Asymptotic results on the length of coalescent trees

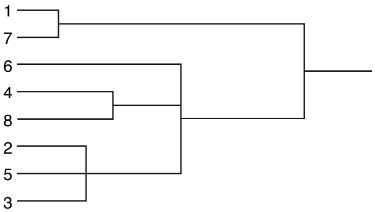
Arno Siri-Jégousse

MAP5 - Université Paris Descartes
joint work with J-F Delmas and J-S Dhersin

CIRM - May 26th 2000

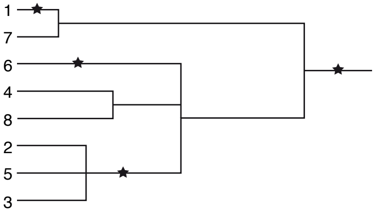
The Infinite Sites Model, Kimura (1969)

- We consider a genealogical tree of n individuals, of total length $L^{(n)}$
- Mutations occur at rate θ
- conditional on $L^{(n)}$, the number of mutations is distributed like Poisson with mean $\theta L^{(n)}$



The Infinite Sites Model, Kimura (1969)

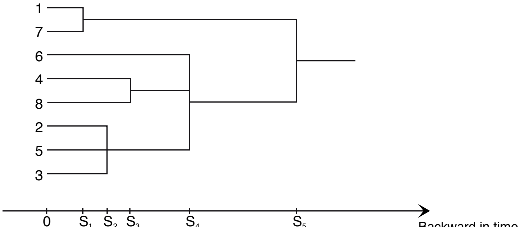
- We consider a genealogical tree of n individuals, of total length $L^{(n)}$
- Mutations occur at rate θ
- conditional on $L^{(n)}$, the number of mutations is distributed like Poisson with mean $\theta L^{(n)}$
- Each mutation appears in a new site, so that we can observe the number of mutations, $S^{(n)}$, as the number of segregating sites in our actual population.



$S^{(n)} = 3$

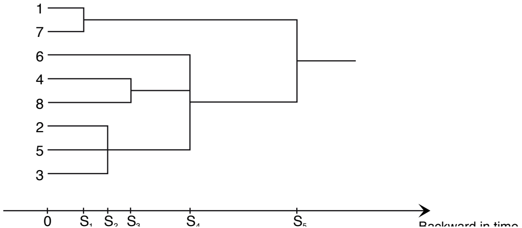
The coalescent

- $(\Pi_t^{(n)}, t \geq 0)$ is a continuous time Markov chain with values in $\mathcal{P}_n$, the set of partitions of $\{1, \ldots, n\}$



The coalescent

- $(\Pi_t^{(n)}, t \geq 0)$ is a continuous time Markov chain with values in $\mathcal{P}_n$, the set of partitions of $\{1, \ldots, n\}$
- $\Pi_0^{(n)} = \{1\}, \ldots, \{n\}$.
- Each block of $\Pi_t^{(n)} \in \mathcal{P}_n$ indicates individuals living at time 0 which have a common ancestor at time $-t$



The Λ-coalescent, Pitman (1999), Sagitov (1999)

If there are b blocks, each k-uplet of them merge to 1 at rate $\lambda_{b,k}$, independent of the current number of blocks :

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx)$$

for $2 \leq k \leq b$, where Λ is a finite measure on $[0,1]$

The Λ-coalescent, Pitman (1999), Sagitov (1999)

If there are b blocks, each k-uplet of them merge to 1 at rate $\lambda_{b,k}$, independent of the current number of blocks :

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx)$$

Definition

The markov process $\Pi^{(n)} = (\Pi_t^{(n)}, t \geq 0)$ with dynamics described above and starting from the trivial partition of $\mathcal{P}_n$ is called the (n-)Λ-coalescent

Consistence : $\Pi^{(n)}$ is the restriction of the so-called Λ-coalescent process Π defined on the set of partitions of $\mathbb{N}^*$.

Examples of Λ-coalescents

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx)$$

- $\Lambda = \delta_0$:
  Kingman's coalescent(1982)
  $\lambda_{b,2} = 1$, $\lambda_{b,k} = 0$ for $k \neq 2$
  Only two blocks can merge at a time.

Examples of Λ-coalescents

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx)$$

- $\Lambda = \delta_0$:
  Kingman's coalescent(1982)
  $\lambda_{b,2} = 1$, $\lambda_{b,k} = 0$ for $k \neq 2$
  Only two blocks can merge at a time.
- $\Lambda =$ Lebesgue on$[0,1]$:
  Bolthausen-Szmitman coalescent(1998)

Examples of Λ-coalescents

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx)$$

- $\Lambda = \delta_0$:
  Kingman's coalescent(1982)
  $\lambda_{b,2} = 1$, $\lambda_{b,k} = 0$ for $k \neq 2$
  Only two blocks can merge at a time.
- $\Lambda =$ Lebesgue on$[0,1]$:
  Bolthausen-Szmitman coalescent(1998)
- Λ is a $\beta(2-\alpha,\alpha)$ distribution, $\alpha \in (1,2)$:
  $\Lambda(dx) = C_0 x^{1-\alpha}(1-x)^{\alpha-1}\mathbf{1}_{[0,1]}(x)dx$.
  Beta-coalescent

