Variations autour de l'inférence de la taille de population effective à partir de données génétiques

Aussois, juin 2021

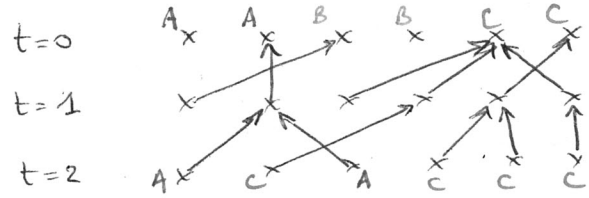
I Motivations

On va chercher à reconstruire l'histoire démographique d'une population, ou du moins des paramètres qui en décrivent les éléments importants dans la perspective de comprendre la diversité génétique actuelle de la population, et ce à partir de données génétiques. Dans la plupart des méthodes utilisées pour ceci (méthodes de vraisemblance), on verra qu'il est nécessaire de décomposer la vraisemblance suivant la généalogie cachée derrière l'échantillon de données dont on dispose et se posera alors la question du niveau d'information optimal sur ces généalogies qui soit nécessaire pour caractériser la loi des données de mutation dont on dispose. On verra plusieurs approches qui utiliseront plusieurs types de données génétiques (travaux développés avec Raazesh Sainudiin - Univ. d'Uppsala et Julia Palacios & Lorenzo Cappello - Stanford), dans le cadre d'une population bien mélangée et en utilisant la diversité génétique à un allèle neutre. On verra au passage quelques applications en exemples.

Plan :
- II Rappels sur le coal. de Kingman avec taille de pop. fluctuante.
- III Méthodes d'inférence de $(N_e(t))_{t>0}$
- IV Un échantillonneur préférentiel d'arbres compatibles avec un SFS donné
- V Inférence basée sur les arbres de Tajima.

II "Rappels" sur le coalescent de Kingman avec taille de population fluctuante

Un modèle vraiment canonique en génétique des populations est le modèle de Wright-Fisher, dans lequel on suppose que la population est bien mélangée, de taille fixée N au cours des générations (supposées discrètes), que les individus sont haploïdes et qu'on s'intéresse à un gène/locus donné dont on suppose que les allèles sont neutres au sens où aucun allèle n'avantage ses porteurs. Transmission génétique : chaque individu à la génération $t+1$ choisit son parent uniformément au hasard à la génération t, indépendamment des autres individus de sa génération et hérite de son allèle :



Supposons que la pop. ait évolué ainsi depuis longtemps et qu'au temps qu'on appelle le présent, on échantillonne 2 individus ($\neq$) uniformément au hasard dans la population actuelle. Notons T_2^N le nombre de générations à remonter pour trouver le 1er ancêtre commun à ces 2 individus.

$$\mathbb{P}(T_2^N = 1) = \mathbb{P}(\text{même parent}) = \sum_{i=1}^{N} \mathbb{P}(\text{même parent, n° } i)$$
$$= \sum_{i=1}^{N} \frac{1}{N^2} = \frac{1}{N}.$$

Puis $\mathbb{P}(T_2^N = k) = \underbrace{\left(1-\frac{1}{N}\right)^{k-1}}_{\substack{k-1 \text{ générations} \\ \text{où les 2 lignées} \\ \text{restent disjointes}}} \times \underbrace{\frac{1}{N}}_{\substack{\text{se rejoignent} \\ \text{au } k^{\text{e}} \text{ essai}}} \quad \forall k \geq 1$,

i.e. $T_2^N \sim \text{Géom}\left(\frac{1}{N}\right)$ ($\Rightarrow \mathbb{E}[T_2^N] = N$, la taille de la population)

Que se passe-t-il quand $N \to +\infty$? Pour tout $t \in \mathbb{R}_+$, on a

$$\mathbb{P}(T_2^N > Nt) = \left(1-\frac{1}{N}\right)^{\lfloor Nt \rfloor} = e^{\lfloor Nt \rfloor \ln\left(1-\frac{1}{N}\right)} = e^{-\frac{1}{N}\lfloor Nt \rfloor + o\left(\frac{1}{N}\right)} \xrightarrow[N\to\infty]{} e^{-t}.$$

Conclusion : la variable aléatoire qui donne combien de "paquets de N" générations on doit remonter pour trouver le 1er ancêtre commun à nos 2 individus choisis uniformément au hasard converge en loi vers une loi Exponentielle (1).

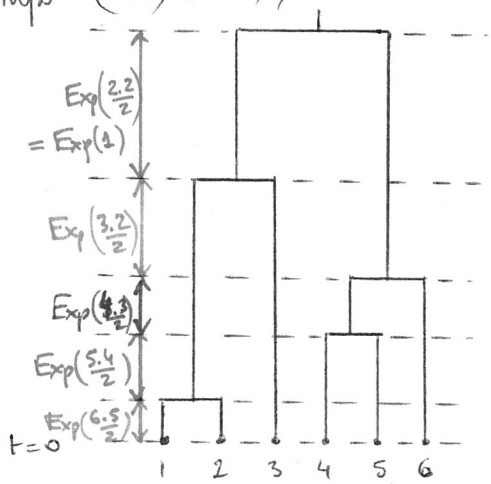
De la même manière, si on prend un échantillon de n individus $\neq$, unif. au hasard dans la population présente, et si on note T_n^N le nombre de générations à remonter avant qu'au moins 2 lignées ancestrales issues de l'échantillon trouve un ancêtre commun, alors $\frac{T_n^N}{N} \xrightarrow[N\to\infty]{\text{(loi)}} \text{Exp}\left(\frac{n(n-1)}{2}\right)$

où $\frac{n(n-1)}{2}$ est le nombre de paires de lignées $\neq$ dans l'échantillon, et que lorsque cette fusion de lignées ancestrales se produit, alors avec une probabilité qui tend vers 1 lorsque $N \to \infty$, seules 2 lignées ancestrales fusionnent en une lignée commune et ces 2 lignées sont choisies unif. au hasard.
(coalescentes)

