A conditional coalescent limit in fixed pedigrees

Matthias Birkner
based on joint work with Andrey Tyukin

Stochastic Dynamics Out of Equilibrium: Life sciences
Institut Henri Poincaré, 16–18 Mai 2017

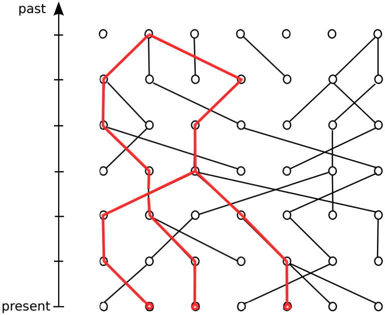
JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Outline

1. Introduction
2. A diploid Cannings model
   - Jointly averaged limit
3. Conditional coalescent limit
4. Outlook

Example: Wright-Fisher model (Sewall Wright, Ronald Fisher, 1930ies)

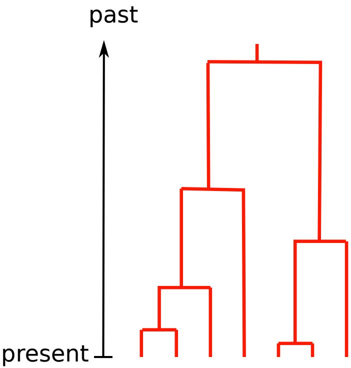
- A (haploid) population of N individuals per generation,
- assign to each individual (gene) in the present generation a ‘parent’ at random from the previous generation (i.i.d. picks),



The most “basic” (and popular) model in mathematical population genetics, allows in particular to model genealogies of samples

∃ very many extensions

The *n*-coalescent



Robust limit model of genealogies of sample in (haploid) exchangeable population models as population size $N \to \infty$

Mathematically, a partition-valued random process, each pair of blocks (=ancestral lines) merges at rate 1

Many extensions, in particular also for *diploid* organisms

Wakeley et al's concern

John Wakeley, Léandra King, Bobbi S. Low and Sohini Ramachandran (*Genetics* 2012) write:

"*We address a conceptual flaw in the backward-time approach to population genetics called coalescent theory as it is applied to diploid biparental organisms. Specifically, the way random models of reproduction are used in coalescent theory is not justified. Instead, the population pedigree for diploid organisms – that is, the set of all family relationships among members of the population – although unknown, should be treated as a fixed parameter, not as a random quantity.*"

Wakeley et al's concern, remarks

- Wakeley et al 2012 consider (simulated) genetic data/gene genealogies in fixed (simulated) pedigrees and observe that for several models, when N is reasonably large
  - various tests (Tajima's D based on independent loci, a test if the pairwise coalescence time is exponential) do not reject the null hypothesis of a standard Kingman coalescent
  - pair coalescence probabilities for genes become 'flat' (as in classical haploid models) after a short initial period

- Relation between pedigree ancestry and genetic ancestry has received some attention in the literature, e.g. Chang (1999) and discussion; Derrida, Manrubia & Zanette (1999,2000); Matsen & Evans (2008); Barton & Etheridge (2011), ...

  Note: different time scales relevant:

$$\approx \log N \text{ (pedigree)} \quad \text{vs.} \quad \approx N \text{ (gene genealogy)}$$

Wakeley et al's concern, remarks 2

- Apparently, so far the issue raised by Wakeley et al has only been considered in 'rigorous' form by Blath, Kadow & Ortgiese (2015), who studied the 'cyclical Wright-Fisher model.'
- This talk (nutshell version): A limit theorem for the conditional distribution of the gene genealogy given the pedigree that corroborates/confirms Wakeley et al's observations in the $N \to \infty$ limit (for a relatively general class of diploid biparental population models).

