsymbol{X} = \mu(\boldsymbol{X})dt + \sigma(\boldsymbol{X})d\boldsymbol{W},$$
$$\boldsymbol{X} = (X_{ij}), \qquad i, j, \in \{A, C, G, T\}.$$



Sampling distribution

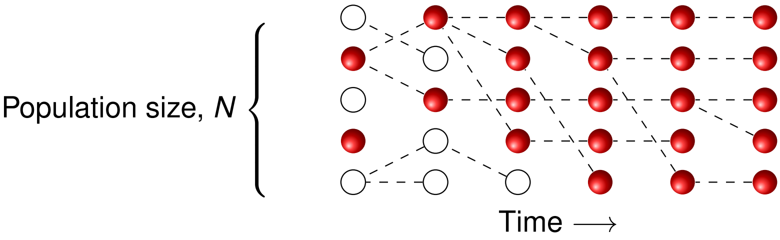
$$q(\boldsymbol{c}) = \mathbb{E}\left[\prod_{i,j} X_{ij}^{c_{ij}}\right].$$

- Using a standard result: $\mathbb{E}[\mathcal{L}f(\boldsymbol{X})] = 0$, we get a linear system of equation for the moments of $\boldsymbol{X}$.
- But this system grows exponentially in the sample size.
- So we need an approximation.

How to derive this diffusion?

Classical approach

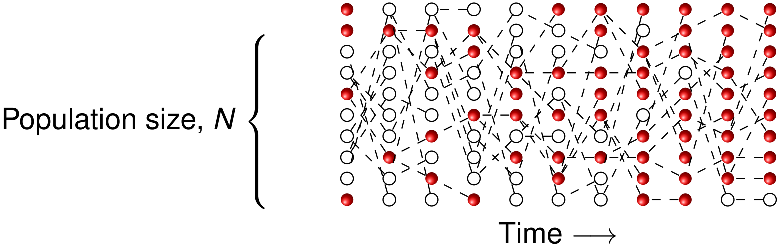
- Start from a finite population model of size N.
- Let $N \to \infty$ (possibly after a rescaling of time).
- Rates of mutation and recombination are assumed to be such that they occur at $O(1)$ in the diffusion limit.



How to derive this diffusion?

Classical approach

- Start from a finite population model of size N.
- Let $N \to \infty$ (possibly after a rescaling of time).
- Rates of mutation and recombination are assumed to be such that they occur at $O(1)$ in the diffusion limit.

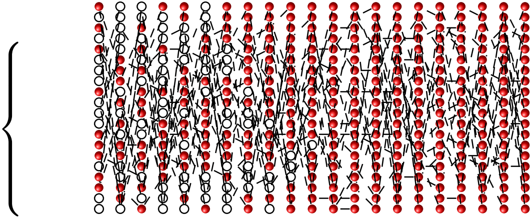


How to derive this diffusion?

Classical approach

- Start from a finite population model of size N.
- Let $N \to \infty$ (possibly after a rescaling of time).
- Rates of mutation and recombination are assumed to be such that they occur at $O(1)$ in the diffusion limit.

Population size, N

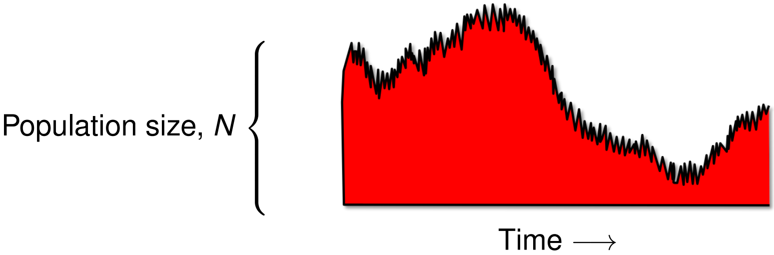


Time $\longrightarrow$

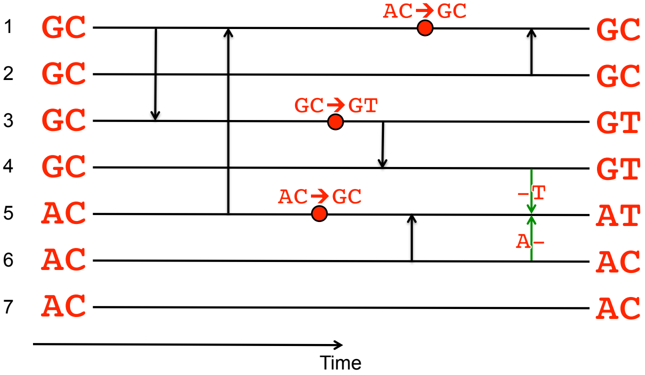
How to derive this diffusion?

Classical approach

- Start from a finite population model of size N.
- Let $N \to \infty$ (possibly after a rescaling of time).
- Rates of mutation and recombination are assumed to be such that they occur at $O(1)$ in the diffusion limit.



1. Moran model



Rates

Resampling	$N^2/2$
Mutation (locus A)	$\theta_A/2$
Mutation (locus B)	$\theta_B/2$
Recombination	$\rho/2$

2. Change coordinates (Ohta & Kimura, 1969)

Old system $\left(X_{ij}^{(N)}\right)$, $i \in \{1, 2, \ldots, K\}, j \in \{1, 2, \ldots, L\}$

2. Change coordinates (Ohta & Kimura, 1969)

Old system $\left(X_{ij}^{(N)}\right), \quad i \in \{1, 2, \ldots, K\}, j \in \{1, 2, \ldots, L\}$

New system $\left((X_{i\cdot}^{(N)}), (X_{\cdot j}^{(N)}), (D_{ij}^{(N)})\right), \quad D_{ij}^{(N)} := X_{ij}^{(N)} - X_{i\cdot}^{(N)} X_{\cdot j}^{(N)}.$

2. Change coordinates (Ohta & Kimura, 1969)

Old system $\left(X_{ij}^{(N)}\right), \qquad i \in \{1, 2, \ldots, K\}, j \in \{1, 2, \ldots, L\}$

New system $\left((X_{i\cdot}^{(N)}), (X_{\cdot j}^{(N)}), (D_{ij}^{(N)})\right), \quad D_{ij}^{(N)} := X_{ij}^{(N)} - X_{i\cdot}^{(N)} X_{\cdot j}^{(N)}.$

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_A}{2}\sum_{k=1}^{K} P_{ki}^{A} X_{k\cdot}^{(N)} - \frac{\theta_A}{2} X_{i\cdot}^{(N)}\right] dt + o(dt),$$

$$\mathbb{E}[\Delta X_{\cdot j}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_B}{2}\sum_{l=1}^{L} P_{lj}^{B} X_{\cdot l}^{(N)} - \frac{\theta_B}{2} X_{\cdot j}^{(N)}\right] dt + o(dt),$$

$$\mathbb{E}[\Delta D_{ij}^{(N)} \mid \boldsymbol{X}] = \left[-\frac{\rho}{2} D_{ij}^{(N)} - D_{ij}^{(N)} + \frac{\theta_A}{2}\sum_{k=1}^{K} P_{ki}^{A} D_{kj}^{(N)} - \frac{\theta_A}{2} D_{ij}^{(N)} \right.$$
$$\left. + \frac{\theta_B}{2}\sum_{l=1}^{L} P_{lj}^{B} D_{il}^{(N)} - \frac{\theta_B}{2} D_{ij}^{(N)} + O(N^{-1})\right] dt + o(dt)$$

3. Rescale recombination, ρ

Suppose $\rho_\beta = \rho N^{\beta-1} = 4N^\beta r$ is fixed as $N \to \infty$, where $0 < \beta < 1$.
Rescale time to capture this fast behaviour: $t_{\text{new}} = N^{1-\beta} t_{\text{old}}$.

