Au-delà du coalescent : quels modèles pour expliquer la diversité génétique ?

Julien Berestycki

LPMA, Université Paris 6, CMAP Polytechnique

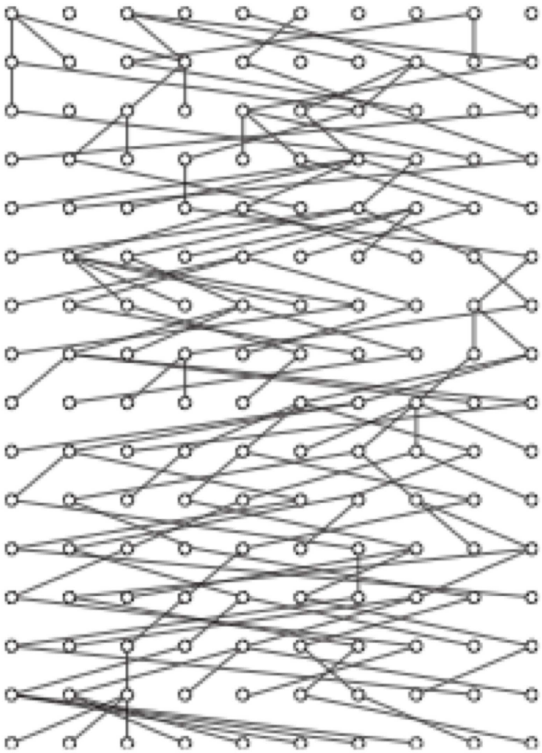
13 octobre 2009

A partir de travaux conjoints avec N. Berestycki, V. Limic, J. Schweinsberg

Outline

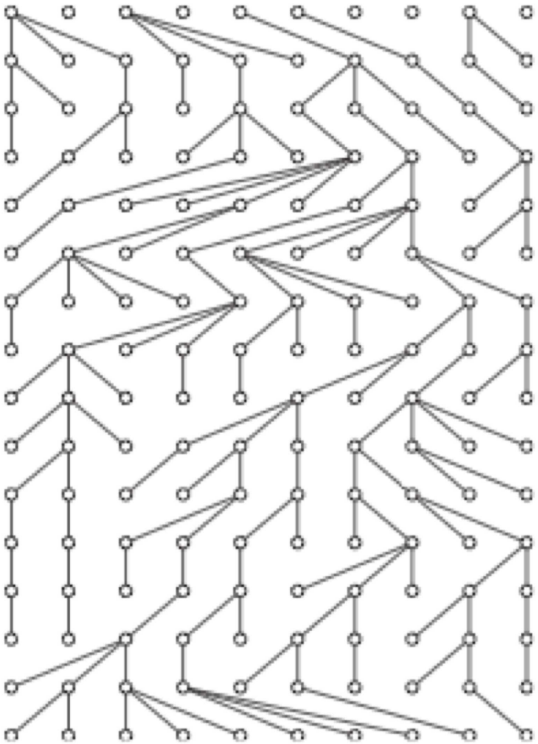
Question : comment expliquer l'évolution ?
Depuis 1859 : La sélection est le moteur de l'évolution (The natural selection, C. Darwin)
Depuis $\sim$1970 : Le moteur de l'évolution moléculaire est le drift génétique (The neutral theory, M Kimura)

Une mutation apparaît et se transmet $\longrightarrow$ modèle de reproduction ;

