Variations sur le thème du coalescent

(Chaire MMB - 11/09/2014)

Modèles de | arbres généalogiques / généalogies assez courants en génétique des populations. Modèles "neutres", taille de pop fixée. Grande variété d'arbres mais qui apparaissent naturellement comme objets limites.

I Zoologie des coalescents

Pour comprendre d'où viennent ces objets, prenons d'abord un point de vue "reproduction". Le modèle suivant s'appelle le modèle de Cannings, c'est une généralisation du modèle de Wright-Fisher à population finie.

On suppose donc que la taille de la pop est fixée à N, que son évolution est neutre et que la population est panmictique.

Pour une génération donnée notons X_i le nombre (aléatoire) de descendants du $i^{ème}$ individu de cette génération.

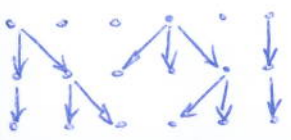
On suppose :

- $\sum X_i = N$ avec probabilité 1
- le vecteur $\underline{X} = (X_1, \dots, X_N)$ est "échangeable" au sens où si on permute ses coordonnées il garde la même loi (les indices ne jouent aucun rôle et n°1 ne va pas forcément avoir plus d'enfants que n°2) $\rightsquigarrow$ neutralité.

Déf : On appelle <u>modèle de Cannings</u> la règle de reproduction suivante : on part de N individus. A chaque génération k on tire une réalisation $\underline{x}^k$ indép. des précédentes du vecteur $\underline{X}$ et on octroie x_1^k enfants à l'individu 1, ..., x_N^k enfants à l'individu N pour former la génération $k+1$.

Rq : Certains individus peuvent avoir 0 enfants

$N=6$



$\underline{x}^1 = (2,0,0,3,0,1)$

$\underline{x}^2 = (1,2,0,0,2,1)$

$\underline{x}^3$ ...

Le modèle de W-F, dans lequel chaque individu choisit son parent unif. au hasard dans la génération précédente, correspond au vecteur $\underline{X} = \text{Mult}\left(N; \frac{1}{N}, \dots, \frac{1}{N}\right)$

$$\Rightarrow \mathbb{P}(X_1 = x_1, \dots, X_N = x_N) = \binom{N}{x_1 \dots x_N} \frac{1}{N^N}.$$

On n'a pas donné d'allèle aux individus, mais on pourrait suivre l'évolution de la diversité génétique. On s'intéresse plutôt aux arbres généalogiques générés.

Commençons par un échantillon de taille 2.

$$\mathbb{P}(\text{ancêtre commun à la gén. précédente}) = \sum_{i=1}^{N} \mathbb{E}\left[\frac{X_i}{N}\frac{(X_i-1)}{N-1}\right] = \frac{\mathbb{E}[X_1(X_1-1)]}{N-1} =: \ell^N_{2,2}$$

$\Rightarrow$ τ_2^N nbre de générations à remonter pour trouver un ancêtre commun

$\hookrightarrow$ $\mathbb{P}(\tau_2^N = k) = (1-\ell^N_{2,2})^{k-1}\,\ell^N_{2,2}$ et $\tau_2^N \sim \text{Geom}(\ell^N_{2,2})$.

En particulier, $\mathbb{E}[\tau_2^N] = \frac{1}{\ell^N_{2,2}}$.

Wright-Fisher : $\ell^N_{2,2} = \frac{1}{N}$, donc on a envie de regarder des échelles de temps d'ordre N pour voir quelque chose.

Quand N est très grand, si $\ell^N_{2,2} \xrightarrow[N\to\infty]{} 0$ alors $\frac{1}{\ell^N_{2,2}}$

$$\mathbb{P}\left(\tau_2^N > t\left(\ell^N_{2,2}\right)^{-1}\right) = \underbrace{(1-\ell^N_{2,2})^{\lfloor t(\ell^N_{2,2})^{-1}\rfloor}}_{(1-\frac{1}{N})^{\lfloor tN\rfloor}} \approx e^{-t} \quad \text{et}$$

Faire pour WF d'abord puis un parallèle.

WF :

$$\frac{\tau_2^N}{1/\ell^N_{2,2}} \xrightarrow[N\to\infty]{\text{(loi)}} \text{Exp}(1) =: T_2$$

Donc $1/\ell^N_{2,2}$ donne l'échelle de temps sur laquelle il se passe qqch.

($= N$ pour W-F)

$\ell^N_{n;k_1,\dots,k_r}$ = proba que $\{1,\dots,k_1\}$ ont un parent, …, $\{k_1+\dots+k_{r-1}+1,\dots,k_1+\dots+k_r\}$ aussi.