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_A}{2}\sum_{k=1}^{K} P_{ki}^{A} X_{k\cdot}^{(N)} - \frac{\theta_A}{2} X_{i\cdot}^{(N)}\right] dt + o(dt),$$

$$\mathbb{E}[\Delta X_{\cdot j}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_B}{2}\sum_{l=1}^{L} P_{lj}^{B} X_{\cdot l}^{(N)} - \frac{\theta_B}{2} X_{\cdot j}^{(N)}\right] dt + o(dt),$$

$$\mathbb{E}[\Delta D_{ij}^{(N)} \mid \boldsymbol{X}] = \left[-\frac{\rho}{2} D_{ij}^{(N)} - D_{ij}^{(N)} + \frac{\theta_A}{2}\sum_{k=1}^{K} P_{ki}^{A} D_{kj}^{(N)} - \frac{\theta_A}{2} D_{ij}^{(N)} + \frac{\theta_B}{2}\sum_{l=1}^{L} P_{lj}^{B} D_{il}^{(N)} - \frac{\theta_B}{2} D_{ij}^{(N)} + O(N^{-1})\right] dt + o(dt)$$

3. Rescale recombination, ρ

Suppose $\rho_\beta = \rho N^{\beta-1} = 4N^\beta r$ is fixed as $N \to \infty$, where $0 < \beta < 1$.
Rescale time to capture this fast behaviour: $t_{\text{new}} = N^{1-\beta} t_{\text{old}}$.

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_A}{2}\sum_{k=1}^{K} P_{ki}^A X_{k\cdot}^{(N)} - \frac{\theta_A}{2} X_{i\cdot}^{(N)}\right] dt + o(dt),$$

$$\mathbb{E}[\Delta X_{\cdot j}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_B}{2}\sum_{l=1}^{L} P_{lj}^B X_{\cdot l}^{(N)} - \frac{\theta_B}{2} X_{\cdot j}^{(N)}\right] dt + o(dt),$$

$$\mathbb{E}[\Delta D_{ij}^{(N)} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta N^{1-\beta}}{2} D_{ij}^{(N)} - D_{ij}^{(N)} + \frac{\theta_A}{2}\sum_{k=1}^{K} P_{ki}^A D_{kj}^{(N)} - \frac{\theta_A}{2} D_{ij}^{(N)} \right.$$
$$\left. + \frac{\theta_B}{2}\sum_{l=1}^{L} P_{lj}^B D_{il}^{(N)} - \frac{\theta_B}{2} D_{ij}^{(N)} + O(N^{-1})\right] dt + o(dt)$$

3. Rescale recombination, ρ

Suppose $\rho_\beta = \rho N^{\beta-1} = 4N^\beta r$ is fixed as $N \to \infty$, where $0 < \beta < 1$.
Rescale time to capture this fast behaviour: $t_{\text{new}} = N^{1-\beta} t_{\text{old}}$.

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_A}{2} \sum_{k=1}^{K} P_{ki}^A X_{k\cdot}^{(N)} - \frac{\theta_A}{2} X_{i\cdot}^{(N)} \right] \frac{dt}{N^{1-\beta}} + o(dt),$$

$$\mathbb{E}[\Delta X_{\cdot j}^{(N)} \mid \boldsymbol{X}] = \left[\frac{\theta_B}{2} \sum_{l=1}^{L} P_{lj}^B X_{\cdot l}^{(N)} - \frac{\theta_B}{2} X_{\cdot j}^{(N)} \right] \frac{dt}{N^{1-\beta}} + o(dt),$$

$$\mathbb{E}[\Delta D_{ij}^{(N)} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta N^{1-\beta}}{2} D_{ij}^{(N)} - D_{ij}^{(N)} + \frac{\theta_A}{2} \sum_{k=1}^{K} P_{ki}^A D_{kj}^{(N)} - \frac{\theta_A}{2} D_{ij}^{(N)} \right.$$
$$\left. + \frac{\theta_B}{2} \sum_{l=1}^{L} P_{lj}^B D_{il}^{(N)} - \frac{\theta_B}{2} D_{ij}^{(N)} + O(N^{-1}) \right] \frac{dt}{N^{1-\beta}} + o(dt)$$

4. Seek a diffusion limit

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot}^{(N)} \mid \boldsymbol{X}] = O\left(\frac{1}{N^{1-\beta}}\right) dt + o(dt),$$

$$\mathbb{E}[\Delta X_{\cdot j}^{(N)} \mid \boldsymbol{X}] = O\left(\frac{1}{N^{1-\beta}}\right) dt + o(dt),$$

$$\mathbb{E}[\Delta D_{ij}^{(N)} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta}{2} D_{ij}^{(N)} + O\left(\frac{1}{N^{1-\beta}}\right)\right] dt + o(dt)$$

4. Seek a diffusion limit

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta X_{\cdot j} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta D_{ij} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta}{2} D_{ij}\right] dt + o(dt)$$

after $N \to \infty$.

4. Seek a diffusion limit

Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta X_{\cdot j} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta D_{ij} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta}{2} D_{ij}\right] dt + o(dt)$$

after $N \to \infty$.

- The description is completed by finding the limiting covariance matrix.

4. Seek a diffusion limit

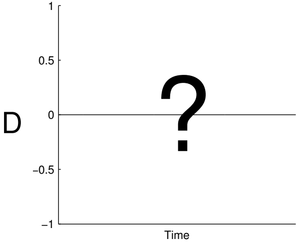
Diffusion limit

$$\mathbb{E}[\Delta X_{i\cdot} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta X_{\cdot j} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta D_{ij} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta}{2} D_{ij}\right] dt + o(dt)$$

after $N \to \infty$.

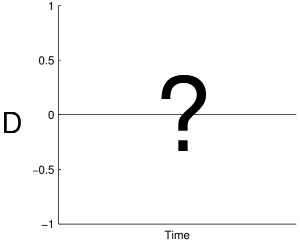
- The description is completed by finding the limiting covariance matrix.
- But—on this timescale it is **0**!

Diffusion limits

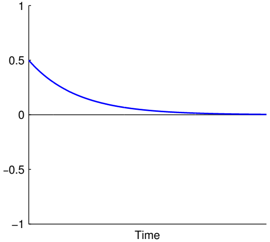


Wright-Fisher
diffusion

Diffusion limits

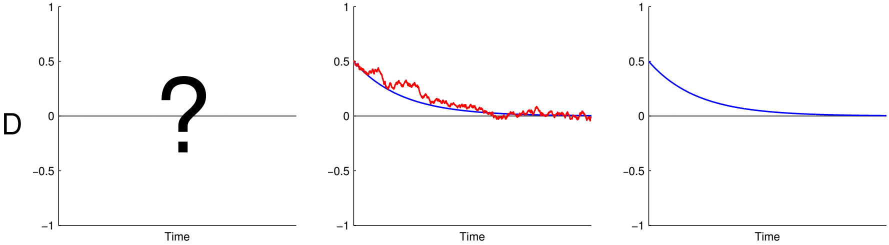


Wright-Fisher diffusion



∞-population

Diffusion limits



Wright-Fisher diffusion

Intermediate limit?

∞-population

Summary so far

If

$$\mathbf{M}^{(N)} = \Big((X_{i\cdot}^{(N)}), (X_{\cdot j}^{(N)}), (D_{ij}^{(N)} = X_{ij} - X_{i\cdot}^{(N)} X_{\cdot j}^{(N)}) \Big)$$

then

$$\mathbf{M}^{(N)} \xrightarrow{d} \mathbf{M} := \Big\{ ((X_{i\cdot}(0)), (X_{\cdot j}(0)), (D_{ij}(0) e^{-\rho_\beta t/2})' : t \geq 0 \Big\},$$

as $N \to \infty$.

