Longueur des arbres de coalescence

Jean-François Delmas

http://cermics.enpc.fr/~delmas

La Londe, Septembre 2007

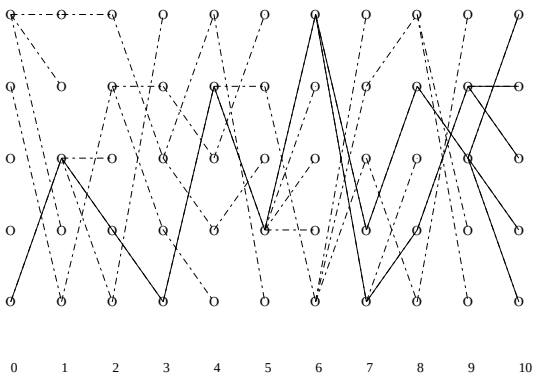
Plan

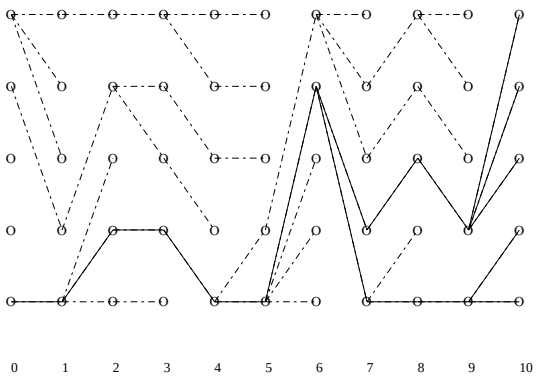
Avec J.-S. Dhersin et A. Siri-Jégousse

Le modèle de Wright (1931) et Fisher (1922)

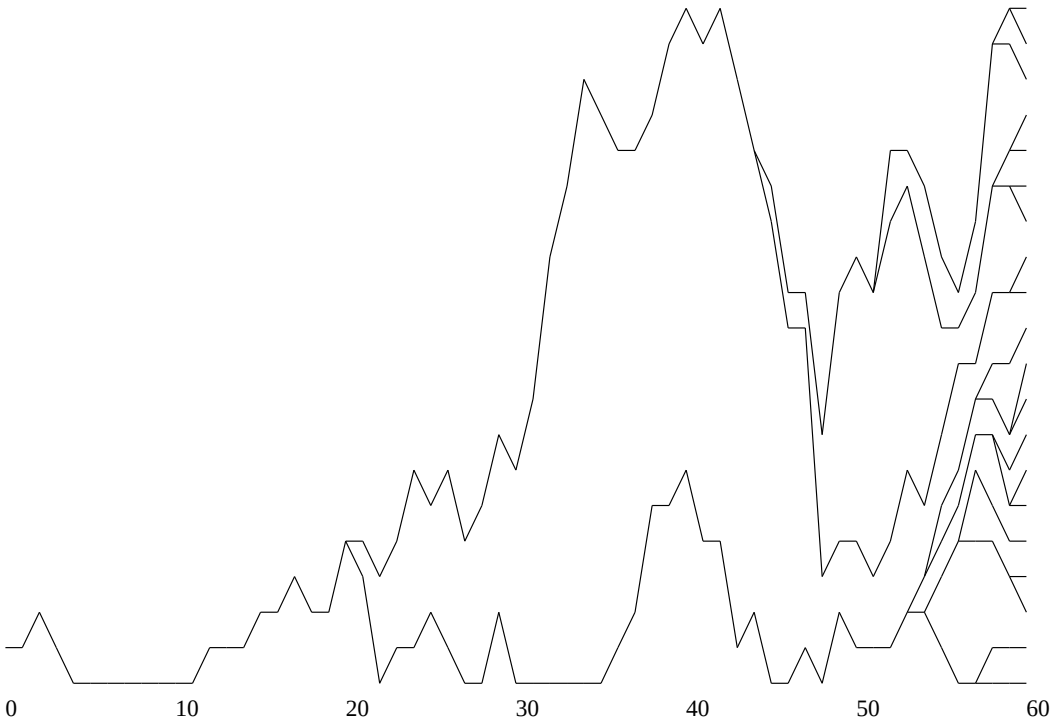
1) N individus asexués.
2) À l'instant k les N individus de la génaration $k-1$ meurent et donnent naissance à N enfants.
3) Le nombre d'enfants de chaque individu est aléatoire: chaque enfant choisit son parent au hasard (absence de sélection).