(pour comprendre pourquoi, on vérifiera que

$$\mathbb{P}\left(\overset{1\text{-}2}{\uparrow} \uparrow \uparrow \uparrow \uparrow \overset{n}{\uparrow} \;\middle|\; T_n^N = 1 \right) \xrightarrow[N\to\infty]{} \frac{2}{n(n-1)} \quad)$$

Au final, si on regarde l'arbre généalogique de l'échantillon sur l'échelle de temps $(Nt, t \geq 0)$, on obtient donc lorsque $N \to +\infty$:



{1,2,3,4,5,6}

{1,2,3}{4,5,6}

{1,2}{3}{4,5,6}

{1,2}{3}{4,5}{6}

{1,2}{3},{4},{5},{6}
{1}{2}{3}{4}{5}{6}

Version arbre — version partitions

On encode cette dynamique par un processus à valeurs dans Π_n, l'ensemble des partitions de $\{1, 2, \dots, n\}$.

Formellement, le <u>coalescent de Kingman</u> pour un échantillon de taille n est le processus markovien à valeurs dans Π_n, de valeur initiale $\{\{1\},\{2\},\dots\{n\}\}$ $(\hat{C}_t)_{t\geq 0}$ et tel que si $|C_t| = k \in \{2,\dots,n\}$, alors à taux $\frac{k(k-1)}{2}$ (i.e. après un temps exponentiel de paramètre $\frac{k(k-1)}{2}$) une paire de blocs fusionne, cette paire étant choisie uniformément au hasard parmi les $\frac{k(k-1)}{2}$ paires possibles.

(Autre formulation possible : chaque paire de blocs fusionne à taux 1 indépendamment des autres).

On notera $\pi \prec \pi'$ si π peut être obtenue à partir de π' en fusionnant 2 blocs. Ex : $\{\{1,2\},\{3\}\} \prec \{\{1\},\{2\},\{3\}\}$. (Inutile dans la suite)

C'est l'objet de base avec lequel on va travailler pour un moment, sous l'hypothèse donc que la taille de population est très grande (infinie, en théorie).

Reprenons les calculs et voyons ce qu'il se passe lorsque la taille de la population s'écrit non plus N, mais $\lfloor N g(t) \rfloor$ où g est une fonction du temps remonté depuis le présent. Taille d'échantillon = 2 :

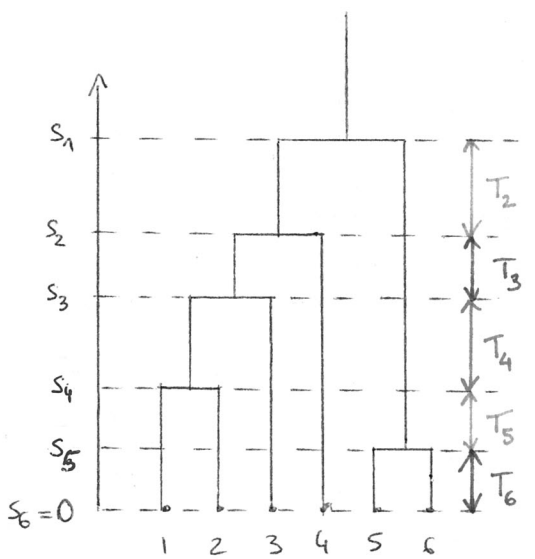
Cette fois on a
$$\mathbb{P}(T_2^N > Nt) = \left(1-\frac{1}{\lfloor Ng(0)\rfloor}\right)\left(1-\frac{1}{\lfloor Ng(1)\rfloor}\right)\cdots\left(1-\frac{1}{\lfloor Ng(\lfloor Nt\rfloor)\rfloor}\right)$$
$$= \exp\left(\sum_{k=0}^{\lfloor Nt\rfloor} \ln\left(1-\frac{1}{\lfloor Ng(k)\rfloor}\right)\right)$$
(en supposant $g>0$ bornée inférieurement par exemple)
$$= \exp\left(-\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor}\frac{1}{g(k)} + O\left(\frac{1}{N}\right)\right)$$

et si g est telle qu'il existe $(N_e(s))_{s\geq 0}$ telle que $\forall t \geq 0$,
$$\frac{1}{N}\sum_{k=0}^{\lfloor Nt\rfloor}\frac{1}{g(k)} \xrightarrow[N\to\infty]{} \int_0^t \frac{1}{N_e(s)}\,ds, \text{ alors}$$
(en gros $g(t) = G\left(\frac{t}{N}\right)$ doit varier très lentement)
$$\mathbb{P}\left(\frac{T_2^N}{N} > t\right) \xrightarrow[N\to\infty]{} e^{-\int_0^t \frac{1}{N_e(s)}ds}.$$

A la limite, on n'obtient plus un temps exponentiel "classique" mais une horloge

qui sonne au taux instantané $\frac{1}{N_e(t)}$ (attention qu'il y a eu un changement d'échelle de temps avant le passage à la limite).

Dans le cas où g satisfait $\forall t$: $\liminf_{N\to\infty} \inf_{k\in[\![0,Nt]\!]} g(k) > 0$, on peut montrer que la généalogie limite d'un échantillon de taille n (lorsque $N\to\infty$) est également un processus markovien en temps continu à valeurs dans Π_n où les fusions de blocs sont binaires, mais chaque paire de blocs fusionne maintenant au taux instantané $\frac{1}{N_e(t)}$, de sorte que l'image à la limite soit



où $\mathbb{P}(T_k > t \mid T_6, T_5, \dots, T_{k+1})$

$$= \exp\left(-\int_{T_6+\dots+T_{k+1}}^{T_6+\dots+T_{k+1}+t} \frac{ds}{N_e(s)}\right)$$

$\forall k \in \{2,\dots,6\}$.