Puisqu'il faut au moins 2 individus qui coalescent pour voir une plus grosse fusion, on a $\ell^N_{2,2} \geq \ell^N_{n;k_1,\dots,k_r}$ $\forall n, k_1,\dots,k_r$

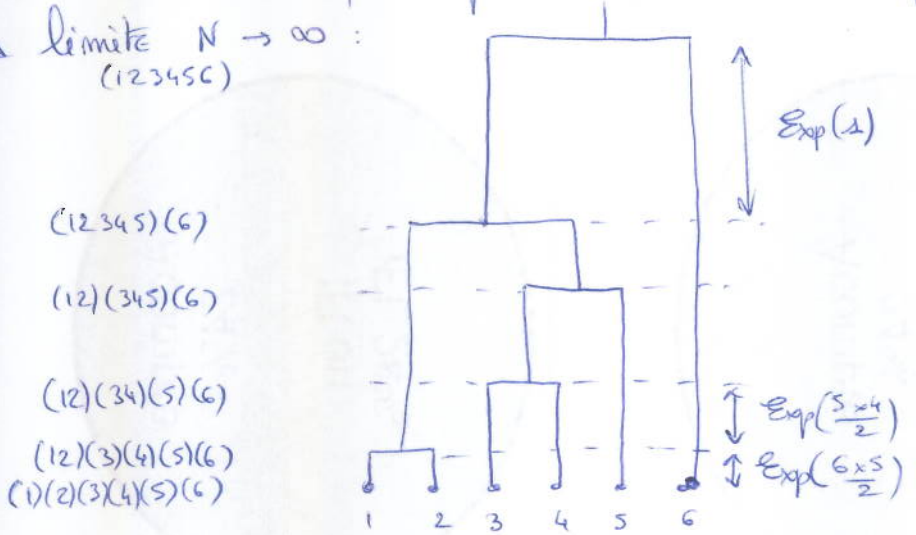
et on veut comparer les autres types de fusion avec celle de 2 individus.

On a $\ell^N_{n;k_1,\dots,k_r} = \underbrace{N(N-1)\dots(N-r+1)}_{\text{nombre de choix de parents}} \underbrace{\mathbb{E}\left[\frac{X_1(X_1-1)\dots(X_1-k_1+1)\quad\dots\quad X_r(X_r-1)\dots(X_r-k_r+1)}{N(N-1)\dots(N-n+1)}\right]}_{\text{proba de l'allocation de parents/enfants}}$

Idée : si $\ell^N_{n;k_1,\dots,k_r} \ll \ell^N_{2,2}$ si l'un des $k_i \geq 3$ ou si au moins 2 k_i sont ≥ 2, alors on ne voit que des fusions de 2 lignées lorsque $N \to \infty$.

A ce moment, on peut montrer que lorsqu'il y a k lignées actives, le nombre de générations à remonter pour voir un 1er événement, divisé par $(1/\ell^N_{2,2})$ converge en loi vers une loi exponentielle de par. $\frac{k(k-1)}{2}$ (le nombre de paires de lignées $\neq$) et la coalescence qui se produit est entre 2 lignées choisies au hasard.

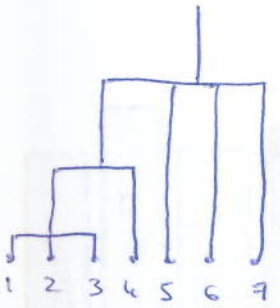
A la limite $N \to \infty$:



C'est le coalescent de Kingman

Si $\frac{\ell^N_{n;k_1,1,\dots,1}}{\ell^N_{2;2}} \xrightarrow[N\to\infty]{} \lambda_{n;k_1} > 0$ mais $\frac{\ell^N_{n;k_1,k_2,\dots,k_r}}{\ell^N_{2,2}} \to 0$ dès que

≥ 2 k_i sont ≥ 2, on obtient un coalescent à collisions multiples



plus fort comme lien avec les reproductions

là un individu peut avoir une grosse famille, mais pas 2 en même temps :

$\ell^N_{2;2} = N\mathbb{E}\left[\frac{X(X-1)}{N(N-1)}\right]$

$\ell^N_{3;3} = N\,\mathbb{E}\left[\frac{X_1(X_1-1)(X_1-2)}{N(N-1)(N-2)}\right]$ si $X_1 = \Theta(1)$ $\sim \frac{1}{N^2}$

$\ell^N_{4;2,2} = N(N-1)\,\mathbb{E}\left[\frac{X_1(X_1-1)X_2(X_2-1)}{N(N-1)(N-2)(N-3)}\right] \sim \frac{1}{N^2}$

(ou Λ-coalescent car les $\lambda_{n;k}$ ont pour forme $\int_0^1 u^k(1-u)^{n-k}\frac{\Lambda(du)}{u^2}$)

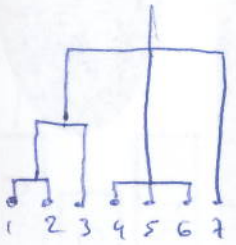
Si $\frac{\ell^N_{n;k_1,\dots,k_r}}{\ell^N_{2,2}} \to \lambda_{n;k_1,\dots,k_r} > 0$ $\forall k_1,\dots,k_r, n$, on obtient un coalescent à fusions multiples et simultanées.