Hypothesis

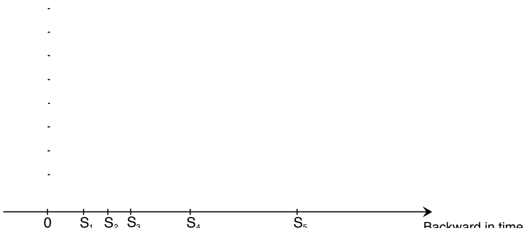
Let $\rho(t) = \int_t^1 \frac{\Lambda(dx)}{x^2}$. We will assume that :

$$\rho(t) = C_0 t^{-\alpha} + O\left(t^{-\alpha+\zeta}\right)$$

with $\alpha \in (1,2)$ and $\zeta > 1 - \frac{1}{\alpha}$.
This includes the Beta-coalescent case.

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

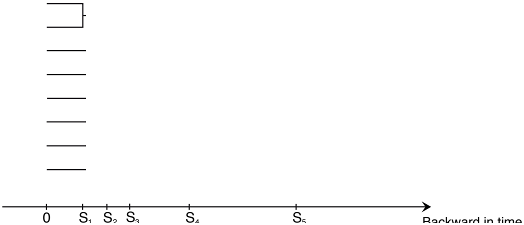
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.



Time of next jump $\sim \mathcal{E}(g_8)$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

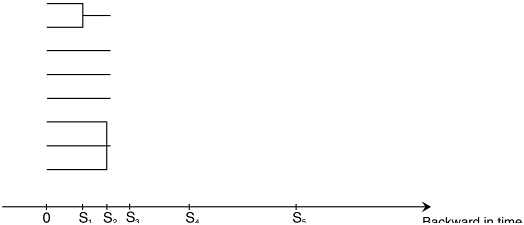
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.



Time of next jump $\sim \mathcal{E}(g_7)$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

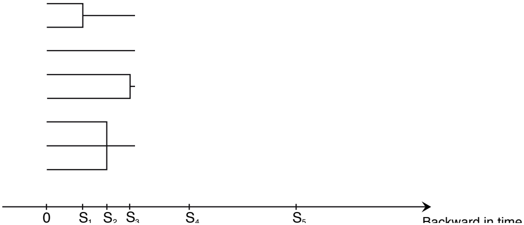
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.



Time of next jump $\sim \mathcal{E}(g_5)$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

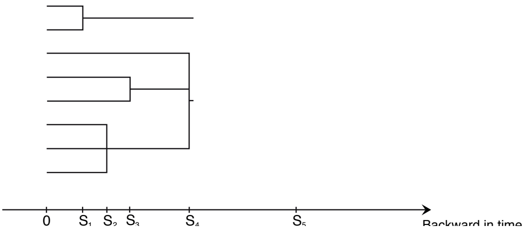
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.



Time of next jump $\sim \mathcal{E}(g_4)$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

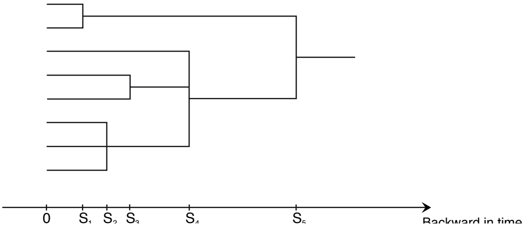
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.



Time of next jump $\sim \mathcal{E}(g_2)$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

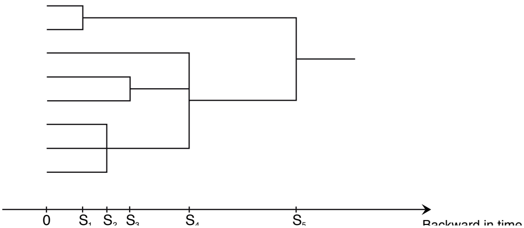
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.



until we reach the common ancestor

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

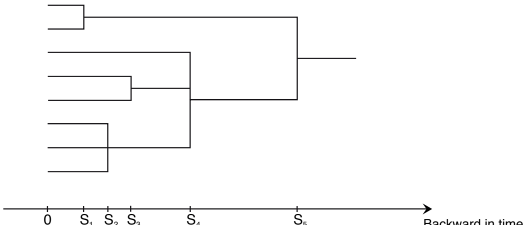
- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.
- $Y_k^{(n)}$: number of blocks after k coalescences.



$Y_0^{(8)} = 8,\ Y_1^{(8)} = 7,\ Y_2^{(8)} = 5,\ Y_3^{(8)} = 4,\ Y_4^{(8)} = 2, Y_5^{(8)} = 1$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.
- $Y_k^{(n)}$: number of blocks after k coalescences.
- τ_n : total number of coalescences.