- This is a law-of-large-numbers result. (Baake & Herms, 2008)

Summary so far

If

$$\boldsymbol{M}^{(N)} = \left((X_{i\cdot}^{(N)}), (X_{\cdot j}^{(N)}), (D_{ij}^{(N)} = X_{ij} - X_{i\cdot}^{(N)} X_{\cdot j}^{(N)})\right)$$

then

$$\boldsymbol{M}^{(N)} \xrightarrow{d} \boldsymbol{M} := \left\{((X_{i\cdot}(0)), (X_{\cdot j}(0)), (D_{ij}(0)e^{-\rho_\beta t/2})' : t \geq 0\right\},$$

as $N \to \infty$.

- This is a law-of-large-numbers result. (Baake & Herms, 2008)
- We really want a central limit theorem.

Summary so far

If

$$\boldsymbol{M}^{(N)} = \Big((X_{i\cdot}^{(N)}), (X_{\cdot j}^{(N)}), (D_{ij}^{(N)} = X_{ij} - X_{i\cdot}^{(N)} X_{\cdot j}^{(N)}) \Big)$$

then

$$\boldsymbol{M}^{(N)} \xrightarrow{d} \boldsymbol{M} := \Big\{ ((X_{i\cdot}(0)), (X_{\cdot j}(0)), (D_{ij}(0) e^{-\rho_\beta t/2})' : t \geq 0 \Big\},$$

as $N \to \infty$.

- This is a law-of-large-numbers result. (Baake & Herms, 2008)
- We really want a central limit theorem.
- So we should be asking: what is the diffusion limit of

$$\boldsymbol{U}^{(N)}(t) := N^{(1-\beta)/2} [\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]?$$

CLTs for density-dependent population processes

Theorem [Ethier & Kurtz, 1986, Ch. 11; Kang *et al.*, 2014]

Suppose that $\boldsymbol{U}^{(N)}(0) \to \boldsymbol{U}(0)$ as $N \to \infty$, and $\boldsymbol{M}(t)$ the solution to

$$\frac{\mathrm{d}\boldsymbol{M}(t)}{\mathrm{d}t} = \boldsymbol{w}(\boldsymbol{M}(t))$$

exists, for some $\boldsymbol{w}$.

CLTs for density-dependent population processes

Theorem [Ethier & Kurtz, 1986, Ch. 11; Kang *et al.*, 2014]

Suppose that $\boldsymbol{U}^{(N)}(0) \to \boldsymbol{U}(0)$ as $N \to \infty$, and $\boldsymbol{M}(t)$ the solution to

$$\frac{\mathrm{d}\boldsymbol{M}(t)}{\mathrm{d}t} = \boldsymbol{w}(\boldsymbol{M}(t))$$

exists, for some $\boldsymbol{w}$. Then [under some regularity conditions]

CLTs for density-dependent population processes

Theorem [Ethier & Kurtz, 1986, Ch. 11; Kang *et al.*, 2014]

Suppose that $\boldsymbol{U}^{(N)}(0) \to \boldsymbol{U}(0)$ as $N \to \infty$, and $\boldsymbol{M}(t)$ the solution to

$$\frac{\mathrm{d}\boldsymbol{M}(t)}{\mathrm{d}t} = \boldsymbol{w}(\boldsymbol{M}(t))$$

exists, for some $\boldsymbol{w}$. Then [under some regularity conditions]

$$\sup_{s \le t} |\boldsymbol{M}^{(N)}(s) - \boldsymbol{M}(s)| \xrightarrow{d} 0,$$

CLTs for density-dependent population processes

Theorem [Ethier & Kurtz, 1986, Ch. 11; Kang *et al.*, 2014]

Suppose that $\boldsymbol{U}^{(N)}(0) \to \boldsymbol{U}(0)$ as $N \to \infty$, and $\boldsymbol{M}(t)$ the solution to

$$\frac{\mathrm{d}\boldsymbol{M}(t)}{\mathrm{d}t} = \boldsymbol{w}(\boldsymbol{M}(t))$$

exists, for some $\boldsymbol{w}$. Then [under some regularity conditions]

$$\sup_{s \le t} |\boldsymbol{M}^{(N)}(s) - \boldsymbol{M}(s)| \xrightarrow{d} 0,$$

and $\boldsymbol{U}^{(N)} \xrightarrow{d} \boldsymbol{U}$, where

$$\boldsymbol{U}(t) = \boldsymbol{U}(0) + \int_0^t [\nabla \boldsymbol{w}(\boldsymbol{M}(s))]\boldsymbol{U}(s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}(s))\mathrm{d}\boldsymbol{W}(s),$$

and $\boldsymbol{\sigma}$ is such that

$$N^{1-\beta}[\boldsymbol{M}^{(N)}]_t - \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}^{(N)}(s))\boldsymbol{\sigma}(\boldsymbol{M}^{(N)}(s))' ds \xrightarrow{d} \boldsymbol{0}.$$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Goals

1. Identify $\boldsymbol{w}$, which supplies the drift part of $\boldsymbol{U}$.

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Goals

1. Identify $\boldsymbol{w}$, which supplies the drift part of $\boldsymbol{U}$.
2. Identify $\boldsymbol{\sigma}$, which supplies the diffusion part of $\boldsymbol{U}$.

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Goals

1. Identify $\boldsymbol{w}$, which supplies the drift part of $\boldsymbol{U}$.
2. Identify $\boldsymbol{\sigma}$, which supplies the diffusion part of $\boldsymbol{U}$.
3. [Check regularity requirements.]

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Goals

1. Identify $\boldsymbol{w}$, which supplies the drift part of $\boldsymbol{U}$.
2. Identify $\boldsymbol{\sigma}$, which supplies the diffusion part of $\boldsymbol{U}$.
3. [Check regularity requirements.]

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Goals

1. Identify $\boldsymbol{w}$, which supplies the drift part of $\boldsymbol{U}$.
2. Identify $\boldsymbol{\sigma}$, which supplies the diffusion part of $\boldsymbol{U}$.
3. [Check regularity requirements.]

Sketch proof.

Recall:

$$\mathbb{E}[\Delta X_{i\cdot} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta X_{\cdot j} \mid \boldsymbol{X}] = o(dt),$$
$$\mathbb{E}[\Delta D_{ij} \mid \boldsymbol{X}] = \left[-\frac{\rho_\beta}{2} D_{ij}\right] dt + o(dt)$$

So: Drift of $\boldsymbol{M}$:

$$\boldsymbol{w}(\boldsymbol{M}) = \left(\boldsymbol{0}, \boldsymbol{0}, -\frac{\rho_\beta}{2}\boldsymbol{D}\right)'$$

Drift of $\boldsymbol{M}^{(N)}$:

$$\boldsymbol{w}^{(N)}(\boldsymbol{M}) = \left(\boldsymbol{0}, \boldsymbol{0}, -\frac{\rho_\beta}{2}\boldsymbol{D}\right)' + O(N^{\beta-1})$$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider: $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2} \Big[\qquad \Big],$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider: $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2} \Big[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] \qquad \Big],$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider:
$$\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}\Big[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] + \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]ds\Big],$$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider:
$$\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}\Big[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] + \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]ds + \boldsymbol{R}^{(N)}(t)\Big],$$

where $\boldsymbol{R}^{(N)}(t) := \boldsymbol{M}^{(N)}(t) - \boldsymbol{M}^{(N)}(0) - \int_0^t \boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s))ds$.

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider:
$$\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}\Big[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] + \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]ds + \boldsymbol{R}^{(N)}(t)\Big],$$

where $\boldsymbol{R}^{(N)}(t) := \boldsymbol{M}^{(N)}(t) - \boldsymbol{M}^{(N)}(0) - \int_0^t \boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s))ds$.

1st term

We assumed $\boldsymbol{U}^{(N)}(0) \to \boldsymbol{U}(0)$ as $N \to \infty$.