A diploid Cannings model

(a very small extension of the model from Möhle & Sagitov, *J. Math. Biol.*, 2003)

- fixed-size, panmictic population, diploid
- N females, N males per generation
- consider one (neutral, autosomal) locus
- each individual has two gene copies (one inherited from each parent), Mendelian inheritance
- $\nu_{ij}^{(r)}$ ... no. of offspring of female i and male j (in generation r), i.i.d. over generations
- $\sum_{i,j=1}^{N} \nu_{ij} = 2N$ (drop superscript $^{(r)}$ for 'generic' case),
- exchangeability condition: $(\nu_{i,j})_{i,j=1,\ldots,N} =^{d} (\nu_{\sigma(i),\sigma'(j)})_{i,j=1,\ldots,N}$ for any permutations $\sigma, \sigma' \in \mathcal{S}_N$
- random sex assignment (i.e., out of the $2N$ children pick N w.o. replacement to be the daughters, the remaining N are the sons)

Examples: diploid two-sex Wright-Fisher model,
Möhle & Sagitov's N couples Cannings model

A diploid Cannings model, bookkeeping

Pedigree:

- enumerate ind.s in each generation randomly (give females numbers $1, \dots, N$, males $N+1, \dots, 2N$, say)
- (k, r) … ind. no. k in gen. r
- $\Phi_{\mathrm{mo}}(k, r)$ … no. of (k, r)'s mother,
  $\Phi_{\mathrm{fa}}(k, r)$ … no. of (k, r)'s father,
  (assigned at random in accordance with offspring numbers $(\nu_{ij}^{(r-1)})_{ij}$)

  i.e. (k, r) is a descendant of $\big(\Phi_{\mathrm{mo}}(k, r), r-1\big)$ and $\big(\Phi_{\mathrm{fa}}(k, r), r-1\big)$
- $\Phi^{(N)} = \big((\Phi_{\mathrm{mo}}(k, r), \Phi_{\mathrm{fa}}(k, r); k = 1, \dots, 2N, r \in \mathbb{Z}\big)$
  is the (population) pedigree (or 'parentship graph')

A diploid Cannings model, bookkeeping 2

Chromosomes and Mendelian randomness:

- Individual (k, r) has chromosomes

  $(k, 1, r)$ [inherited from the mother, $(\Phi_{\text{mo}}(k, r), r-1)$] and

  $(k, 2, r)$ [inherited from the father, $(\Phi_{\text{fa}}(k, r), r-1)$]

- $M_{k,c,r}$ i.i.d., $\mathbb{P}(M_{k,c,r} = 1) = \frac{1}{2} = \mathbb{P}(M_{k,c,r} = 2)$, $k = 1, \dots, 2N$; $c = 1, 2$; $r \in \mathbb{Z}$

  $$(k, c, r) \text{ descends from } \begin{cases} \big(\Phi_{\text{mo}}(k, r), M_{k,1,r}, r-1\big), & c = 1, \\ \big(\Phi_{\text{fa}}(k, r), M_{k,2,r}, r-1\big), & c = 2. \end{cases}$$

Note: Two levels of randomness in this model

Random pedigree $\Phi^{(N)}$, 'Mendelian randomness' $(M_{k,c,r})$

A diploid Cannings model, notation/bookkeeping 3

n-sample:

pick n chromosomes at random from generation 0
(We may think of n chromosomes from n distinct randomly chosen individuals or both chromosomes from $n/2$ individuals or s.th. in-between; this will not matter in the limit.)

For $1 \leq i, j \leq n$, $t \in 0, 1, 2, \ldots$ write $i \sim_t j$
if sampled chromosomes i and j descend from the same chromosome in generation $-t$

and $R_t^{(N,n)} := \{\text{equivalence classes under } \sim_t\}$.

$(R_t^{(N,n)})_{t \in \mathbb{N}_0}$ is a stochastic process with values in $\mathcal{E}_n$, the set of partitions of $\{1, \ldots, n\}$,
$R_0^{(N,n)} = \big\{\{1\}, \{2\}, \ldots, \{n\}\big\}$.