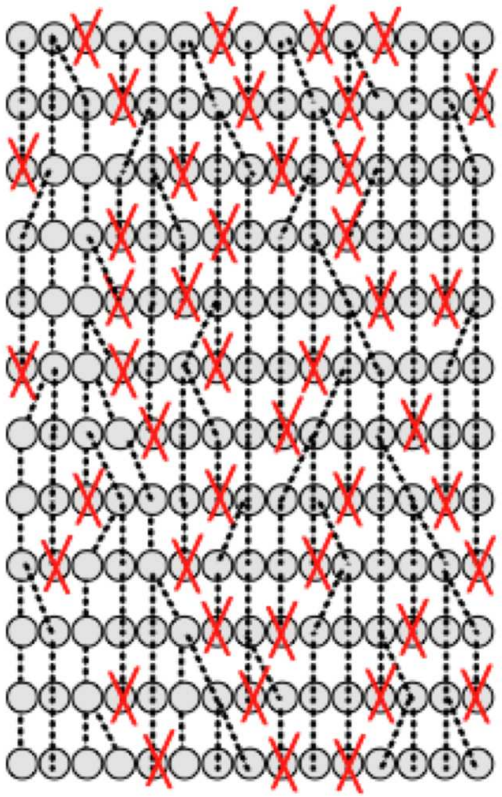


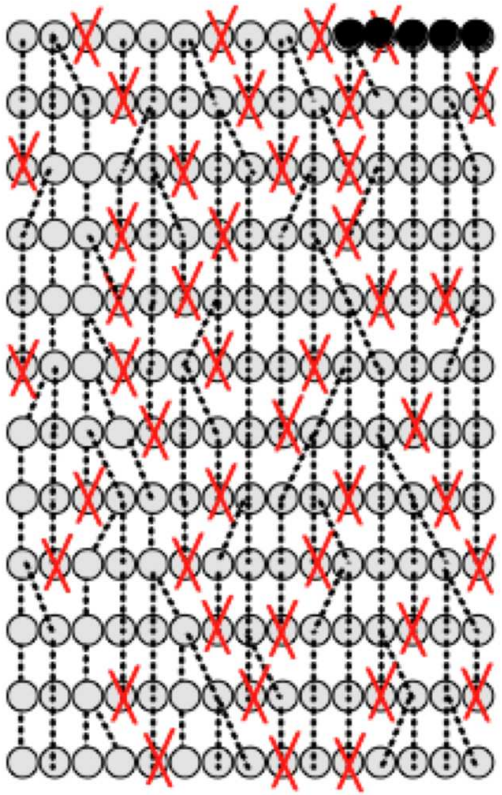
Hypothèses : Pas de sélection, Population de taille N constante, panmictique. Générations discrètes et sans recouvrements, haploide.
Pas d'individus favorisés : si $N = 6, (\nu_1, \ldots, \nu_6) = \#$ d'enfants alors $P(3, 1, 0, 0, 2, 0) = P(1, 0, 0, 0, 3, 2)$ (échangeable). A chaque génération on tire un vecteur ν iid. Modèle de Cannings

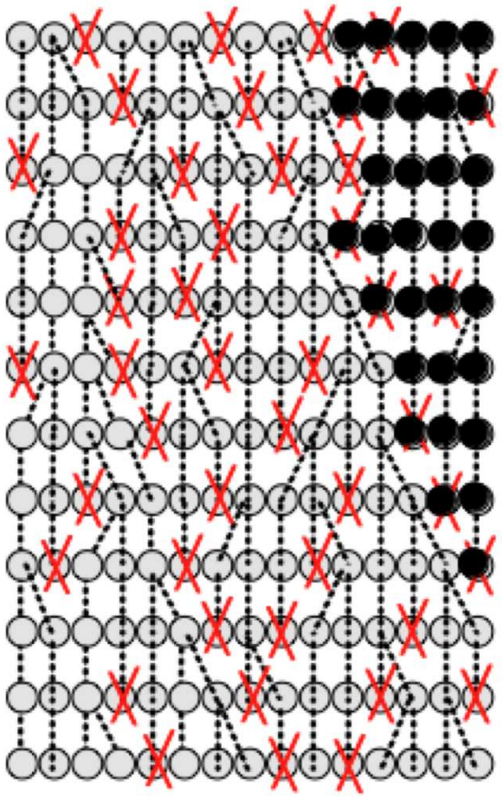
Pas d'ordre sur les individus.