$\tau_8 = 5$

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

- $g_b = \sum_{l=1}^{b-1} \binom{b}{l+1} \lambda_{b,l+1}$: rate of the next jump of the coalescent when there are b blocks. E_k are i.i.d rate 1 exponential r.v.
- $Y_k^{(n)}$: number of blocks after k coalescences.
- τ_n : total number of coalescences.

Question

What is the asymptotic behavior of $L^{(n)}$?

Approximations

$$L^{(n)} = \sum_{k=0}^{\tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

$$g_n \overset{+\infty}{\sim} C_0 \Gamma(2-\alpha) n^{\alpha}$$

Replacing E_k's by their man,1, we approximate $L^{(n)}$ by

$$\hat{L}^{(n)} = \sum_{k=0}^{\tau_n - 1} \left(Y_k^{(n)}\right)^{1-\alpha}$$

Asymptotics of τ_n

Proposition

$$n^{-\frac{1}{\alpha}}\left(n - \frac{\tau_n}{\alpha - 1}\right) \xrightarrow{\mathcal{L}} V_{\alpha-1}$$

where $(V_t, t \geq 0)$ *is an* α*-stable Lévy process with non positive jumps with Laplace exponent* $\psi(u) = u^\alpha/(\alpha - 1)$*.*

This result was also obtained by Iksanow and Möhle (2007) and Gnedin and Yakubovich (2008) with quite similar hypothesis.

Asymptotics of the length

Let $\gamma = \alpha - 1$.
We establish a first step to convergence and asymptotics of $L^{(n)}$ by giving results for $L_t^{(n)}$, the length of the coalescent tree up to the $\lfloor nt \rfloor$-th coalescence, for $t \in (0, \gamma)$.

$$L_t^{(n)} = \sum_{k=0}^{\lfloor nt \rfloor \wedge \tau_n - 1} Y_k^{(n)} \frac{E_k}{g_{Y_k^{(n)}}}$$

As $\tau_n \sim \gamma n$, intuitively we have $L_\gamma^{(n)}$ close to $L^{(n)}$. This gives an idea of the results we should obtain for $L^{(n)}$.

Main result

Theorem

Let $v(t) = \int_0^t (1 - \frac{r}{\gamma})^{-\gamma} dr$ *and* $V_t^* = \int_0^t (1 - \frac{r}{\gamma})^{-\gamma} V_r dr$ *Under our conditions, for all* $t \in (0, \gamma)$*,*

1. $n^{-2+\alpha} L_t^{(n)} \xrightarrow{P} \frac{v(t)}{C_0 \Gamma(2-\alpha)}$
2. *For* $\alpha \in (1, \frac{1+\sqrt{5}}{2})$

$$n^{-1+\alpha-\frac{1}{\alpha}} (L_t^{(n)} - \frac{v(t)}{C_0 \Gamma(2-\alpha)} n^{2-\alpha}) \xrightarrow{\mathcal{L}} V_t^*$$

3. *For* $\alpha \in [\frac{1+\sqrt{5}}{2}, 2)$*, if* $\varepsilon > 0$

$$n^{-\varepsilon} (L_t^{(n)} - \frac{v(t)}{C_0 \Gamma(2-\alpha)} n^{2-\alpha}) \xrightarrow{P} 0$$

Remarks

1. In the Beta-coalescent case, Berestycki et al. (2007) have already shown that

$$n^{-2+\alpha} L^{(n)} \xrightarrow{P} \frac{\Gamma(\alpha)\alpha(\alpha-1)}{2-\alpha}$$

2. Moreover in this case, we have $C_0 = \frac{1}{\alpha\Gamma(2-\alpha)\Gamma(\alpha)}$, and so

$$\frac{v(\gamma)}{C_0\Gamma(2-\alpha)} = \frac{\Gamma(\alpha)\alpha(\alpha-1)}{2-\alpha}$$

which means that the (coarse) approximation of $L^{(n)}$ by $L_\gamma^{(n)}$ leads to the good limit.

Let's go back to the infinite sites model.
$S^{(n)}$ is closely related to $L^{(n)}$ so we can obtain an asymptotic result for $S_t^{(n)}$, the number of mutations in the tree up to $\lfloor nt \rfloor$th coalescence.

Asymptotics of $S_t^{(n)}$

Let $a(t) = v(t)/C_o\Gamma(2-\alpha)$.

Corollary

Under our hypothesis, let $t \in (0,\gamma)$ and G be a standard gaussian r.v. independant of V

1. *For $\alpha \in (1, \sqrt{2})$*

$$n^{-1+\alpha-\frac{1}{\alpha}}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) \xrightarrow{\mathcal{L}} \theta V_t^*$$

2. *For $\alpha \in (\sqrt{2}, 2)$*

$$n^{-1+\alpha/2}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) \xrightarrow{\mathcal{L}} \sqrt{\theta a(t)}G$$

3. *For $\alpha = \sqrt{2}$*

$$n^{-1+\alpha/2}(S_t^{(n)} - \theta a(t) n^{2-\alpha}) \xrightarrow{\mathcal{L}} \theta V_t^* + \sqrt{\theta a(t)}G$$

Outlooks

- we now have an idea of the behavior of the total length
- Parametric estimation (of θ, of α)