(Averaged) coalescent limit, preliminaries

Put $\nu_i^{\mathrm{f}} = \sum_{j=1}^{N} \nu_{ij}, \quad \nu_j^{\mathrm{m}} = \sum_{i=1}^{N} \nu_{ij}$

($\sum_{i=1}^{N} \nu_i^{\mathrm{f}} = 2N = \sum_{j=1}^{N} \nu_j^{\mathrm{f}}$ and $(\nu_1^{\mathrm{f}}, \dots, \nu_N^{\mathrm{f}})$, $(\nu_1^{\mathrm{m}}, \dots, \nu_N^{\mathrm{m}})$ are jointly exchangeable in the sense that $\big((\nu_1^{\mathrm{f}}, \dots, \nu_N^{\mathrm{f}}), (\nu_1^{\mathrm{m}}, \dots, \nu_N^{\mathrm{m}})\big) =^d \big((\nu_{\sigma(1)}^{\mathrm{f}}, \dots, \nu_{\sigma(N)}^{\mathrm{f}}), (\nu_{\sigma'(1)}^{\mathrm{m}}, \dots, \nu_{\sigma'(N)}^{\mathrm{m}})\big)$ for any permutations $\sigma, \sigma' \in \mathcal{S}_N$.)

Let

$$c_N = \tfrac{1}{16(2N-1)} \mathbb{E}[\nu_1^{\mathrm{f}}(\nu_1^{\mathrm{f}} - 1)] + \tfrac{1}{16(2N-1)} \mathbb{E}[\nu_1^{\mathrm{m}}(\nu_1^{\mathrm{m}} - 1)]$$

be the pair coalescence probability (over one generation for two randomly chosen chromosomes from different individuals),

$$d_N = \frac{\mathbb{E}[\nu_1^{\mathrm{f}}(\nu_1^{\mathrm{f}} - 1)(\nu_1^{\mathrm{f}} - 2)]}{64(2N-1)(2N-2)} + \frac{\mathbb{E}[\nu_1^{\mathrm{m}}(\nu_1^{\mathrm{m}} - 1)(\nu_1^{\mathrm{m}} - 2)]}{64(2N-1)(2N-2)}$$

the triple coalescence probability.

(Averaged) coalescent limit

Lemma (a tiny variation on Möhle & Sagitov 2003)**.**
If $0 < c_N \to 0$ and $d_N / c_N \to 0$,

$$\mathscr{L}\big((R^{(N,n)}_{\lfloor t/c_N \rfloor})_{t \geq 0}\big) \Longrightarrow \mathscr{K}^{(n)} \quad \text{as } N \to \infty$$

(convergence in distribution on $\mathcal{D}([0,\infty), \mathcal{E}_n)$), where $\mathscr{K}^{(n)}$ is (the law of) Kingman's n-coalescent.

Note:

Distributional convergence refers to averaging over both levels of randomness, the random pedigree $\Phi^{(N)}$ and the 'Mendelian randomness' $(M_{k,c,r})$.

(Averaged) coalescent limit, remarks on the proof

$$(R^{(N,n)}_{\lfloor t/c_N \rfloor})_{t \geq 0} \underset{N\to\infty}{\Longrightarrow} \mathscr{K}^{(n)}$$

- $R^{(N,n)}$ is not a Markov chain
- but a suitably enriched version $\widetilde{R}^{(N,n)}$ – which keeps track of the grouping of blocks ($\widehat{=}$ancestral chromosomes) into diploid individuals – is
- $\widetilde{R}^{(N,n)}$ has state space $\widetilde{\mathcal{E}}_n = \{\text{partitions of } \{1,\dots,n\}, \text{ possibly grouped into ordered pairs}\}$, which canonically 'contains' $\mathcal{E}_n$
- separation of time scales: breaking up diploid grouping takes $O(1)$, non-trivial coalescences takes $\Theta(1/c_N) \gg O(1)$
- Möhle's (1998) lemma
  (A, B_N ($m \times m$-)matrices such that $P := \lim_{r\to\infty} A^r$, $G := \lim_{N\to\infty} PB_NP$ exist. Then $\lim_{N\to\infty}(A + c_N B_N)^{\lfloor t/c_N \rfloor} = Pe^{tG}$.)