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider: $$\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}\Big[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] + \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]ds + \boldsymbol{R}^{(N)}(t)\Big],$$

2nd term

$$\begin{aligned} N^{(1-\beta)/2} \int_0^t [\boldsymbol{w}_3^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}_3(\boldsymbol{M}(s))]ds &= N^{(1-\beta)/2} \int_0^t \Big[-\frac{\rho_\beta}{2}[\boldsymbol{D}^{(N)}(s) - \boldsymbol{D}(s)] + O(N^{\beta-1})\Big] ds \\ &= \int_0^t \Big[-\frac{\rho_\beta}{2}\boldsymbol{U}_3^{(N)}(s) + O(N^{(\beta-1)/2})\Big] ds \\ &\xrightarrow{d} -\frac{\rho_\beta}{2}\int_0^t \boldsymbol{U}_3(s)ds, \qquad N \to \infty. \end{aligned}$$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Sketch proof (*cont.*).

Consider: $$\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}\left[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] + \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]ds + \boldsymbol{R}^{(N)}(t)\right],$$

where $$\boldsymbol{R}^{(N)}(t) := \boldsymbol{M}^{(N)}(t) - \boldsymbol{M}^{(N)}(0) - \int_0^t \boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s))ds.$$

3rd term

- “The difference between the evolution of the Moran process and its expectation.” Key observation: $\boldsymbol{R}^{(N)}(t)$ is a martingale.
- Appeal to the martingale CLT to characterise its limit.
- In other words: we know $\boldsymbol{\sigma}(\boldsymbol{M}(t))$. □

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Putting all this together:

$$\boldsymbol{U}^{(N)}(t) \to \left[\boldsymbol{U}(0) - \frac{\rho_\beta}{2}\int_0^t (\mathbf{0},\mathbf{0},\mathbf{1})' \circ \boldsymbol{U}(s)ds + \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}(s))d\boldsymbol{W}(s)\right].$$

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

Putting all this together:

$$\boldsymbol{U}^{(N)}(t) \to \left[\boldsymbol{U}(0) - \frac{\rho_\beta}{2}\int_0^t (\mathbf{0},\mathbf{0},\mathbf{1})' \circ \boldsymbol{U}(s)ds + \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}(s))d\boldsymbol{W}(s)\right].$$

Apart from a (complicated, time-evolving) covariance term, $D_{ij}(t)$ follows an Ornstein-Uhlenbeck process!

Main aim

Find the diffusion limit of $\boldsymbol{U}^{(N)}(t) = N^{(1-\beta)/2}[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)]$.

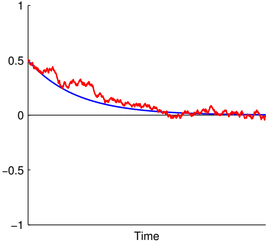
Putting all this together:

$$\boldsymbol{U}^{(N)}(t) \rightarrow \left[\boldsymbol{U}(0) - \frac{\rho_\beta}{2}\int_0^t (\mathbf{0},\mathbf{0},\mathbf{1})' \circ \boldsymbol{U}(s)ds + \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}(s))d\boldsymbol{W}(s)\right].$$

Apart from a (complicated, time-evolving) covariance term, $D_{ij}(t)$ follows an Ornstein-Uhlenbeck process!

Retracing our steps…

$$\boldsymbol{D}^{(N)}(t) \approx \boldsymbol{D}(0)e^{-\rho_\beta t/2} + N^{(\beta-1)/2}\boldsymbol{U}_{\boldsymbol{D}}(t).$$



Stationary distribution

Tracing our steps backwards, we can derive an approximate stationary distribution:

$$\boldsymbol{D} \sim \text{Normal}\left(\mathbf{0}, \frac{1}{\rho}[X_{i\cdot}(0)X_{\cdot j}(0)(\delta_{ik} - X_{k\cdot}(0))(\delta_{jl} - X_{\cdot l}(0))]_{ij,kl}\right).$$

Stationary distribution

Tracing our steps backwards, we can derive an approximate stationary distribution:

$$\boldsymbol{D} \sim \text{Normal}\left(\mathbf{0}, \frac{1}{\rho}[X_{i\cdot}(0)X_{\cdot j}(0)(\delta_{ik} - X_{k\cdot}(0))(\delta_{jl} - X_{\cdot l}(0))]_{ij,kl}\right).$$

Sampling distribution

Tracing our steps further, we can obtain a sampling distribution:

$$q_{\text{Gaussian}}(\boldsymbol{c}) = \mathbb{E}\left[\prod_{i,j} X_{ij}^{c_{ij}}\right] = \mathbb{E}\left[\prod_{i,j}(D_{ij} + X_{i\cdot}X_{\cdot j})^{c_{ij}}\right] = \ldots.$$

$$= q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \ldots.$$

Stationary distribution

Tracing our steps backwards, we can derive an approximate stationary distribution:

$$\boldsymbol{D} \sim \text{Normal}\left(\mathbf{0}, \frac{1}{\rho}[X_{i\cdot}(0)X_{\cdot j}(0)(\delta_{ik} - X_{k\cdot}(0))(\delta_{jl} - X_{\cdot l}(0))]_{ij,kl}\right).$$

Sampling distribution

Tracing our steps further, we can obtain a sampling distribution:

$$q_{\text{Gaussian}}(\boldsymbol{c}) = \mathbb{E}\left[\prod_{i,j} X_{ij}^{c_{ij}}\right] = \mathbb{E}\left[\prod_{i,j}(D_{ij} + X_{i\cdot}X_{\cdot j})^{c_{ij}}\right] = \ldots.$$

$$= q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \ldots.$$

Accuracy

"Truth": $$q(\boldsymbol{c}) \approx q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \ldots + \frac{q_\lambda(\boldsymbol{x})}{\rho^\lambda},$$

Gaussian model: $$q^{(G)}(\boldsymbol{c}) \approx q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2^{(G)}(\boldsymbol{c})}{\rho^2} + \ldots + \frac{q_\lambda^{(G)}(\boldsymbol{x})}{\rho^\lambda}.$$

		$\rho = 100$			$\rho = 200$		
λ	Type of sum	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$
0	True	0.50	0.72	1.00	0.54	0.95	1.00
	Gaussian	0.50	0.72	1.00	0.54	0.95	1.00
1	True	0.74	0.95	1.00	0.90	0.99	1.00
	Gaussian	0.74	0.95	1.00	0.90	0.99	1.00
2	True	0.95	1.00	1.00	1.00	1.00	1.00
	Gaussian	0.64	0.99	1.00	0.85	1.00	1.00
4	True	1.00	1.00	1.00	1.00	1.00	1.00
	Gaussian	0.64	0.99	1.00	0.83	1.00	1.00
6	True	1.00	1.00	1.00	1.00	1.00	1.00
	Gaussian	0.64	0.99	1.00	0.83	1.00	1.00

Accuracy

$$\text{“Truth”:} \quad q(\boldsymbol{c}) \approx q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{c})}{\rho^2} + \ldots + \frac{q_\lambda(\boldsymbol{x})}{\rho^\lambda},$$

$$\text{Gaussian model:} \quad q^{(G)}(\boldsymbol{c}) \approx q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + \frac{q_2^{(G)}(\boldsymbol{c})}{\rho^2} + \ldots + \frac{q_\lambda^{(G)}(\boldsymbol{x})}{\rho^\lambda}.$$

		$\rho = 25$			$\rho = 50$		
λ	Type of sum	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$
0	True	0.39	0.58	1.00	0.49	0.63	1.00
	Gaussian	0.39	0.58	1.00	0.49	0.63	1.00
1	True	0.51	0.75	0.96	0.59	0.84	0.99
	Gaussian	0.51	0.75	0.96	0.59	0.84	0.99
2	True	0.59	0.91	0.97	0.77	0.98	1.00
	Gaussian	0.50	0.73	0.97	0.50	0.86	1.00
4	True	0.83	0.99	1.00	0.95	1.00	1.00
	Gaussian	0.51	0.72	1.00	0.50	0.80	1.00
6	True	0.89	0.99	1.00	0.99	1.00	1.00
	Gaussian	0.49	0.71	0.99	0.50	0.79	1.00

Remarks

- No dependence on β in these expressions.