Généalogie avec $N = 5$ sur 10 générations





Généalogie de $N = 20$ individus sur 60 générations



Le coalescent de Kingman (1982)

- Temps de coalescence de 2 individus: $T_N \sim \text{Géom}(1/N)$.
- Asymptotique $N \to \infty$ et changement de temps en $1/N$:

$$T_N/N \simeq \mathcal{E}(1).$$

- 1er temps de coalescence pour n individus: minimum d'$\mathcal{E}(1)$ sur toutes les paires d'individus soit $\mathcal{E}\left(\frac{n(n-1)}{2}\right)$.
- À la limite: coalescence binaire et pas de coalescences simultanées.

Longueur de l'arbre de coalescence de Kingman

- Nombre d'ancêtres (des n individus) à la k-ème coalescence: Y_k,

$$Y_0 = n, \quad Y_{k+1} = Y_k - 1 \quad \text{soit} \quad Y_k = n - k \quad (k < n).$$

- Si r ancêtres, temps avant la prochaine coalescence: $\mathcal{E}\left(\frac{r(r-1)}{2}\right)$.
- Nombre de coalescences: $\tau_n = n - 1$.
- Longueur de l'arbre de coalescence: $L^{(n)}$

$$\sum_{r=n}^{2} r\mathcal{E}\left(\frac{r(r-1)}{2}\right) = \sum_{r=n}^{2} 2\mathcal{E}(r-1) = 2\sum_{\ell=1}^{n-1} \mathcal{E}(\ell) \simeq 2(\log(n-1) + \text{Gumbel}).$$

Mutations neutres (pas d'avantage sélectif)

- Toute nouvelle mutation affecte un nouveau site de l'ADN: modèle avec infinité de sites.
- Observation du nombre total de mutations sur n individus: $S^{(n)}$.
- Taux de mutation θ. $S^{(n)}$ de loi $\mathcal{P}oi(\theta L^{(n)})$.

- But: **estimer** θ.

Estimateur de Watterson (1975)

- Conditionnellement à $L^{(n)}$, $S^{(n)}$ de loi $\mathcal{P}oi(\theta L^{(n)})$ et donc

$$\frac{S^{(n)} - \theta L^{(n)}}{\sqrt{\theta L^{(n)}}} \xrightarrow[n\to\infty]{\textbf{loi}} \mathcal{N}(0,1).$$

- $L^n \simeq \mathbb{E}[L^{(n)}] + 2\text{Gumbel}$ avec $\mathbb{E}[L^{(n)}] = 2\sum_{r=1}^{n-1} \frac{1}{r}$.
- Estimateur de Watterson de θ: $S^{(n)}/\mathbb{E}[L^{(n)}]$ est convergent et asymptotiquement normal

$$\frac{S^{(n)} - \theta \mathbb{E}[L^{(n)}]}{\sqrt{\theta \mathbb{E}[L^{(n)}]}} \xrightarrow[n\to\infty]{\textbf{loi}} \mathcal{N}(0,1).$$

Coalescences multiples

- Modèle pertinent pour les huîtres et certains poissons (Sagitov 1999, Schweinsberg 2003).
- Pitman (1999) et Sagitov (1999): parmi b individus, ℓ coalescent avec taux

$$\lambda_{b,\ell} = \int_{[0,1]} x^{\ell-2}(1-x)^{b-\ell}\Lambda(dx), \quad 2 \leq \ell \leq b,$$

où Λ est une mesure finie sur $[0,1]$.

- $\binom{b}{\ell}$ choix possibles pour les ℓ individus qui coalescent.
- Pour b individus, le temps d'attente avant une coalescence est $\mathcal{E}(g_b)$ avec

$$g_b = \sum_{\ell=2}^{b} \binom{b}{\ell} \lambda_{b,\ell} = \int_{[0,1]} \left(1-(1-x)^b - bx(1-x)^{b-1}\right) \frac{\Lambda(dx)}{x^2}.$$

Des mesures de coalescence remarquables

$$\lambda_{b,\ell} = \int_{[0,1]} x^{\ell-2}(1-x)^{b-\ell}\Lambda(dx), \quad 2 \leq \ell \leq b,$$

- Coalescent de Kingman (1982): $\Lambda = \delta_0$.
- Coalescent de Bolthausen-Sznitman (1998): $\Lambda(dx) = \mathbf{1}_{[0,1]}(x)dx$.
- Beta$(2-\alpha, \alpha)$-coalescent, Schweinsberg (2003); Birkner et al. (2005); Bertoin et Le Gall (2006); Berestycki et al. (2007):

$$\Lambda(dx) = C(\alpha)x^{1-\alpha}(1-x)^{\alpha-1}\mathbf{1}_{]0,1[}(x)\ dx.$$

- But: **estimer** α.