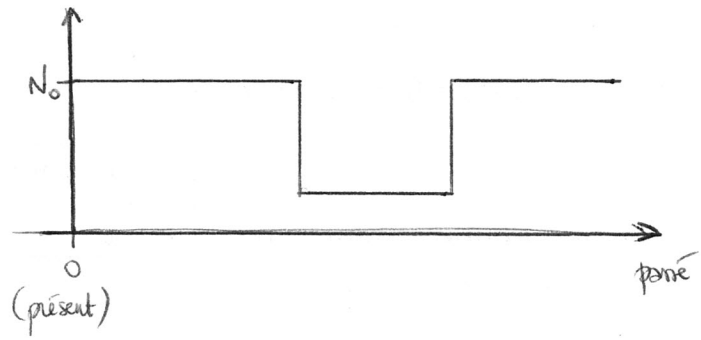
La trajectoire $(N_e(s))_{s\geq 0}$ est appelée <u>trajectoire de taille de population efficace ou effective</u> dans les applications (notamment lorsqu'on s'intéresse à son inférence) car il ne s'agit pas nécessairement des variations démographiques mais elle peut servir à modéliser de faibles déviations de la neutralité (panmixie, etc) qui ont conduit à la diminution transitoire du nombre d'individus ayant réellement contribué à la diversité génétique de la population.

" effectif / efficace " = parmi la (sous-) population des individus ayant pu contribuer réellement au pool de reproducteurs de leur génération.

<u>Exemples</u> : " croissance " exponentielle (croissance du passé vers le présent, donc décroissance quand on remonte le temps pour les généalogies)

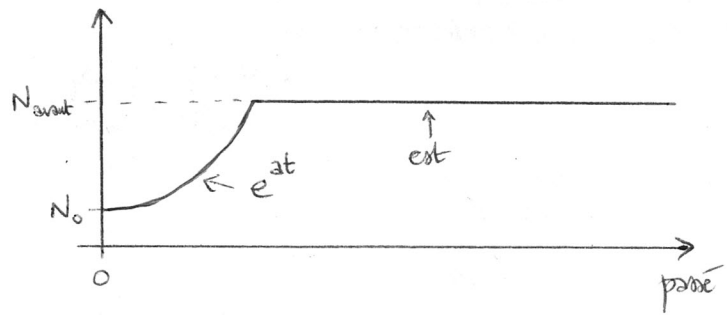
$N_e(t) = N_0 e^{-at}$ ⟶ coalescences de + en + rapides qd on remonte dans le passé

"bottleneck"



(ou sans le rétablissement à la fin)

"déclin"

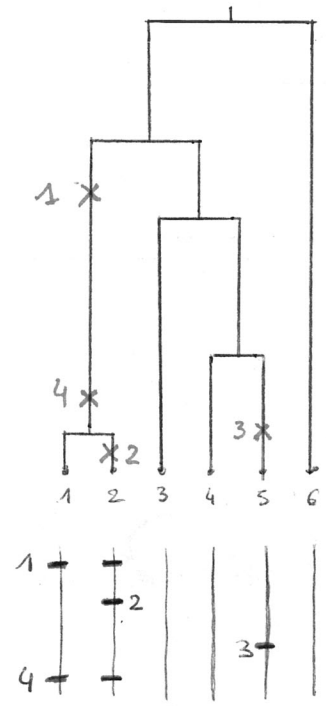


Dernier point modélisation

On veut inclure des mutations au gène d'intérêt et pour ça on va considérer que celui-ci correspond à une très longue région du génome ($\sim 10^5$ bp ou plus) et que chaque nouvelle mutation qui apparaît sur une paire de bases différente des précédentes → modèle à infinité de sites de mutation. C'est une hypothèse assez restrictive pour les jeux de données modernes mais il faut bien commencer quelque part (et pour certaines applications ce type de données est fréquemment utilisé).

Plus précisément, on suppose que les généalogies et mutations peuvent être co-construits de la manière suivante : étant donnés le taux de mutation μ et la trajectoire de taille de population effective $(N_e(t), t \geq 0)$, chaque lignée ancestrale mute à taux μ et chaque paire de lignées fusionne à un taux instantané $\frac{1}{N_e(t)}$, le tout indépendamment des autres. Chaque mutation touche un site nouveau de l'ADN et donc toutes les mutations sont visibles (aucune n'est effacée par une mutation plus proche du présent).

Exemple :



Généalogie + mutations

Alignement de séquences

Binary incidence matrix (BIM) correspondante =

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

site 1 | site 2 | site 3 | site 4 ↙ présence/absence de mutation

séq 1
séq 2

Spectre des fréquences de sites : pour $i \in \{1, \dots, n-1\}$

S_i = # sites ségrégeants où la mutation est portée par i individus

$\underline{S} = (2, 2, 0, 0, 0)$

↑ mut. 2,3 ↑ mut 1,4

Nombre de sites ségrégeants / mutations (équivalent sous ce modèle) = 4

⚠ BIM et $\underline{S}$ supposent de savoir quelle est la mutation et quelle est la base originelle, ce qu'on ne sait pas toujours.

III Méthodes d'inférence de $(N_e(t))_{t \geq 0}$.

Supposons qu'on dispose de données génétiques de type alignement de séquences ou SFS et que l'on veut les utiliser pour reconstruire la trajectoire de taille de pop efficace. De nombreuses approches existent déjà, on va en citer quelques unes avant de décrire la philosophie qu'on suivra ensuite :

1) L'approche Poisson Random Field : on considère une série de SNP qu'on suppose indépendants et on suppose que le taux de mutation est tellement bas qu'on peut approcher le nombre de mutations portées par k individus de l'échantillon par une loi de Poisson de paramètre la proportion moyenne de la longueur de l'arbre formée d'arêtes sous-tendant k individus de

l'échantillon. Cette dernière peut être estimée par simulation et on utilise le produit sur toutes les valeurs de k possibles pour obtenir une vraisemblance approchée pour $(N(t))_{t\geq 0}$. Ref : Gutenkunst et al 2009, Nielsen 2000, Sawyer & Hartl 1992