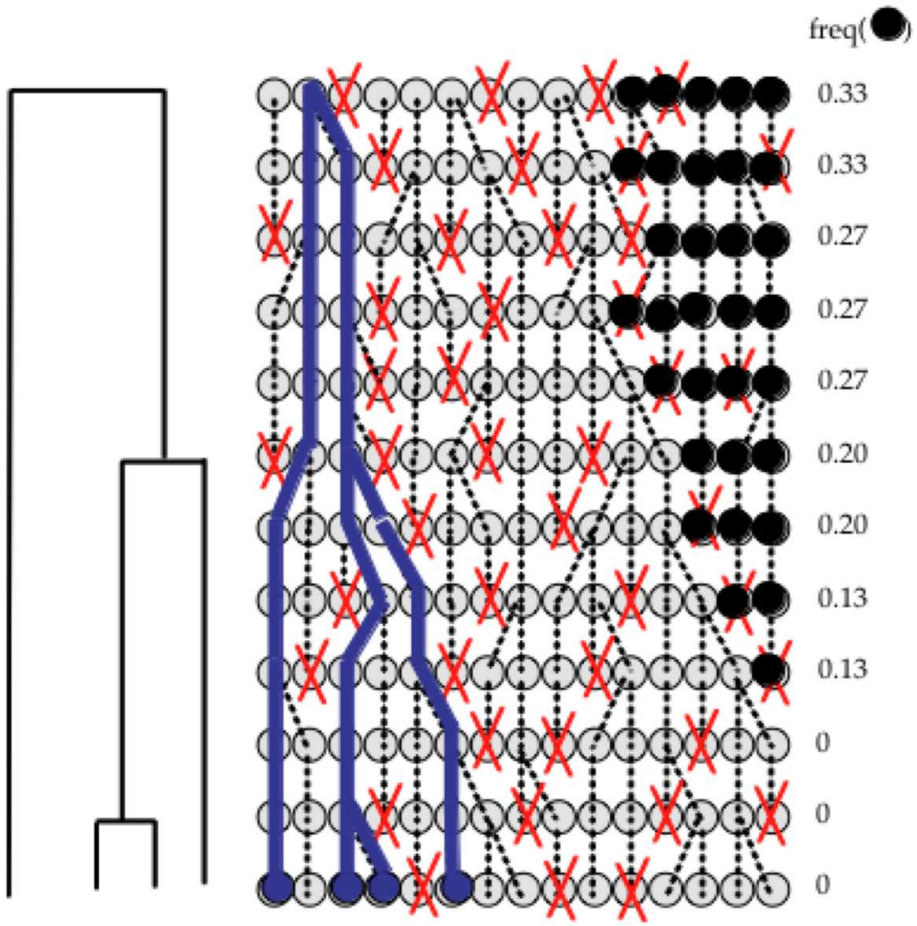


Wright-Fisher : La loi du vecteur ν est multinomiale $\Leftrightarrow$ chaque individu choisit son parent dans la génération précédente uniformément.
Question 1 : Proportion d'allèle (en avant dans le temps)
Question 2 : Généalogie de k indiv. quand $N \longrightarrow \infty$ (en remontant dans le temps)

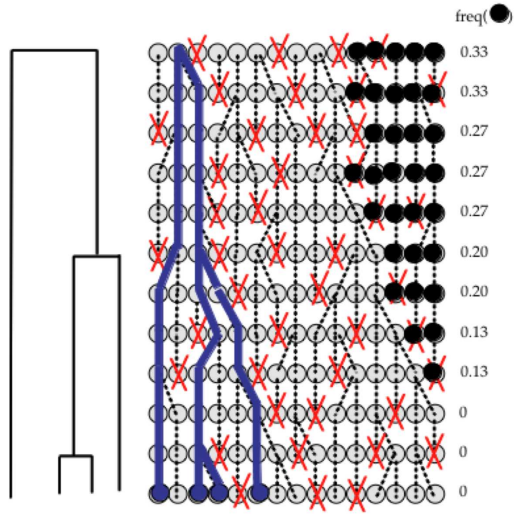







freq(●)
0.33
0.33
0.27
0.27
0.27
0.20
0.20
0.13
0.13
0
0
0

Quand $N \longrightarrow \infty$



Proportion de $(\circ) \longrightarrow$ diffusion de Wright Fisher

Généalogie de k indiv. $\longrightarrow$ objet limite = le coalescent de Kingman. (temps mesuré en unités de N générations.