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.
- This strong recombination result complements analogous results for strong mutation and strong selection
  - (Feder *et al.*, 2014; Feller, 1951; Norman, 1972, 1975; Kaplan *et al.*, 1988; Nagylaki, 1986, 1990; Wakeley & Sargsyan, 2009).

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.
- This strong recombination result complements analogous results for strong mutation and strong selection
  - (Feder *et al.*, 2014; Feller, 1951; Norman, 1972, 1975; Kaplan *et al.*, 1988; Nagylaki, 1986, 1990; Wakeley & Sargsyan, 2009).

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.
- This strong recombination result complements analogous results for strong mutation and strong selection
  - (Feder *et al.*, 2014; Feller, 1951; Norman, 1972, 1975; Kaplan *et al.*, 1988; Nagylaki, 1986, 1990; Wakeley & Sargsyan, 2009).

Wright-Fisher model

- One could obtain the same diffusion limit starting from a Wright-Fisher model.

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.
- This strong recombination result complements analogous results for strong mutation and strong selection
  - (Feder *et al.*, 2014; Feller, 1951; Norman, 1972, 1975; Kaplan *et al.*, 1988; Nagylaki, 1986, 1990; Wakeley & Sargsyan, 2009).

Wright-Fisher model

- One could obtain the same diffusion limit starting from a Wright-Fisher model.
- CLTs for the Wright-Fisher model have been studied extensively by Norman (1972, 1975) and Nagylaki (1986, 1990).

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.
- This strong recombination result complements analogous results for strong mutation and strong selection
  - (Feder *et al.*, 2014; Feller, 1951; Norman, 1972, 1975; Kaplan *et al.*, 1988; Nagylaki, 1986, 1990; Wakeley & Sargsyan, 2009).

Wright-Fisher model

- One could obtain the same diffusion limit starting from a Wright-Fisher model.
- CLTs for the Wright-Fisher model have been studied extensively by Norman (1972, 1975) and Nagylaki (1986, 1990).
- Additional complication: the Wright-Fisher model in continuous time is non-Markovian.

Remarks

- No dependence on β in these expressions.
- Reduced a difficult likelihood computation to the moments of a Normal distribution.
- This strong recombination result complements analogous results for strong mutation and strong selection
  - (Feder *et al.*, 2014; Feller, 1951; Norman, 1972, 1975; Kaplan *et al.*, 1988; Nagylaki, 1986, 1990; Wakeley & Sargsyan, 2009).

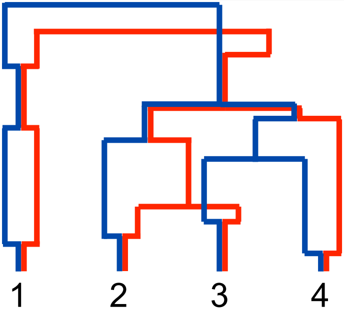
Wright-Fisher model

- One could obtain the same diffusion limit starting from a Wright-Fisher model.
- CLTs for the Wright-Fisher model have been studied extensively by Norman (1972, 1975) and Nagylaki (1986, 1990).
- Additional complication: the Wright-Fisher model in continuous time is non-Markovian.
- Q: Are there simple, general CLTs for non-Markovian density-dependent population processes?

A new coalescent model

Question

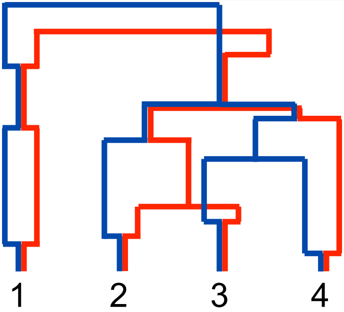
Can we give a similar treatment to the ancestral recombination graph?



A new coalescent model

Question

Can we give a similar treatment to the ancestral recombination graph? Yes—via a coupling argument.



A new coalescent model

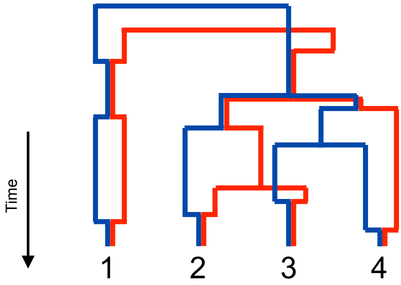
Question

Can we give a similar treatment to the ancestral recombination graph? Yes—via a coupling argument.

Toy example: sample size $c = 4$.

Blue: Lineages ancestral to the sample at locus A.

Red: Lineages ancestral to the sample at locus B.



A new coalescent model

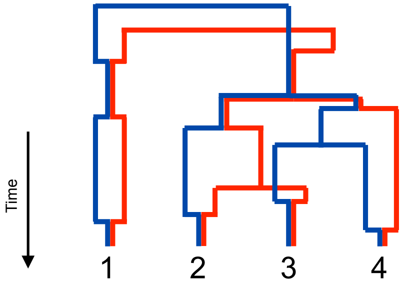
Question

Can we give a similar treatment to the ancestral recombination graph? Yes—via a coupling argument.

Toy example: sample size $c = 4$.

Blue: Lineages ancestral to the sample at locus A.

Red: Lineages ancestral to the sample at locus B.



Reminder: $q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c}) = q^A(\boldsymbol{c}_A)q^B(\boldsymbol{c}_B)$ corresponds to unlinked loci ($\rho = \infty$).

A new coalescent model

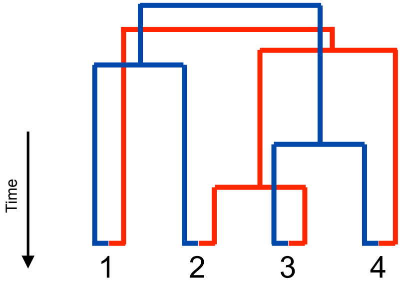
Question

Can we give a similar treatment to the ancestral recombination graph? Yes—via a coupling argument.

Toy example: sample size $c = 4$.

Blue: Lineages ancestral to the sample at locus A.

Red: Lineages ancestral to the sample at locus B.



Reminder: $q_0(\boldsymbol{c})$

$q_0(\boldsymbol{c}) = q^A(\boldsymbol{c}_A)q^B(\boldsymbol{c}_B)$ corresponds to unlinked loci ($\rho = \infty$).

A new coalescent model

What about $q_1(\boldsymbol{c})$?

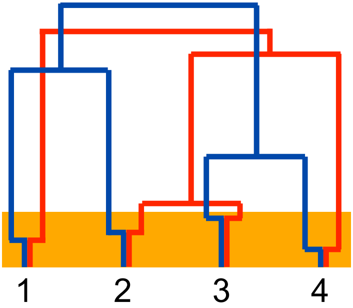
Consider what happens if we start to reduce ρ down from ∞.

A new coalescent model

What about $q_1(\boldsymbol{c})$?

Consider what happens if we start to reduce ρ down from ∞.

There is a short delay going backwards before lineages all recombine apart.

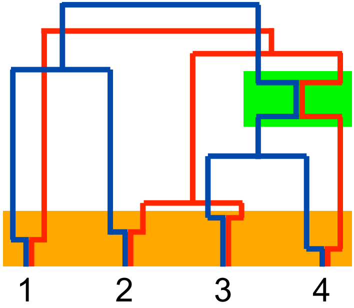


A new coalescent model

What about $q_1(\boldsymbol{c})$?

Consider what happens if we start to reduce ρ down from ∞.

There is a short delay going backwards before lineages all recombine apart. Some lineages may recoalesce further back in time.



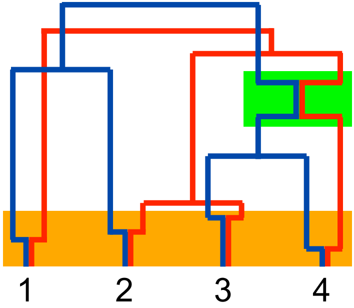
A new coalescent model

What about $q_1(\boldsymbol{c})$?