Conditional (‘quenched’) coalescent limit

Theorem.
If $0 < c_N \to 0$ and $d_N/c_N \to 0$,

$$\mathscr{L}\big((R^{(N,n)}_{\lfloor t/c_N \rfloor})_{t\geq 0} \,\big|\, \Phi^{(N)}\big) \Longrightarrow \mathscr{K}^{(n)} \quad \text{in probability as } N \to \infty$$

where $\mathscr{K}^{(n)}$ is (the law of) Kingman’s n-coalescent.

This implies in particular (for N suff. large):

- the rescaled pair coalescence time for two genes given $\Phi^{(N)}$ is approximately exponential
- statistical tests for the Kingman coalescent as a null hypothesis cannot reject the null hyp. based on a gene genealogy drawn from a ‘typical’ pedigree (with higher prob. than the significance level)

Conditional ('quenched') coalescent limit, proof idea

Encode $(R^{(N,n)}_{\lfloor t/c_N \rfloor})_{t\geq 0}$ via jump times

$$0 = T^{(N,n)}_n < T^{(N,n)}_{n-1} < T^{(N,n)}_{n-2} < \cdots < T^{(N,n)}_1$$

and skeleton chain

$$S^{(N,n)}_j := R^{(N,n)}_{T^{(N,n)}_j}, \quad j = n, n-1, \dots, 1.$$

Want to show

$$\mathbb{E}\Big[e^{-\lambda_1 T_1^{(N,n)} - \cdots - \lambda_n T_n^{(N,n)}} 1(S^{(N,n)}_1 = s_1, \dots, S^{(N,n)}_n = s_n) \Big| \Phi^{(N)}\Big] \xrightarrow[N\to\infty]{\mathbb{P}} \mathbb{E}_{\mathscr{K}^{(n)}}\Big[e^{-\lambda_1 T_1^{(n)} - \cdots - \lambda_n T_n^{(n)}} 1(S^{(n)}_1 = s_1, \dots, S^{(n)}_n = s_n)\Big]$$

for any $\lambda_1, \dots, \lambda_n \geq 0$, $s_1, \dots, s_n \in \mathcal{E}_n$.

We have (by the jointly averaged limit lemma)

$$\mathbb{E}\Big[\mathbb{E}\Big[e^{-\lambda_1 T_1^{(N,n)} - \cdots - \lambda_n T_n^{(N,n)}} 1(S^{(N,n)}_1 = s_1, \dots, S^{(N,n)}_n = s_n) \Big| \Phi^{(N)}\Big]\Big] \xrightarrow[N\to\infty]{\mathbb{P}} \mathbb{E}_{\mathscr{K}^{(n)}}\Big[e^{-\lambda_1 T_1^{(n)} - \cdots - \lambda_n T_n^{(n)}} 1(S^{(n)}_1 = s_1, \dots, S^{(n)}_n = s_n)\Big].$$

Conditional ('quenched') coalescent limit, proof idea 2

To show that

$$\mathrm{Var}\left[\mathbb{E}\left[e^{-\lambda_1 T_1^{(N,n)}-\cdots-\lambda_n T_n^{(N,n)}} 1(S_1^{(N,n)}=s_1,\ldots,S_n^{(N,n)}=s_n)\middle|\Phi^{(N)}\right]\right] \xrightarrow[N\to\infty]{} 0$$