2) Méthodes basées sur les fonctions génératrices multivariées des longueurs de branches de la généalogie, mais impossible à utiliser au-delà de $n = 5$.
Ref : Bunnefeld et al 2015

3) Skyline plots : méthodes non paramétriques que je n'ai jamais vraiment comprises qui supposent que le taux de mutation est suffisamment élevé par qu'observer le spectre de mutations laisse peu de variabilité possible par l'arbre généalogique caché derrière elles. Mais ce ne sera pas le cas pour nous donc on ne donnera pas plus de détails. Ref : Ho & Shapiro 2011 (review)

4) Pour utiliser des séquences très longues sans avoir à supposer qu'elles ne recombinent pas, le "sequential Markov coalescent" a été introduit dans Marjoram & Wall 2006, McVean & Cardin 2005. Plusieurs raffinements existent mais l'idée est qu'on scanne le génome en supposant que les arbres généalogiques correspondant à chaque site changent de temps en temps à cause d'une recombinaison. La fonction de vraisemblance intègre donc la suite des généalogies et la liste des points de recombinaison le long du génome.

5) Méthodes Approximate Bayesian Computation (ABC) : on simule un très grand nombre d'arbres avec mutations pour chaque jeu de paramètres dans l'espace des paramètres d'intérêt et on calcule une vraisemblance approchée en ne gardant que les réalisations qui sont proches (pour une métrique bien choisie) des données. Jean-Michel Marin est l'un des fondateurs de ces méthodes et des refs récentes : Baitard et al 2016, Peter et al. 2010.

Ces méthodes sont très exigeantes du point de vue computationnel car elles se basent sur de grands nombres de simulations, notamment pour explorer l'espace des arbres généalogiques cachés derrière les données mutationnelles. Pour ces méthodes, les arbres en question sont modélisés par le coalescent de Kingman (avec taille de population variable qu'on souhaite reconstruire).

Pourquoi passer par une variable cachée ? Parce qu'on sait donner la loi des mutations conditionnelle à l'arbre sous-jacent mais pas la loi des données mutationnelles dans un modèle démographique sans passer par les arbres.

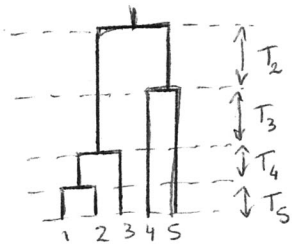
Supposons qu'on a un échantillon de taille n avec des données génétiques qu'on note D_n (à un locus non-recombinant). Soit Φ l'ensemble qui indexe une famille de modèles démographiques (pas nécessairement paramétrique). Alors $\forall \varphi \in \Phi$,

$$\mathbb{P}_\varphi(D_n) = \int_{(A,T) \in \mathcal{A}_n \times \mathcal{T}_n} \mathbb{P}_\varphi(D_n \mid (A,T)) \, d\mathbb{P}_\varphi((A,T)) \qquad (*)$$

A : topologie d'arbre étiquetée ; T : vecteur de temps inter-coal.

Rg : manière pas très rigoureuse de l'écrire $\mathbb{P}_\varphi(D_n) = \sum_{(A,T)} \mathbb{P}_\varphi(D_n \mid (A,T)) \, \mathbb{P}_\varphi((A,T))$.

Rappel :



époque j = période pendant laquelle l'échantillon a j ancêtres $\neq$.

Connaître chaque terme de $(*)$ ne veut pas dire qu'on sait calculer la somme. Calculer $\mathbb{P}_\varphi(D_n)$ requiert d'explorer l'espace $\mathcal{A}_n$. Comme la taille de l'espace $\mathcal{A}_n$ augmente sur-exponentiellement, même pour $n = 10$ on ne peut pas faire ce calcul et on approche $\mathbb{P}_\varphi(D_n)$ en explorant $\mathcal{A}_n$ par des méthodes MCMC ou d'échantillonnage préférentiel.

Pour info, $|\mathcal{A}_n| = \dfrac{n!\,(n-1)!}{2^{n-1}}$

$n = 5$	$n = 10$	$n = 20$
$\lvert\mathcal{A}_5\rvert = 180$	$2{,}5 \cdot 10^9$	$5{,}64 \cdot 10^{29}$

Analysons ce qu'il se passe lorsque D_n = #mutations / sites ségrégeants.
Une description équivalente de la manière dont les mutations apparaissent sur l'arbre est de dire que conditionnellement à la réalisation de l'arbre généalogique, les mutations tombent sur l'arbre suivant un processus ponctuel de Poisson (équivalente à la manière dont on a introduit les mutations plus tôt).

En particulier, conditionnellement à la longueur L_n de l'arbre, on a

$$D_n \sim \text{Poisson}(\mu L_n),$$

i.e. la loi de D_n ne dépend que de $L_n = \sum_{j=2}^{n} j T_j$ et donc il est inutile de scanner un immense espace de topologies étiquetées. Il suffit dans ce cas de simuler le vecteur $(T_2, T_3, \dots, T_n)$ sous $\mathbb{P}_\varphi$ pour calculer ou approcher $\mathbb{P}_\varphi(D_n)$.

Dans la suite, on va se demander quel niveau d'information sur l'arbre généalogique, autrement dit quel niveau de résolution de l'arbre est optimal pour le calcul (même approché) de $\mathbb{P}_\varphi(D_n)$ en fonction de la nature des données D_n. On va le voir d'abord lorsque D_n est le spectre des fréquences de sites, puis lorsque D_n est un alignement de séquences (et pour finir lorsqu'il s'agit d'alignements de séquences échantillonnés à différents temps mais en réalité on n'aura pas le temps).

La suite émane d'une idée de Raazesh Sainudiin qu'on a développée ensemble puis avec d'autres.

IV Un échantilleonneur préférentiel d'arbres compatibles avec un SFS donné.