Consider what happens if we start to reduce ρ down from ∞.

There is a short delay going backwards before lineages all recombine apart. Some lineages may recoalesce further back in time.

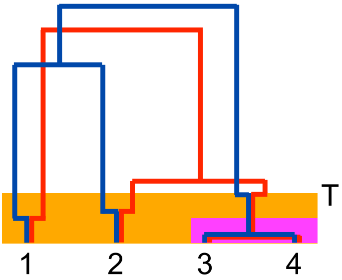
$q_1(\boldsymbol{c})$ represents the effects of any single nontrivial event in the ARG that could distinguish its sampling distribution from that of two independent coalescent trees.



A new coalescent model

Possible "nontrivial events"

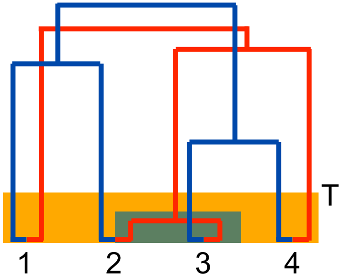
1. A coalescence prior to the first time all lineages have recombined (T).



A new coalescent model

Possible "nontrivial events"

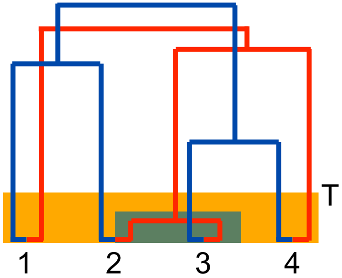
1. A coalescence prior to the first time all lineages have recombined (T).
2. A coalescence that would have happened had the marginal trees been coalescing independently, but could not have happened in our ARG before time T. (Call these "prohibited coalescences".)



A new coalescent model

Possible "nontrivial events"

1. A coalescence prior to the first time all lineages have recombined (T).
2. A coalescence that would have happened had the marginal trees been coalescing independently, but could not have happened in our ARG before time T. (Call these "prohibited coalescences".)

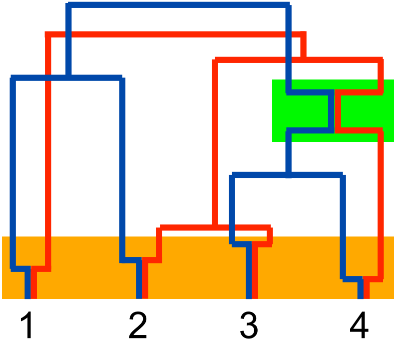


- In fact, these are the only events (or nonevents) of relevance.

Trivial event

Another "nontrivial" event?

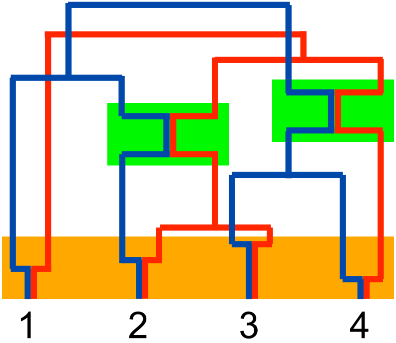
1. First coalescence: $O(1)$.



Trivial event

Another "nontrivial" event?

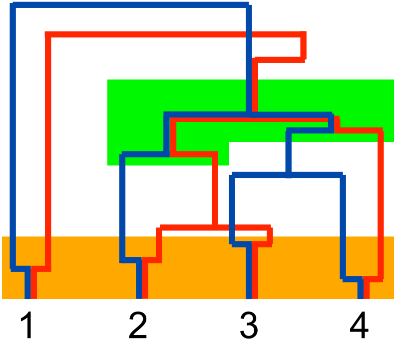
1. First coalescence: $O(1)$.
2. Second coalescence: $O(\rho^{-1})$.



Trivial event

Another "nontrivial" event?

1. First coalescence: $O(1)$.
2. Second coalescence: $O(\rho^{-1})$.
3. Third coalescence: $O(\rho^{-1})$.

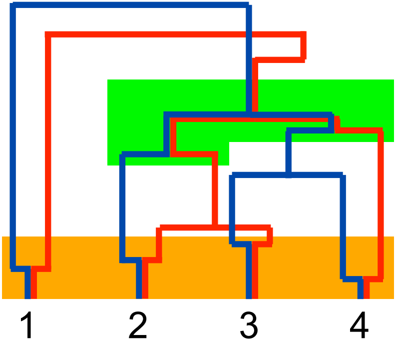


Trivial event

Another "nontrivial" event?

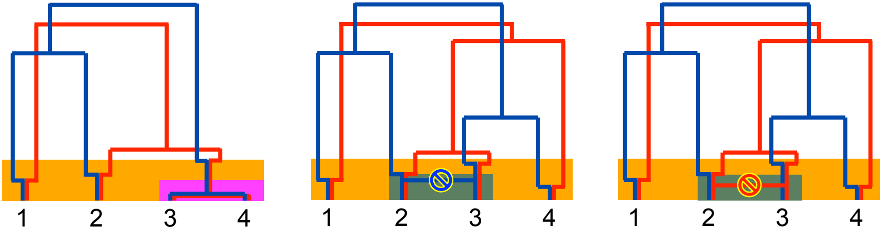
1. First coalescence: $O(1)$.
2. Second coalescence: $O(\rho^{-1})$.
3. Third coalescence: $O(\rho^{-1})$.

Overall probability of this event is $O(\rho^{-2})$—i.e. negligible.



- A coupling between the ARG and a pair of independent coalescent trees can make these arguments rigorous.

Coupling argument (outline)



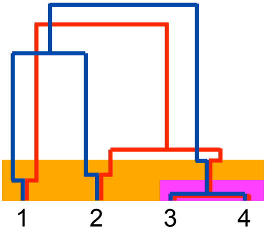
F_1: Type 1 failure F_2: Type 2 failure F_3: Type 3 failure

Outline of argument

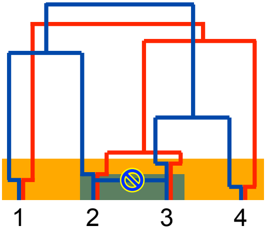
Show that:

- $\mathbb{P}(F_1) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,

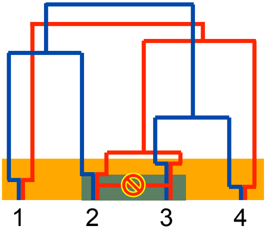
Coupling argument (outline)



F_1: Type 1 failure



F_2: Type 2 failure



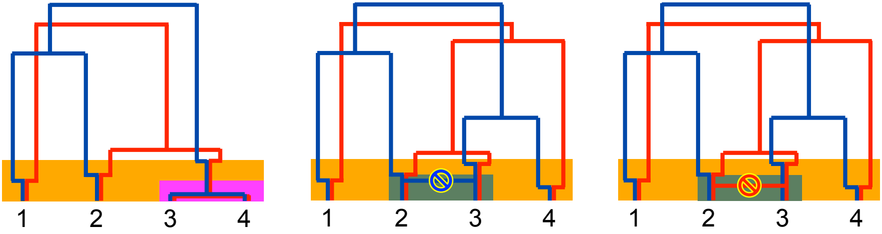
F_3: Type 3 failure

Outline of argument

Show that:

- $\mathbb{P}(F_1) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_2) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,

Coupling argument (outline)



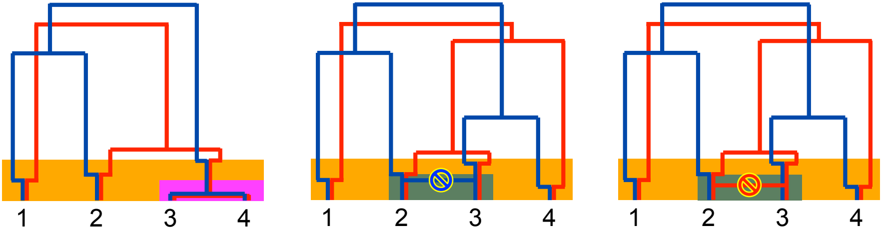
F_1: Type 1 failure F_2: Type 2 failure F_3: Type 3 failure