note that

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbb{E}\left[e^{-\lambda_1 T_1^{(N,n)}-\cdots-\lambda_n T_n^{(N,n)}} 1(S_1^{(N,n)}=s_1,\ldots,S_n^{(N,n)}=s_n)\middle|\Phi^{(N)}\right]\right)^2\right]\\ &= \mathbb{E}\Big[e^{-\lambda_1 T_1^{(N,n)}-\cdots-\lambda_n T_n^{(N,n)}} 1(S_1^{(N,n)}=s_1,\ldots,S_n^{(N,n)}=s_n)\\ &\qquad\qquad \times e^{-\lambda_1 \widehat{T}_1^{(N,n)}-\cdots-\lambda_n \widehat{T}_n^{(N,n)}} 1(\widehat{S}_1^{(N,n)}=s_1,\ldots,\widehat{S}_n^{(N,n)}=s_n)\Big]\end{aligned}$$

where $\widehat{T}_j^{(N,n)}$, $\widehat{S}_j^{(N,n)}$ refer to a copy $(\widehat{R}_{\lfloor t/c_N\rfloor}^{(N,n)})_{t\geq 0}$ which uses the *same* pedigree $\Phi^{(N)}$ but *independent* 'Medelian coin flips' $(\widehat{M}_{k,c,r})$.

Conditional ('quenched') coalescent limit, proof idea 2

a suitably enriched version of $(R_s^{(N,n)}, \widehat{R}_s^{(N,n)})_{s\in\mathbb{N}_0}$ is a Markov chain on (a suitable enrichment of) $\widetilde{\mathcal{E}}_n \times \widetilde{\mathcal{E}}_n$
(which keeps track of grouping into diploid individuals and also of 'labelling' if a chromosome counts for $R^{(N,n)}$, for $\widehat{R}^{(N,n)}$ or for both)

joint dynamics 'factorises' as $N \to \infty$ (separation of time scales, and $R^{(N,n)}, \widehat{R}^{(N,n)}$ interact rarely), this yields

$$\mathbb{E}\left[\left(\mathbb{E}\left[e^{-\lambda_1 T_1^{(N,n)}-\cdots-\lambda_n T_n^{(N,n)}}1(S_1^{(N,n)}=s_1,\dots,S_n^{(N,n)}=s_n)\middle|\Phi^{(N)}\right]\right)^2\right]$$

$$\xrightarrow[N\to\infty]{}\left(\mathbb{E}\left[\mathbb{E}\left[e^{-\lambda_1 T_1^{(N,n)}-\cdots-\lambda_n T_n^{(N,n)}}1(S_1^{(N,n)}=s_1,\dots,S_n^{(N,n)}=s_n)\middle|\Phi^{(N)}\right]\right]\right)^2$$

which gives the claim.

Conditional ('quenched') coalescent limit

$$\mathscr{L}\big((R^{(N,n)}_{\lfloor t/c_N \rfloor})_{t\geq 0} \,\big|\, \Phi^{(N)}\big) \Longrightarrow \mathscr{K}^{(n)} \quad \text{in probability as } N \to \infty$$

Remark.
One can think of the theorem as a quenched limit result for systems of directed random walks in random environment (RWRE) on $\{1, \dots, 2N\} \times \{1, 2\} \times \mathbb{Z}_-$.

In fact, the idea to strengthen an averaged central limit theorem to a quenched central limit theorem by suitably controlling two copies of the random walk in the same random medium appears in the literature on RWRE, cf Bolthausen & Sznitman (2002), B., Černý, Depperschmidt & Gantert (2013).

Outlook

Possible extensions:

- local structure
- unequal sex ratios
- varying population sizes
- (partial) selfing [coalescent limit after possibly a first ‘scattering phase’]
- diplo-/haploid systems
- monoecious populations
- several unlinked/linked loci [a ‘quenched ARG’]
- but not: highly skewed offspring laws [there is no ‘Λ-coalescent analogue’]

Question

- behaviour on shorter timescales $1 \ll t \ll 1/c_N$?

Thank you for your attention!