Longueur de l'arbre et nombre de mutations

- Conditionnellement à $L^{(n)}$, $S^{(n)}$ de loi $\mathcal{P}oi(\theta L^{(n)})$ et donc

$$\frac{S^{(n)} - \theta L^{(n)}}{\sqrt{\theta L^{(n)}}} \xrightarrow[n\to\infty]{\textbf{loi}} \mathcal{N}(0,1).$$

- Coalescent de Kingman: vu en introduction.
- Coalescent de Bolthausen-Sznitman: Drmota, Iksanov, Möhle et Rösler (2007).

$$\frac{L^{(n)} - a_n}{b_n} \xrightarrow[n\to\infty]{\textbf{loi}} Z \quad \text{et} \quad \frac{S^{(n)} - \theta a_n}{\theta b_n} \xrightarrow[n\to\infty]{\textbf{loi}} Z,$$

avec $a_n \simeq \dfrac{n}{\log(n)}$ et $b_n = \dfrac{n}{\log(n)^2}$.

- Cas $\int_{]0,1]} x^{-1}\Lambda(dx) < \infty$: Möhle (2006).

Un peu plus général que le Beta coalescent

Soit $\rho(t) = \int_t^1 x^{-2}\Lambda(dx)$ pour $t \in (0,1]$. On suppose que

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

pour $\alpha \in (1,2)$, $C_0 > 0$ et $\zeta > 1 - 1/\alpha$.

- Couvre le cas du Beta-$(2-\alpha, \alpha)$ coalescent et plus généralement le cas $\Lambda(dx) \simeq cx^{1-\alpha}dx$ en 0.

Notations

- n individus à l'instant initial.
- Nombre d'ancêtres à la k-ème coalescence: Y_k (processus de mort: chaîne de Markov de transition P).
- On pose $X_k = Y_{k-1} - Y_k$ $(1 \leq k \leq \tau_n = \inf\{i; Y_i = 1\})$.

$$\mathbb{P}(X_k = \ell | Y_k = b) = P(b, b-\ell) = \frac{\binom{b}{\ell+1} \lambda_{b,\ell+1}}{g_b}.$$

- Temps d'attente dans l'état b: $\mathcal{E}(1)/g_b$.

Comportement asymptotique sous (H)

- $g_b \underset{+\infty}{\sim} C_0\Gamma(2-\alpha)b^\alpha$.
- $\mathbb{P}(X_k = \ell | Y_k = b) \xrightarrow[b\to\infty]{} \dfrac{\alpha}{\Gamma(2-\alpha)} \dfrac{\Gamma(\ell+1-\alpha)}{(\ell+1)!}$
- $X_k \xrightarrow[n\to\infty]{\textbf{loi}} X$ (k fixé) avec

$$\mathbb{E}[X] = \frac{1}{\alpha-1}, \quad \mathbb{E}[X^2] = +\infty, \quad \mathbb{P}(X \geq \ell) \underset{+\infty}{\sim} \frac{1}{\Gamma(2-\alpha)} \ell^{-\alpha}.$$

$$\phi(u) = \mathbb{E}[\exp -uX] \simeq 1 - \frac{u}{\alpha-1} + \frac{u^\alpha}{\alpha-1} \quad \text{(pour } u \text{ petit)}.$$

Un résultat sur le nombre de coalescences

Théorème

Nombre de coalescences: $\tau_n = \inf\{k, Y_k = 1 | Y_0 = n\}$. *On a*

$$n^{-1/\alpha}(n - \frac{\tau_n}{\alpha - 1}) \xrightarrow[n\to\infty]{\textbf{loi}} V_{\alpha-1},$$

où $(V_t, t \geq 0)$ *est un processus de Lévy d'exposant de Laplace* $\psi(u) = u^\alpha$.

Résultat obtenu également par Iksanov et Möhle (2007) et Gnedin et Yakubovich (2007).

Preuve

On pose $X_k = Y_k - Y_{k-1}$ (avec $Y_0 = n$). Pour u petit et ℓ grand:

$$\phi_\ell(u) = \mathbb{E}[\exp(-uX_1)|Y_0 = \ell] \simeq \phi(u) \simeq 1 - \frac{u}{\alpha - 1} + \frac{u^\alpha}{\alpha - 1}.$$

On considère la martingale $M_{v,k} = \exp\left(-\sum_{i=1}^k \left[vX_i + \log \phi_{Y_{i-1}}(v)\right]\right)$.