Travaux avec R. Sainudiin (Univ. d'Uppsala)

Supposons que D_n correspond ici au spectre des fréquences de sites :

$$D_n = \underline{S}_n = (S_1, S_2, \dots, S_{n-1}) \quad \text{où } S_i = \#\text{ mutations portées par } i \text{ individus de l'échantillon.}$$

à un locus non recombinant.

Dans ce paragraphe, on ne va pas supposer la panmixie ou la neutralité de l'évolution, seulement que les généalogies peuvent être modélisées par des arbres binaires (pas de fusions multiples / simultanées) $\rightsquigarrow$ la loi de l'arbre généalogique n'est pas forcément le coal. de Kingman.

Si $L_n^{(j)}$ est la longueur totale des branches de l'arbre qui sous-tendent j individus de l'échantillon, alors conditionnellement à $L_n^{(j)}$, on a

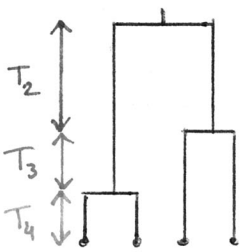
$S_j \sim \text{Poisson}(\mu L_n^{(j)})$ et cond à $(L_n^{(2)}, \dots, L_n^{(n-1)})$, les S_j sont indépendantes.

En particulier, on n'a pas vraiment besoin de savoir quelle lignée est étiquetée par quel nombre, simplement combien d'arêtes à chaque étage de l'arbre sous-tendent j individus (ainsi que la longueur de ces arêtes).

On va donc utiliser la matrice $(F_{ij})_{\substack{2\le i\le n \\ 1\le j\le n-1}}$ définie par

$F_{ij} = \#\{\text{branches arêtes/ancêtres à l'époque } i \text{ qui sous-tendent } j \text{ individus de l'échantillon}\}$.

Ex:



$n=4$

$$F = \begin{pmatrix} 0 & 2 & 0 \\ 2 & 1 & 0 \\ 4 & 0 & 0 \end{pmatrix} \qquad T = (T_2, T_3, T_4)$$

On a alors $L_n^{(j)} = \sum_{i=2}^{n} T_i F_{ij} = (TF)_j \quad \forall j \in \{1, \dots, n-1\}$ et donc la paire (F,T) est nécessaire et suffisante pour caractériser la loi de S_n.

F vit dans l'espace des "tree shapes" (traduction ?...), en bijection avec l'ensemble des matrices $(n-1)\times(n-1)$ telles que $F_{ij} \in \mathbb{N} \ \forall i,j$,

$\sum_{j=1}^{n-1} F_{ij} = i \quad \forall i \in \{2, \dots, n\}$ (attention à la numérotation peu canonique)

et $\sum_{j=1}^{n-1} j F_{ij} = n$.

A chaque "tree shape" sont associées $\frac{n!}{2^{n-1-b(F)}} \prod_{i=2}^{n} j_i!$ topologies étiquetées où $b(F) = \sum_{i=2}^{n} \mathbb{1}_{\{0\}}(\max\{F_{i,\cdot} - F_{i-1,\cdot}\})$ et j_i est le nombre de blocs ayant la même taille que celui qui est créé au début de l'époque i.

$n=10$: $\frac{n!}{2^{n-1}} = 7100$, $n=20 \to 4{,}6 \cdot 10^{12}$, mais attention que le facteur combinatoire supplémentaire peut être très grand aussi.

Mathématiquement, pour passer des topologies étiquetées aux "tree shapes" on définit une relation d'équivalence sur les topologies étiquetées telle que 2 topologies étiquetées sont équivalentes si elles conduisent aux mêmes nombres F_{ij} lorsqu'on retire les étiquettes.

On va utiliser cette résolution plus grossière des arbres binaires étiquetés pour mettre au point un échantillonneur préférentiel de généalogies qui sont toujours compatibles avec un SFS donné. Ceci permettra de ne pas échantillonner des topologies pour lesquelles la probabilité des données observées est nulle et qui donc n'apportent rien au calcul de $\mathbb{P}_\theta(D_n)$.

On va utiliser une famille de lois sur les tree shapes plus large que le coalescent de Kingman et différents modèles démographiques et on souhaite inférer ces 2 composantes (la loi des topologies des arbres généalogiques sous-jacents et le scénario démographique).

On utilise le modèle Beta-splitting d'Aldous, qui a un seul paramètre $\beta > -2$ qui permet de régler l'équilibre de l'arbre : dans ce modèle,

- $\beta \to -2$ donne l'arbre en peigne 
- $\beta = 0$ donne la topologie du coalescent de Kingman
- $\beta \to +\infty$ donne des arbres de plus en plus équilibrés

⚠ Le modèle d'Aldous donne des cladogrammes, dans lesquels l'ordre des coalescences n'est pas enregistré (ex : ) donc il faudra ensuite donner un ordre aux événements. On construit ces arbres de la racine vers les feuilles.

Si une arête sous-tend b individus, la probabilité que cette arête se "sépare" en une arête sous-tendant $x \geq \frac{b}{2}$ feuilles et une arête sous-tendant $b-x$ feuilles vaut

$$\lambda_{b,x} = 2\, a_b^{-1} \binom{b}{x} \int_0^1 u^{x+\beta}(1-u)^{b-x+\beta}\, du \quad \text{si } x > \frac{b}{2}$$

$$= a_b^{-1} \binom{b}{x} \int_0^1 u^{x+\beta}(1-u)^{b-x+\beta}\, du \quad \text{si } x = \frac{b}{2} \quad (b \text{ pair})$$

où $a_b = \int_0^1 (1-u^b-(1-u)^b)\, u^\beta (1-u)^\beta\, du$.

En particulier, pour $\beta = 0$ on obtient $\lambda_{b,x}^{\text{Kingman}} = \dfrac{2 - \mathbb{1}_{\{x=b/2\}}}{b-1}$.