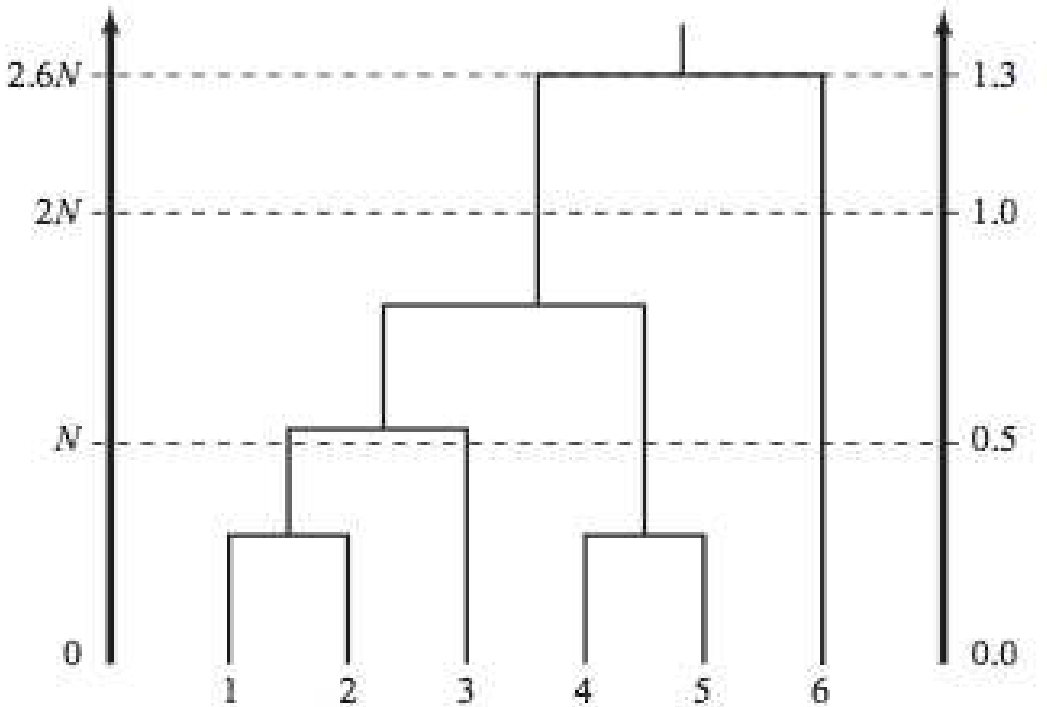
Robuste Reste vrai pour de nombreuses lois des vecteurs ν.

Un modèle naturel de Cannings $\neq$ Wright Fisher.

À Chaque génération :

- reproduction libre Chaque individ. produit Y_i enfants, Y_i iid. Avec proba $\sim_{N\to 1} 1$ on a $\sum_{i=1}^{N} Y_i > N$.
- limitation (ressources,...). Seuls N enfants survivent, tirés uniforméments parmi $\sum_{i=1}^{N} Y_i$.

Si $E(Y_i^2) < \infty$ alors on est dans le régime Kingman / Wright Fisher.

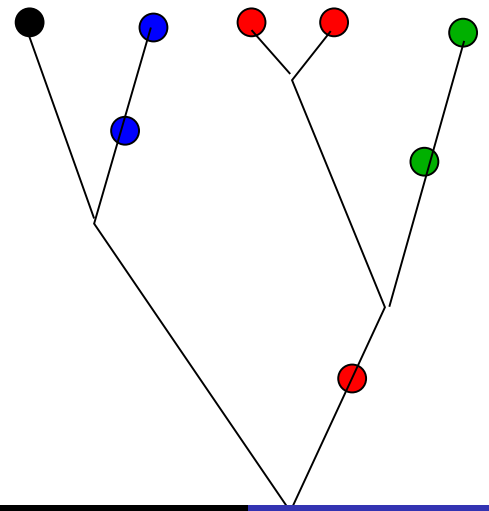


Dynamique : $\forall$ paire de lignés coalesce à taux 1.
T_i Temps durant lequel exactement i lignées actives.
$E(T_i) = 2/(i(i-1))$
$(E(T_2) = 1, E(T_3) = 1/3, \ldots)$.
Plus récent ancêre commun (MRCA)
$E(MRCA) = 2(1 - 1/n)$.