$$\begin{aligned} 1 = \mathbb{E}[M_{v,\tau_n}] &= \mathbb{E}\left[\exp\left(-\sum_{i=1}^{\tau_n}\left[vX_i + \log\phi_{Y_{i-1}}(v)\right]\right)\right] \\ &\simeq \mathbb{E}\left[\exp\left(-\sum_{i=1}^{\tau_n}\left[v(X_i - \frac{1}{\alpha-1}) + \frac{v^\alpha}{\alpha-1}\right]\right)\right] \\ &= \mathbb{E}\left[\exp -v(n - \frac{\tau_n}{\alpha-1}) - v^\alpha \frac{\tau_n}{\alpha-1}\right]. \end{aligned}$$

$v = \frac{1}{n} \Rightarrow \frac{\tau_n}{\alpha-1} \sim n$ et $v = un^{-\frac{1}{\alpha}} \Rightarrow \mathbb{E}\left[\exp -un^{-1/\alpha}(n - \frac{\tau_n}{\alpha-1}) - u^\alpha\right] \simeq 1.$

Un résultat sur la longueur partielle de l'arbre (I)

Pour $t \in (0, \alpha - 1]$, longueur de l'arbre jusqu'à la $\lfloor nt \rfloor$-ème coalescence

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

et son approximation déterministe ($Y_k \sim n - \frac{k}{\alpha - 1}$ et $g_b \sim C_0 \Gamma(2-\alpha) b^\alpha$)

$$\mathcal{L}_t^{(n)} = \sum_{k=0}^{\lfloor nt \rfloor} \left(n - \frac{k}{\alpha - 1} \right) \frac{1}{C_0 \Gamma(2-\alpha) \left(n - \frac{k}{\alpha - 1} \right)^\alpha} \sim n^{2-\alpha} a(t).$$

Théorème

Pour $t_0 \in (0, \alpha - 1)$, on a: $n^{-2+\alpha} \sup_{0 \le t \le t_0} |L_t^{(n)} - \mathcal{L}_t^{(n)}| \xrightarrow[n \to \infty]{\mathbb{P}} 0.$

Berestycki et al. (2007): $n^{-2+\alpha} |L^{(n)} - \mathcal{L}_{\alpha-1}^{(n)}| \xrightarrow[n \to \infty]{\mathbb{P}} 0.$

Un résultat sur la longueur partielle de l'arbre (II)

Théorème

Soit $\alpha_0 = \dfrac{1+\sqrt{5}}{2}$ *et* $V_t^* = \dfrac{\alpha-1}{C_0\Gamma(2-\alpha)} \displaystyle\int_0^t (1 - \frac{s}{\alpha-1})^{-\alpha} V_s \, ds.$

Si $\alpha \in (1, \alpha_0)$ *(et donc* $-1+\alpha-1/\alpha < 0$*), alors pour* $t \in (0, \alpha-1)$,

$$n^{-1+\alpha-1/\alpha} \left(L_t^{(n)} - \mathcal{L}_t^{(n)} \right) \xrightarrow[n\to\infty]{\textbf{loi}} V_t^*.$$

Nombre de mutations jusqu'à la $\lfloor nt \rfloor$-ème coalescence: $S_t^{(n)} \sim \mathcal{P}oi(\theta L_t^{(n)})$.
Soit $G \sim \mathcal{N}(0,1)$ indép. de V_t^*. Rappel: $\mathcal{L}_t^{(n)} \sim n^{2-\alpha} a(t)$.

Théorème ($0 < t < \alpha - 1$)

Si $\alpha \in (1, \sqrt{2})$, *alors*

$$n^{-1+\alpha-1/\alpha}(S_t^{(n)} - \theta \mathcal{L}_t^{(n)}) \xrightarrow[n\to\infty]{\textbf{loi}} \theta V_t^*.$$

Si $\alpha \in (\sqrt{2}, 2)$, *alors*

$$n^{-1+\alpha/2}(S_t^{(n)} - \theta \mathcal{L}_t^{(n)}) \xrightarrow[n\to\infty]{\textbf{loi}} \sqrt{\theta a(t)} G.$$

Si $\alpha = \sqrt{2}$, *alors* $-1 + \alpha - \frac{1}{\alpha} = 1 - \frac{\alpha}{2}$ *et*

$$n^{-1+\alpha-1/\alpha}(S_t^{(n)} - \theta \mathcal{L}_t^{(n)}) \xrightarrow[n\to\infty]{\textbf{loi}} \theta V_t^* + \sqrt{\theta a(t)} G.$$

Objectifs

- Conjectures faciles pour l'arbre total (prendre $t = \alpha - 1$), mais ...
- Ces résultats $\Longrightarrow$ estimation de θ + intervalle de confiance quand α connu.
- Et ces résultats $\Longrightarrow$ estimation de α + intervalle de confiance quand θ connu (+ intéressant).