On obtient ainsi une famille à 1 paramètre telle que pour $\beta < 0$ les arbres sont plutôt déséquilibrés (ce qu'on attend avec de la sélection naturelle directionnelle par ex.)

alors que $\beta \gg 1$ donne des arbres plus équilibrés (par exemple ce qu'il se passe si on choisit bien son plan d'échantillonnage dans des populations avec structure spatiale).

Rappel : principe de base de l'échantillonnage préférentiel (importance sampling).

On suppose qu'on veut estimer une quantité de la forme $\mathbb{E}[f(X)]$ pour une variable X qu'on ne sait pas simuler mais dont on connaît la loi. On supposera pour simplifier que cette loi est à densité par rapport à la mesure de Lebesgue ou discrète, de sorte qu'on peut définir une fonction $p(x)$ qui soit la densité de X au point $x \in \mathbb{R}^d$ ou la probabilité de l'événement $X = x$ dans le cas discret. On note E l'espace d'états de X.

Supposons que l'on sait simuler une variable $\tilde{X}$, de densité ou proba $q(x)$ (définie sur le même espace que X). Alors naturellement on a envie d'écrire que

$$\begin{aligned}\mathbb{E}[f(X)] &= \int_E f(x)\, p(x)\, dx \\ &= \int_E f(x) \frac{p(x)}{q(x)}\, q(x)\, dx \\ &= \mathbb{E}\left[f(\tilde{X}) \frac{p(\tilde{X})}{q(\tilde{X})} \right]\end{aligned}$$

version à densité

$$\begin{aligned}\mathbb{E}[f(X)] &= \sum_{x \in E} f(x)\, p(x) \\ &= \sum_{x \in E} \left[f(x) \frac{p(x)}{q(x)} \right] q(x) \\ &= \mathbb{E}\left[f(\tilde{X}) \frac{p(\tilde{X})}{q(\tilde{X})} \right]\end{aligned}$$

Pour donner un sens à ceci il faut que

$$\mathbb{E}[|f(X)|] < \infty, \quad \mathbb{E}\left[|f(\tilde{X})| \frac{p(\tilde{X})}{q(\tilde{X})} \right] < \infty$$

et en particulier il faut que

$q(x) = 0 \Rightarrow p(x) = 0$ (soit $p \ll q$) pour que le ratio $\frac{p(\tilde{X})}{q(\tilde{X})}$ ait un sens.

p absolument continue par rapport à q

En supposant ces conditions soient remplies, la loi des grands nombres nous dit alors que si $(\tilde{X}_1, \tilde{X}_2, \dots)$ est une suite de tirages i.i.d de même loi que $\tilde{X}$, alors

$$\frac{1}{N} \sum_{i=1}^{N} f(\tilde{X}_i) \frac{p(\tilde{X}_i)}{q(\tilde{X}_i)} \xrightarrow[N \to \infty]{} \mathbb{E}\left[f(\tilde{X}) \frac{p(\tilde{X})}{q(\tilde{X})} \right] = \mathbb{E}[f(X)].$$

Si en outre $\mathbb{E}\left[\left(f(\tilde{X}) \frac{p(\tilde{X})}{q(\tilde{X})} \right)^2 \right] < \infty$, on a un théorème de la limite centrale qui nous donne une vitesse de convergence en $\sqrt{N}$.

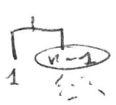
L'idée derrière cette technique est qu'on simule les variables que l'on veut (tant qu'on ne peut pas lister de valeurs pouvant être prises par X) mais ensuite il faut corriger le poids qu'on leur donne dans l'estimation de $\mathbb{E}[f(X)]$.

Revenons à l'échantillonneur de topologies + vecteur de temps d'intercoalescence qui soient compatibles avec le spectre des fréquences de sites observé.

- <u>Vecteur de temps</u> : L'échantillonneur suppose a priori que les T_j sont indépendants et suivent une loi exponentielle de paramètre A_j où $(A_2, A_3, \dots, A_n)$ sont des taux donnés en arguments à l'échantillonneur.

En gros, si on souhaite se rapprocher d'une simulation sous le scénario démographique $\mathbb{P}_\varphi$ (par $\varphi \in \Phi$), alors on prend A_j tel que $\frac{1}{A_j}$ soit l'espérance de T_j sous ce scénario (on peut l'obtenir facilement par simulation de la loi des T_j non-biaisée par le SFS qu'on veut expliquer).

Le reste de la philosophie est le suivant : on donne à l'échantillonneur les paramètres $\beta, \underline{A}, \mu, \underline{S}$ et il produit un triplet (F, M, T) accompagné d'un poids w. F et T sont les tree shape et vecteur de temps d'intercoalescence et M est une matrice qui encode la manière dont les mutations sont situées sur l'arbre. M_{ij} = # mutations portées par j individus de l'échantillon apparaissant sur une branche à l'époque i.

La construction itérative de F se fait en considérant le nombre de mutations portées par $n-1$ individus, puis par $n-2$ individus, etc., de sorte que si $S_{n-1} > 0$ alors forcément il y a une branche de l'arbre qui sous-tend $n-1$ individus quand on le construit (i.e., en haut de l'arbre on a 

), si S_{n-2} alors on force l'existence d'au moins une branche sous-tendant $n-2$ individus de l'échantillon, etc. Les mutations sont aussi placées de manière incrémentale sur l'arbre formé peu à peu et à la fin de chaque itération, le vecteur des temps des différentes époques est mis à jour suivant le principe que

si $T \sim \text{Gamma}(k, \lambda)$, alors T conditionné à ce que $\text{Poisson}(\mu T) = m$ suit une loi $\text{Gamma}(k+m, \lambda+\mu)$ $\rightsquigarrow$ plus on a mis de mutations sur une branche et plus elle aura tendance à être longue.

Le poids w associé à (F, M, T) est lui aussi calculé de manière incrémentale en multipliant les probabilités de chaque opération effectuée dans la construction de (F, M, T).

On ne va pas trop s'apesantir sur le pseudo code de cette méthode (cf article Sainudiin R et V.A. (2018). Full likelihood inference from the site frequency spectrum at a non-recombining locus. Theor. Pop. Biol. 124: 1-15).