Quelques réalisations

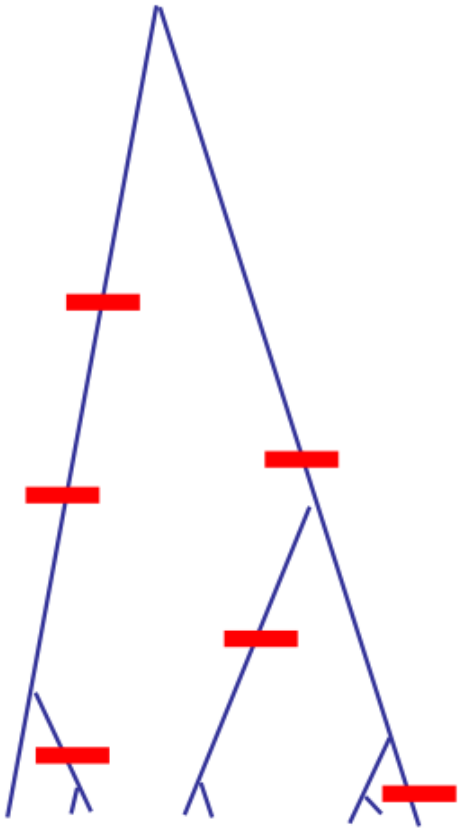
ISM et IAM

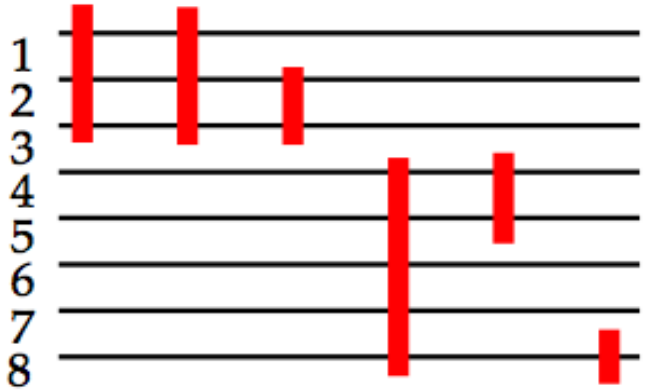
Dans le modèle de WF, chaque indiv.a une proba θ/N de muter/ son parent.
Les mutations arrivent le long d'une séquence ADN (. . .ATTGTCA . . .) : chaque mutation affecte un nouveau site. Si la séquence code pour une couleur, chaque mutation crée une nouvelle couleur jamais vue.
Quand $N \to \infty$ les mutations surviennent à taux θ le long de chaque lignée.



Arbres et séquences

On peut reconstruire l'arbre à partir des séquences





ESF

Ewens sampling formula : la loi de la partition de l'échantillon en couleurs.
Si $\forall j, \pi$ a a_j familles de taille j, alors

$$P(\Pi_n = \pi) = \frac{n!}{\theta(\theta+1)\dots(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!}.$$

$\#$ de couleurs $\sim \theta \log n$
E ($\#$ de couleurs portées par 1 indiv.) $= n\theta/(n+\theta-1)$.

Taille effective de population

Population effective = la taille qui fait que ça marche
Dans le modèle de WF : age MRCA $= 2N, P(T_2 = 1) = 1/N, \ldots$
Donc on définit $N_e = 1/P(T_2 = 1)$ ou $N_e = ageMRCA/2$, etc...
Si dans chaque gén. trois gènes génèrent chacun 1/3 de la pop. à la gén. suivante alors $N_e = 3$;

- Pop. humaine : pour de nombreux gènes MRCA ≤ 400.000 ans. Une gén. $= 20$ ans et $E(\text{MRCA}) = 2N \Rightarrow N \leq 10000\,!\,!\,!$ La population est modélisée par un WF avec $N = 10000$.
- D. Melanogaster gen ~ 15 jours, $N_e = 10^4$ MRCA ~ 820 ans.
- HIV (dans un patient) gen ~ 1.5 jours, $N_e \sim 10^3 - 10^6$ donc MRCA 4 - 400 ans.

HIV

Pour estimer la population effective on utilise le *turn over* (combien de temps pour une population génétiquement distincte ?) observé $\sim$ 2 ans.
$\Rightarrow N_e \sim 10^3 - 10^4$ alors que la pop. virale $\sim 10^8$.