Ex n°2 : Si 1 individu a $\Theta(N)$ descendants et les autres $\Theta(1)$

$\ell^N_{2;2} \propto \frac{1}{N^2}N^2 = 1$

$\ell^N_{3,3} \propto \frac{N^3}{N^3} = 1$

$\ell^N_{4;2,2} = 0$ ou $\Theta(\frac{1}{N})$ si on est un peu flexible



(ou Ξ-coalescent parce que la paramétrisation des fusions dépend d'une mesure Ξ).

On peut décrire précisément la loi des objets limites, mais l'important est de garder la philosophie de leur construction en tête.

II Comment prendre de bonnes résolutions (avec R. Sainudiin et T. Stadler)

Le § précédent n'était pas très formel, mais une manière d'encoder les coalescents est de les considérer comme des processus à valeurs dans les partitions de $\{1,\dots,n\}$, où n est la taille de l'échantillon.

(figure).

On se restreint maintenant au coalescent de Kingman. La description la plus précise de ce modèle de généalogie est celle à base de partitions, puisqu'on décrit qui partage les mêmes ancêtres à chaque instant. Mais on pourrait aussi compter :

→ le nombre de lignées ancestrales $\neq$:

$n \to n-1$ après un temps $\mathrm{Exp}\left(\frac{n(n-1)}{2}\right)$

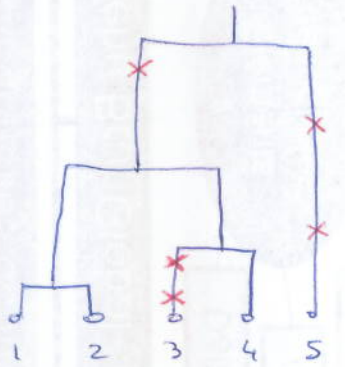
$n-1 \to n-2$ $\mathrm{Exp}\left(\frac{(n-1)(n-2)}{2}\right)$

$\vdots$

$2 \to 1$ $\mathrm{Exp}(1)$

→ le nombre de lignées en vie qui sont ancêtres de $1, 2, \dots, n$ individus de l'échantillon. → vecteur $(f_1(t), \dots, f_n(t))$ qui part de $(n, 0, \dots, 0)$ et se fixe en $(0, \dots, 0, 1)$

Ex :



$(0,0,0,0,1)$

$(1,0,0,1,0)$

$(1,2,0,0,0)$
$(3,1,0,0,0)$
$(5,0,\dots,0)$

D'autres "résolutions" sont possibles. L'intérêt du F-processus apparaît quand on ajoute des mutations sur l'arbre. En effet, ce qu'on observe réellement n'est pas l'arbre généalogique, mais qui porte les mêmes mutations. On note S_i le nombre de mutations portées par i individus dans l'échantillon. Le vecteur

$$S = (S_1, \dots, S_{n-1})$$ est appelé spectre des fréquences de mutations/site

et c'est quasiment l'info la plus précise qu'on peut avoir.

Mais S_i ne dépend que de la longueur totale de branches qui sous-tendent i individus de l'échantillon, pas de qui précisément sont ces individus.

On peut donc travailler directement à la résolution des F-processus, notamment dans les calculs de vraisemblance ou pour utiliser des méthodes ABC ou échantillonnage d'importance.

Plusieurs gains :

* simuler un F-processus est moins coûteux que simuler le coalescent complet.
* les probas qui apparaissent dans les calculs sont des fonctions plus directes du F-processus que du coalescent-partition.

Ex : On suppose que les mutations apparaissent à taux $\mu > 0$ sur l'arbre. On cherche à estimer μ à partir du SFS.

La vraisemblance de $(S_1, \dots, S_{n-1})$ s'écrit

$$\mathcal{L}(\mu) = \mathbb{E}\left[\prod_{i=1}^{n-1} \left(e^{-\mu L_i} \frac{(\mu L_i)^{S_i}}{S_i!} \right)\right]$$

L_i est une fonction simple du processus F, tandis qu'il faut commencer par vérifier la taille de chaque bloc dans le cas du coalescent.

Ensuite, l'espérance n'est pas simple à calculer directement. On utilise alors

$$\mathcal{L}(\mu) = \mathbb{P}_\mu(S = s) = \sum_{f \sim s} \mathbb{P}_\mu(S = s \mid F = f)\, \mathbb{P}(F = f)$$

(optimal, mais a priori marche sur la somme complète)

$$\approx \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}_\mu(S = s \mid F = f_i)$$ où les f_i sont iid de loi Kingman

→ soit la taille de l'échantillon est suffisamment petite et on peut calculer la 1ère somme en considérant le f de manière exhaustive (exclus pour le coal de Kingman)

→ soit on produit un grand nombre de f_i (moins coûteux que produire des arbres complets) et la fonction $\mathbb{P}_\mu(S = s \mid F = f_i)$ est une fonction + simple de f_i que de l'arbre.

Dernier avantage :

→ on peut vouloir mettre en place un algorithme contrôlé qui ne produit des Kingmans compatibles avec le spectre observé (cf. Sainudiin & al 2011) → Plus simple avec une résolution du coalescent adaptée.

A retenir : réfléchir à la bonne résolution à laquelle on doit travailler peut permettre d'économiser plein d'énergie !!