Applications :

1) Scénarios pour lequels les vraisemblances sont connues explicitement

Pour tout scénario structurel et démographique $\bar{\varphi} \in \bar{\bar{\Phi}}$ (← pour le départager de Φ par lequel on avait appelé l'ensemble des scénarios démographiques considérés, avec le coalescent de Kingman comme modèle de topologie des généalogies), on a

$$\mathbb{P}_{\bar{\varphi}}(\underline{S}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{SFS(M_i) = \underline{S}\}} \frac{\mathbb{P}_{\bar{\varphi}}((F_i, M_i, T_i))}{w_i}$$

($\underline{S}$: SFS observé)

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{P}_{\bar{\varphi}}((F_i, M_i, T_i))}{w_i}$$

puisque par construction tous les triplets (F_i, M_i, T_i) échantillonnés sont compatibles avec $\underline{S}$.

Puis on utilise la forme du modèle de mutations poissonniennes pour écrire

$$\mathbb{P}_{\bar{\varphi}}((F_i, M_i, T_i)) = \mathbb{P}_{\bar{\varphi}}(M_i \mid (F_i, T_i)) \underbrace{\mathbb{P}_{\bar{\varphi}}((F_i, T_i))}_{\text{supposé explicite}}$$

$$\downarrow$$

$$= \prod_{k=2}^{n} \prod_{j=1}^{m-1} e^{-\mu F_{kj} T_k} \frac{(\mu F_{kj} T_k)^{M_{kj}}}{M_{kj}!}$$

(indépendant de $\bar{\varphi}$ en fait, sauf si on cherche ρ également et qu'on la considère comme étant dans $\bar{\varphi}$).

2) Scénarios pour lesquels les vraisemblances ne sont pas connues explicitement

Dans ce cas, on peut chercher les arguments $(\underline{A}, \beta, \mu)$ qui maximisent la probabilité des données $\underline{S}$ en supposant que la généalogie de l'échantillon suit le modèle du Beta-splitting de paramètre β, les temps entre coalescences sont indépendants et suivent une loi exponentielle de paramètre A_i et les mutations tombent sur l'arbre à taux μ.

On obtient alors un modèle "statistique" basé sur l'équilibre de l'arbre et le nombre de mutations comme proxy pour la longueur des branches, au contraire des modèles plus mécanistes qu'on cherchait à reconstruire en 1):

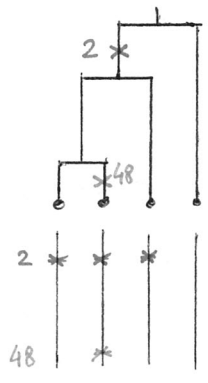
$$\mathbb{P}_{(\underline{A},\beta,\mu)}(\underline{S}) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{P}_{(\underline{A},\beta,\mu)}((F_i, M_i, T_i))}{\omega_i}$$

(Montrer figures de l'article)

V Inférence basée sur les arbres de Tajima (avec J. Palacios, L. Capello, J. Wakeley, S. Ramachandran)

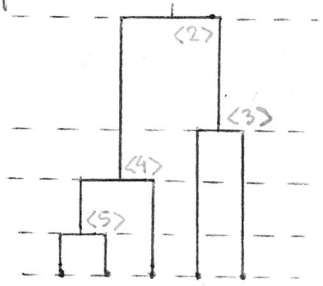
Supposons maintenant que nos données D_n sont constituées d'alignements de séquences (qu'on suppose non-étiquetés dans ce qui suit, i.e. les séquences ne sont pas numérotées 1, 2, ..., n elles-mêmes). Cette fois la résolution des "tree shapes" n'est plus suffisante pour caractériser la loi des données car on a besoin de pouvoir encoder (le fait) qu'une certaine mutation portée par un individu (par exemple) est "sous" une autre mutation dans l'arbre de sorte que la séquence correspondant à cet individu porte les 2 mutations.

Ex :



L'information dont on dispose est que 3 séquences portent la mutation 2 et, parmi elles, une séquence porte aussi la mutation repérée au site 48.

Mais d'un autre côté on n'a pas non plus besoin de donner une étiquette à tous les individus de l'échantillon (et donc à toutes les feuilles de l'arbre) parce que si on permute les étiquettes on ne change rien au fait qu'on a une séquence qui porte les mutations 2 et 48, et 2 autres qui ne portent que la mutation 2. La résolution optimale consiste à n'étiqueter que les nœuds internes pour garder l'ordre des événements de coalescence. C'est plus simple sur un exemple :



$n=5$

$\Leftrightarrow$

. 1er événement est la coalescence de 2 singletons qui crée une branche de "vintage" / "millésime" / étiquette interne $<n>=<5>$

. 2ème évé^t, qui crée une branche d'étiquette $<4>$, est la fusion du bloc $<5>$ de la branche avec un autre singleton.

. 3ème évé^t, créant une branche d'étiquette $<3>$, est la fusion de 2 autres lignées singleton.

. 4ème évé^t, créant la branche "racine de l'arbre" d'étiquette $<2>$ est la fusion des blocs (branches) $<3>$ et $<4>$.

On se replace dans le cadre où chaque paire de lignées fusionne à taux égaux, indép. les unes des autres

Ce faisant, on obtient une résolution connue dans la littérature sous le nom de <u>Coalescent de Tajima</u>, ou "vintaged and sized" coalescent. Une manière de l'encoder mathématiquement est par exemple de définir

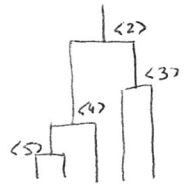
$$D_{ij} = \begin{cases} 1 & \text{s'il existe une branche à l'époque } i \text{ qui porte l'étiquette } <j> \\ 0 & \text{sinon} \end{cases}$$

(parce qu'en connaissant la matrice $(D_{ij})_{\substack{2\le i\le n \\ 1\le j\le n-1}}$ complète, on peut reconstruire l'histoire des coalescences de blocs singletons ou déjà étiquetés). Une manière différente d'encoder une telle topologie "millésimée" pour laquelle on a opté dans les travaux avec Julia Palacios est la suivante : pour $i \ge j$

H_{ij} = # blocs/lignées qui ne coalescent pas dans l'intervalle de temps $[t_{i+1}, t_j[$, où

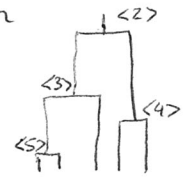
$t_i = T_n + T_{n-1} + \dots + T_i$ est le temps total qu'il faut pour tomber à $i-1$ lignées $\neq$. Par convention, $t_{n+1} = 0$.