Outline of argument

Show that:

- $\mathbb{P}(F_1) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_2) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_3) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,

Coupling argument (outline)



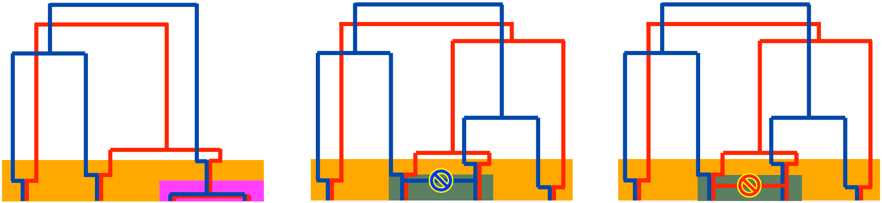
F_1: Type 1 failure F_2: Type 2 failure F_3: Type 3 failure

Outline of argument

Show that:

- $\mathbb{P}(F_1) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_2) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_3) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_i \cap F_j) = O\left(\frac{1}{\rho^2}\right), i \neq j$,

Coupling argument (outline)



F_1: Type 1 failure F_2: Type 2 failure F_3: Type 3 failure

Outline of argument

Show that:

- $\mathbb{P}(F_1) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_2) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_3) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right)$,
- $\mathbb{P}(F_i \cap F_j) = O\left(\frac{1}{\rho^2}\right), i \neq j$,
- $\mathbb{P}(\text{any other type of failure}) = O\left(\frac{1}{\rho^2}\right)$.

Outline of argument (*cont.*)

$$q(\boldsymbol{c};\rho) = \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid F_1^{\complement};\rho)$$

Outline of argument (*cont.*)

$$q(\boldsymbol{c};\rho) = \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid F_1^{\complement};\rho)$$
$$= \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid (F_2 \cup F_3)^{\complement};\infty)$$

Outline of argument (*cont.*)

$$q(\boldsymbol{c};\rho) = \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid F_1^{\complement};\rho)$$
$$= \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid (F_2 \cup F_3)^{\complement};\infty)$$

$$q(\boldsymbol{c} \mid F_1;\rho) = \sum_{i,j} \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q(\boldsymbol{c} - \boldsymbol{e}_{ij};\infty),$$

Outline of argument (*cont.*)

$$q(\boldsymbol{c};\rho) = \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid F_1^{\complement};\rho)$$
$$= \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid (F_2 \cup F_3)^{\complement};\infty)$$

$$q(\boldsymbol{c} \mid F_1;\rho) = \sum_{i,j} \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q(\boldsymbol{c} - \boldsymbol{e}_{ij};\infty),$$

$$q(\boldsymbol{c} \mid F_2;\rho) = \sum_{i} \frac{\binom{c_{i\cdot}}{2}}{\binom{c}{2}} q(\boldsymbol{c}_A - \boldsymbol{e}_i;\infty)q(\boldsymbol{c}_B;\infty),$$

$$q(\boldsymbol{c} \mid F_3;\rho) = \sum_{j} \frac{\binom{c_{\cdot j}}{2}}{\binom{c}{2}} q(\boldsymbol{c}_A;\infty)q(\boldsymbol{c}_B - \boldsymbol{e}_j;\infty),$$

Outline of argument (*cont.*)

$$q(\boldsymbol{c};\rho) = \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid F_1^{\complement};\rho)$$
$$= \mathbb{P}(F_1)q(\boldsymbol{c} \mid F_1;\rho) + \mathbb{P}(F_1^{\complement})q(\boldsymbol{c} \mid (F_2 \cup F_3)^{\complement};\infty)$$

$$q(\boldsymbol{c} \mid F_1;\rho) = \sum_{i,j} \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q(\boldsymbol{c} - \boldsymbol{e}_{ij};\infty),$$

$$q(\boldsymbol{c} \mid F_2;\rho) = \sum_{i} \frac{\binom{c_{i\cdot}}{2}}{\binom{c}{2}} q(\boldsymbol{c}_A - \boldsymbol{e}_{i};\infty)q(\boldsymbol{c}_B;\infty),$$

$$q(\boldsymbol{c} \mid F_3;\rho) = \sum_{j} \frac{\binom{c_{\cdot j}}{2}}{\binom{c}{2}} q(\boldsymbol{c}_A;\infty)q(\boldsymbol{c}_B - \boldsymbol{e}_{j};\infty),$$

$$q(\boldsymbol{c} \mid (F_2 \cup F_3)^{\complement};\infty) = \left[\frac{1}{1 - \mathbb{P}(F_2) - \mathbb{P}(F_3)}\right][q(\boldsymbol{c};\infty)$$
$$- \mathbb{P}(F_2)q(\boldsymbol{c} \mid F_2;\infty) - \mathbb{P}(F_3)q(\boldsymbol{c} \mid F_3;\infty)].$$

Theorem.

The sampling distribution of the loose linkage coalescent is

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + O\left(\frac{1}{\rho^2}\right).$$

Theorem.

The sampling distribution of the loose linkage coalescent is

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + O\left(\frac{1}{\rho^2}\right).$$

Explanation for the simple form of $q_1(\boldsymbol{c})$

$$q_1(\boldsymbol{c}) = \sum_{i,j} \overbrace{\binom{c_{ij}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{c}_B - \boldsymbol{e}_j)}^{\text{A randomly chosen pair of haplotypes coalesces before time } T}$$

Theorem.

The sampling distribution of the loose linkage coalescent is

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + O\left(\frac{1}{\rho^2}\right).$$

Explanation for the simple form of $q_1(\boldsymbol{c})$

$$q_1(\boldsymbol{c}) = \sum_{i,j} \overbrace{\binom{c_{ij}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{c}_B - \boldsymbol{e}_j)}^{\text{A randomly chosen pair of haplotypes coalesces before time } T}$$

Theorem.

The sampling distribution of the loose linkage coalescent is

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + O\left(\frac{1}{\rho^2}\right).$$

Explanation for the simple form of $q_1(\boldsymbol{c})$

$$q_1(\boldsymbol{c}) = \overbrace{\sum_{i,j} \binom{c_{ij}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{c}_B - \boldsymbol{e}_j)}^{\text{A randomly chosen pair of haplotypes coalesces before time } T} + \overbrace{\binom{c}{2} q^A(\boldsymbol{c}_A) q^B(\boldsymbol{c}_B)}^{\text{Otherwise, the trees are independent}}$$

Theorem.

The sampling distribution of the loose linkage coalescent is

$$q(\boldsymbol{c}) = q_0(\boldsymbol{c}) + \frac{q_1(\boldsymbol{c})}{\rho} + O\left(\frac{1}{\rho^2}\right).$$

Explanation for the simple form of $q_1(\boldsymbol{c})$

$$q_1(\boldsymbol{c}) = \overbrace{\sum_{i,j}\binom{c_{ij}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{c}_B - \boldsymbol{e}_j)}^{\text{A randomly chosen pair of haplotypes coalesces before time } T} + \overbrace{\binom{c}{2} q^A(\boldsymbol{c}_A) q^B(\boldsymbol{c}_B)}^{\text{Otherwise, the trees are independent}}$$

$$\underbrace{- q^B(\boldsymbol{c}_B)\sum_i \binom{c_{i\cdot}}{2} q^A(\boldsymbol{c}_A - \boldsymbol{e}_i) - q^A(\boldsymbol{c}_A)\sum_j \binom{c_{\cdot j}}{2} q^B(\boldsymbol{c}_B - \boldsymbol{e}_j)}_{\text{...with the restriction that no "prohibited coalescences" occur before time } T}.$$

A new “loose linkage” coalescent

- The previous decomposition picks out the important events in the ARG.