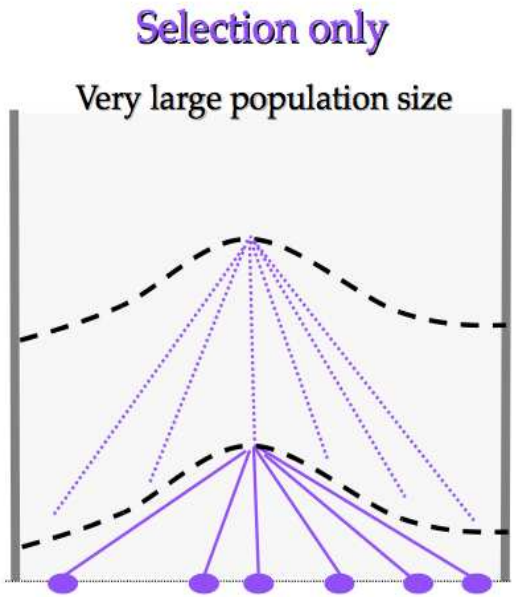
Quand utiliser N_e ?

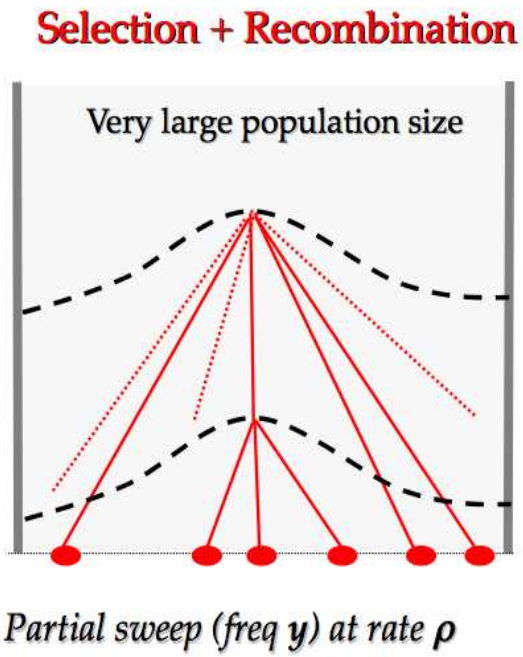
Utiliser N_e dans le modèle de WF ou Kingman n'est pas toujours justifié. Probablement ok pour les pop. ou la généalogie est de *type* Kingman (i.e. pop. humaine), mais pas quand on a des évènements de coal. multiples

Quelques exemples suivent :

HIV intra patient

La population de virus est soumise à une très forte sélection.
Conduit à des balayages sélectifs répétés.

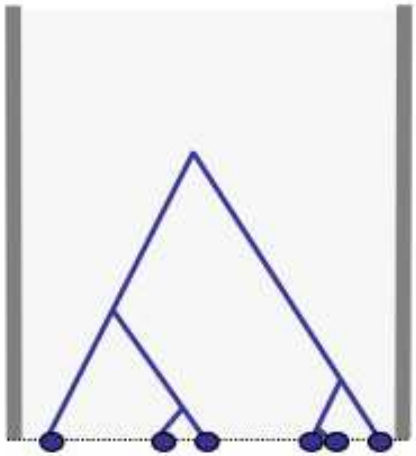




(Gillespie 2000a, 2000b, ...)

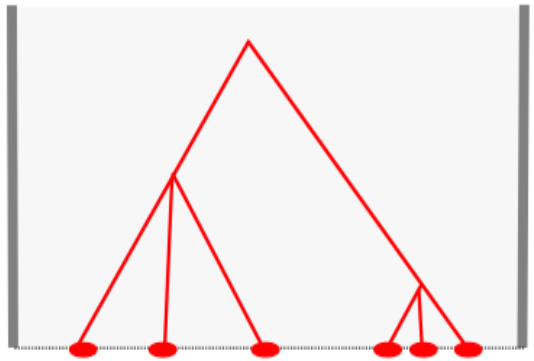
HIV intra patient

Coalescence = Inbreeding



Drift-Mutation
(the "reference" model)

Coalescence = hitchhiking



Selection-Recombination-Mutation
(an "alternative" model)

$$E(\pi) = 2N\theta, N_e = \pi/2\theta$$
$$E(\pi) = 2\theta/\rho y^2$$

Biologie marine

Li and Hedgehock (1998) *Genetic heterogeneity [...] among samples of larval Pacific oysters (Crassostrea gigas) supports the hypothesis of large variance in reproductive success* Can. J. Fish. Aquat. Sci.