Ex: Si on revient à l'exemple précédent, on a



$$\begin{bmatrix} 2 & & & \\ 1 & 3 & & \\ 0 & 2 & 4 & \\ 0 & 2 & 3 & 5 \end{bmatrix}$$

et pour comparaison



$$\begin{bmatrix} 2 & & & \\ 1 & 3 & & \\ 0 & 2 & 4 & \\ 0 & 1 & 3 & 5 \end{bmatrix}$$

Donc ce n'est peut-être pas facile à voir, mais en fait ces matrices permettent vraiment d'encoder l'ordre des coalescences.

L'ensemble des arbres de Tajima à n feuilles a en cardinal donné par les nombres de zigzags, ou nombres d'Euler, qui se calculent par récurrence. On peut montrer qu'il y a $\frac{n!}{2^{c(\underline{t})}}$ arbres avec feuilles étiquetées (type Kingman) correspondant à un arbre de Tajima $\underline{t}$ donné où $c(\underline{t})$ est le nombre de "cerises" de l'arbre, i.e. le nombre de paires de singletons qui coalescent dans $\underline{t}$.

(*)

A nouveau, on a réussi à faire sérieusement diminuer le cardinal de l'ensemble des arbres cachés à explorer pour calculer la vraisemblance des scénarios considérés :

$$\mathbb{P}_\varphi(D_n) = \int_{(H,T)\in\mathcal{H}_n\times\mathcal{T}_n} \mathbb{P}_\varphi(D_n \mid (H,T))\, d\mathbb{P}_\varphi((H,T))$$

Dans les travaux avec Julia et Lorenzo, on considère que H a la loi du coalescent de Tajima et que $(N_e(t))_{t\geq 0}$ suit la loi de l'exponentielle d'un processus Gaussien (de moyenne 0 et de covariance $(C(\tau), \tau\geq 0)$).

On a alors $\mathbb{P}_\varphi(D_n) = \int_{(H,T)} \mathbb{P}_\varphi(D_n \mid (H,T))\, \mathbb{P}_{\text{Taj.}}(H)\, d\mathbb{P}_\varphi(T)$

$$= \int_{(H,T)} \mathbb{P}_\varphi(D_n \mid (H,T))\, \frac{2^{n-1-c(H)}}{(n-1)!}\, \mathbb{E}\left[\prod_{k=2}^{n} \left\{ \frac{\binom{k}{2}}{N_e(t_k)} \exp\left(-\int_{t_{k+1}}^{t_k} \frac{\binom{k}{2}}{N_e(s)}\, ds\right)\right\}\right] dt_2 \dots dt_n$$

où l'espérance correspond au processus gaussien que modélise $(\ln N_e(t))_{t\geq 0}$.

Le reste du travail, qu'on ne verra pas ici, consiste à trouver une manière algorithmiquement simple

(*) Ex :

n	#Kingman	#Tajima
3	3	1
5	180	5
7	56 700	61
10	$2{,}5 \cdot 10^9$	7936
12	$9{,}3 \cdot 10^{12}$	353 792

de calculer la probabilité des données sachant le couple (H, T) (qui nécessite de scanner toutes les manières d'allouer les mutations observées aux branches de l'arbre de sorte que l'alignement de séquences observé soit respecté).

Par ailleurs, malgré la réduction de la taille de l'espace où vit la variable généalogique cachée, celui reste très grand même pour des valeurs de n pas très grandes et donc on calcule une valeur approchée de la proba des données via des méthodes MCMC.

Diapos avec l'exemple des bisons $\rightsquigarrow$ on a étendu cette méthodologie à des échantillons pris à plusieurs temps.

Bibliographie (non exhaustive) :

Approche « Poisson Random Field »

- R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson, and C.D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics, 5(10) : e1000695, 2009.
- R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics, 154 : 931-942, 2000.
- S.A. Sawyer and D.L. Hartl. Population genetics of polymorphism and divergence. Genetics, 132 : 1161-1176, 1992.

Approche « fonctions génératrices »

- L. Bunnefeld, L.A.F. Frantz, and K. Lohse. Inferring bottlenecks from genome-wide samples of short sequence blocks. Genetics, 201 : 1157-1169, 2015.
- K. Lohse, R.J. Harrison, and N.H. Barton. A general method for calculating likelihoods under the coalescent process. Genetics, 189 : 977-987, 2011.

Approche « skyline plots »

- S.Y.W. Ho and B. Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Res., 11 : 423-434, 2011.

Approche « Sequential Markov coalescent »

- P. Marjoram and J. Wall. Fast coalescent simulation. BMC Genetics, 7 :16, 2006.
- G. McVean and N. Cardin. Approximating the coalescent with recombination. Phil. Trans. Royal Soc. B, 360 : 1387-1393, 2005.

Approche « Approximate Bayesian Computation (ABC) »

- S. Boitard, W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz. Inferring population size history from large samples of genome-wide molecular data - An Approximate Bayesian Computation approach. PLoS Genetics, 12(3) : e1005877, 2016.
- J.-M. Marin et ses collaborateurs ont plein de travaux très intéressants sur cette approche.
- B.M. Peter, D.Wegmann, and L. Exco_er. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol. Ecol., 19 :4648-4660, 2010.