A new "loose linkage" coalescent

- The previous decomposition picks out the important events in the ARG.
- We can define a new coalescent process which keeps only these events.

A new "loose linkage" coalescent

- The previous decomposition picks out the important events in the ARG.
- We can define a new coalescent process which keeps only these events.

A new "loose linkage" coalescent

- The previous decomposition picks out the important events in the ARG.
- We can define a new coalescent process which keeps only these events.

Algorithm: Loose linkage coalescent

1. With probability $\frac{1}{\rho}\binom{c}{2}$:
    - Choose a pair (uniformly) from the c haplotypes to coalesce.
    - Every lineage undergoes recombination until time T, with this sole coalescence inserted randomly into the sequence of recombinations.
    - Simulate the rest as two independent coalescent trees.

A new "loose linkage" coalescent

- The previous decomposition picks out the important events in the ARG.
- We can define a new coalescent process which keeps only these events.

Algorithm: Loose linkage coalescent

1. With probability $\frac{1}{\rho}\binom{c}{2}$:
   - Choose a pair (uniformly) from the c haplotypes to coalesce.
   - Every lineage undergoes recombination until time T, with this sole coalescence inserted randomly into the sequence of recombinations.
   - Simulate the rest as two independent coalescent trees.
2. Otherwise:
   - Simulate from two independent coalescent trees conditioned not to have any prohibited coalescences before time T, as described earlier.

Summary

1. Both the Wright-Fisher diffusion with recombination and the ARG possess a deep and regular structure when the recombination rate increases, which we have described.

Summary

1. Both the Wright-Fisher diffusion with recombination and the ARG possess a deep and regular structure when the recombination rate increases, which we have described.
2. This structure can be exploited to derive simple approximations to these models.

Summary

1. Both the Wright-Fisher diffusion with recombination and the ARG possess a deep and regular structure when the recombination rate increases, which we have described.
2. This structure can be exploited to derive simple approximations to these models.
3. Our work also provides the first closed-form extension of Ewens sampling formula for multilocus models.

Summary

1. Both the Wright-Fisher diffusion with recombination and the ARG possess a deep and regular structure when the recombination rate increases, which we have described.
2. This structure can be exploited to derive simple approximations to these models.
3. Our work also provides the first closed-form extension of Ewens sampling formula for multilocus models.

Summary

1. Both the Wright-Fisher diffusion with recombination and the ARG possess a deep and regular structure when the recombination rate increases, which we have described.
2. This structure can be exploited to derive simple approximations to these models.
3. Our work also provides the first closed-form extension of Ewens sampling formula for multilocus models.

Future work

- Further generalizations:
  - More than two loci
  - Natural selection

Summary

1. Both the Wright-Fisher diffusion with recombination and the ARG possess a deep and regular structure when the recombination rate increases, which we have described.
2. This structure can be exploited to derive simple approximations to these models.
3. Our work also provides the first closed-form extension of Ewens sampling formula for multilocus models.

Future work

- Further generalizations:
    - More than two loci
    - Natural selection
- Better tools:
    - Duality between the two models?
    - “Separation of timescales” (cf. Möhle, 1998)

References

- Jenkins, P.A., Fearnhead, P., and Song, Y.S. (2015). "Tractable stochastic models of evolution for weakly correlated loci." *Electronic Journal of Probability*, **20** (58): 1–26.
- Jenkins, P.A. and Song, Y.S. (2012). "Padé approximants and exact two-locus sampling distributions." *Ann. Appl. Prob.*, 22(2): 576–607.

Acknowledgements

- Discussions with Song lab, Bob Griffiths, Charles Langley, Ben Peter, John Pool, Nadia Singh
- Simons Institute for the Theory of Computing
- Isaac Newton Institute

Research supported in part by EPSRC (PAJ, PF), NIH (PAJ, YSS), Alfred P. Sloan Research Fellowship (YSS), and a Packard Fellowship for Science and Engineering (YSS).

Covariances of the Moran model

$$\lim_{\mathrm{d}t\to 0}(\mathrm{d}t)^{-1}\mathbb{E}[\Delta \boldsymbol{M}^{(N)}(\tau) \mid \boldsymbol{M}^{(N)}(\tau)=\boldsymbol{m}]$$
$$= N^{\beta-1}\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{E}[\Delta \boldsymbol{M}^{(N)}(\tau) \mid \boldsymbol{M}^{(N)}(\tau)=\boldsymbol{m}] =: \boldsymbol{w}^{(N)}(\boldsymbol{m}),$$

$$\lim_{\mathrm{d}t\to 0}(\mathrm{d}t)^{-1}\mathrm{cov}[\Delta \boldsymbol{M}^{(N)}(\tau) \mid \boldsymbol{M}(\tau)=\boldsymbol{m}]$$
$$= N^{\beta-1}\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathrm{cov}[\Delta \boldsymbol{M}^{(N)}(\tau) \mid \boldsymbol{M}^{(N)}(\tau)=\boldsymbol{m}] =: N^{\beta-1}\boldsymbol{s}^{(N)}(\boldsymbol{m}),$$

Thus, with $\boldsymbol{m}=(x_1,\ldots,x_K,y_1,\ldots,y_L,d_{11},\ldots,d_{KL})$, we have

$$\boldsymbol{w}^{(N)}(\boldsymbol{m}) = \boldsymbol{w}(\boldsymbol{m}) + O(N^{\beta-1}),$$

where
$$\boldsymbol{w}(\boldsymbol{m}) = \Big(\underbrace{0,\ldots 0}_{K},\underbrace{0,\ldots 0}_{L},\underbrace{-\frac{\rho_\beta}{2}d_{11},\ldots,-\frac{\rho_\beta}{2}d_{KL}}_{K\times L}\Big)',$$

Covariances of the Moran model (II)

$\boldsymbol{s}^{(N)}(\boldsymbol{m}) = \boldsymbol{s}(\boldsymbol{m}) + O(N^{-\beta})$ is determined in a similar fashion:

$$\boldsymbol{s}(\boldsymbol{m}) = \begin{bmatrix} \boldsymbol{s}_{\mathbf{XX}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{XY}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{XD}}(\boldsymbol{m}) \\ \boldsymbol{s}_{\mathbf{XY}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{YY}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{YD}}(\boldsymbol{m}) \\ \boldsymbol{s}_{\mathbf{XD}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{YD}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{DD}}(\boldsymbol{m}) \end{bmatrix},$$

where

$$\begin{aligned} [\boldsymbol{s}_{\mathbf{XX}}(\boldsymbol{m})]_{ik} &= x_i(\delta_{ik} - x_k), \\ [\boldsymbol{s}_{\mathbf{YY}}(\boldsymbol{m})]_{jl} &= y_j(\delta_{jl} - y_l), \\ [\boldsymbol{s}_{\mathbf{XY}}(\boldsymbol{m})]_{ij} &= d_{ij}, \\ [\boldsymbol{s}_{\mathbf{XD}}(\boldsymbol{m})]_{i,kl} &= d_{kl}(\delta_{ik} - x_i) - x_k d_{il}, \\ [\boldsymbol{s}_{\mathbf{YD}}(\boldsymbol{m})]_{j,kl} &= d_{kl}(\delta_{jl} - y_j) - y_l d_{kj}, \\ [\boldsymbol{s}_{\mathbf{DD}}(\boldsymbol{m})]_{ij,kl} &= x_i y_j(\delta_{ik} - x_k)(\delta_{jl} - y_l) + d_{kj} x_i y_l + d_{il} x_k y_j \\ &\quad + d_{ij}(x_k y_l - \delta_{ik} y_l - \delta_{jl} x_k) \\ &\quad + d_{kl}(x_i y_j - \delta_{ik} y_j - \delta_{jl} x_i) + d_{ij}(\delta_{ik}\delta_{jl} - d_{kl}). \end{aligned}$$