Analyse un échantillon de $n = 666$ larves d'huitre. Trouve $S_n = 67$ haplotypes différents, et parmi eux 72% ne se trouvent que dans 1 indiv.

Prediction pour Kingman : $\approx \theta/\theta \log n \approx 15\%$.

Eldon Wakeley *Coalescent processes when the distribution of offspring number among individuals is highly skewed* et Birkner Blath *Inference for Λ-coalescents*.

Retour sur le modèle de Cannings

À Chaque génération :

- reproduction libre Chaque individ. produit Y_i enfants, Y_i iid. Avec proba $\sim_{N\to 1} 1$ on a $\sum_{i=1}^{N} Y_i > N$.
- limitation (ressources,...). Seuls N enfants survivent, tirés uniforméments parmi $\sum_{i=1}^{N} Y_i$.

Si $P(Y_i > k) \sim k^{\alpha}$ avec $\alpha \in (1,2)$ alors on ne converge plus vers Kingman. Autre arbre limite : le Beta-coalescent (Schweinsberg 2000). (collisions multiples possibles).

Sélection

Conjecture de Brunnet Derrida Simon. Population de N individus sur la droite réelle. Position = fitness. À chaque gén. les indiv.produisent 2 enfants situés à une dist. Z du parent. Les Z sont iid. On garde les N plus à droite pour faire la gén. suivante.

Conjecture

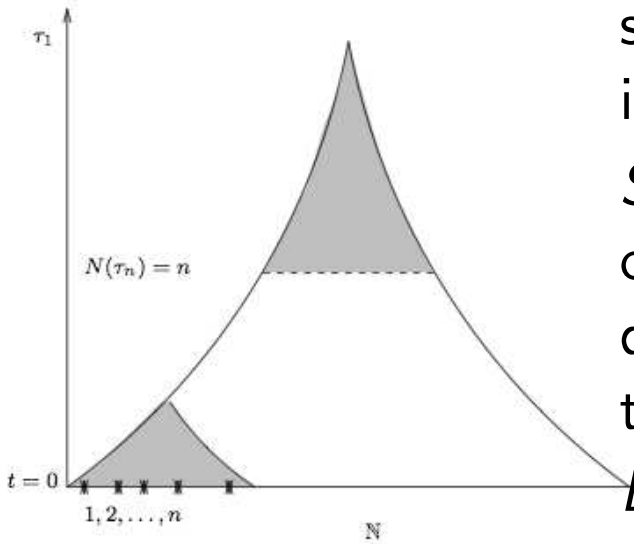
Bien renormalisée, la généalogie converge vers le Coaescent de Bolthausen Sznitman.

Progrès récents en ce sens pas B., Berestycki et Schweinsberg.

To do list

Objectif

Développer une description du polymorphisme génétique attendu pour ces autres coalescents.



Ex : l'asymptotique du # SNP sites de ségregation S_n, ou # allèles K_n parmi n indiv. (Kingman $\sim \theta \log n$)

S_n est proportionel à $L_n \sim \int_0^{T_1^n} N^n(t)dt$ où $N^n(t) = \#$ de blocs au temps t quand on part de n blocs et T_1 premier temps où un seul bloc. On a aussi $L_n \sim \int_{T_n^\infty}^{T_1^\infty} N^\infty(t)dt$.

Descente de l'infini :
formulation biologique l'age du MRCA converge-t-il quand $n \to \infty$?
formulation mathématique si $T = \inf\{t : N(t) = 1\}$ est ce que T fini ?

Théorème

Pitman '99 $\Lambda(\{1\}) = 0. N(0) = \infty$. *Avec proba 1 on a l'un des deux cas suivants*

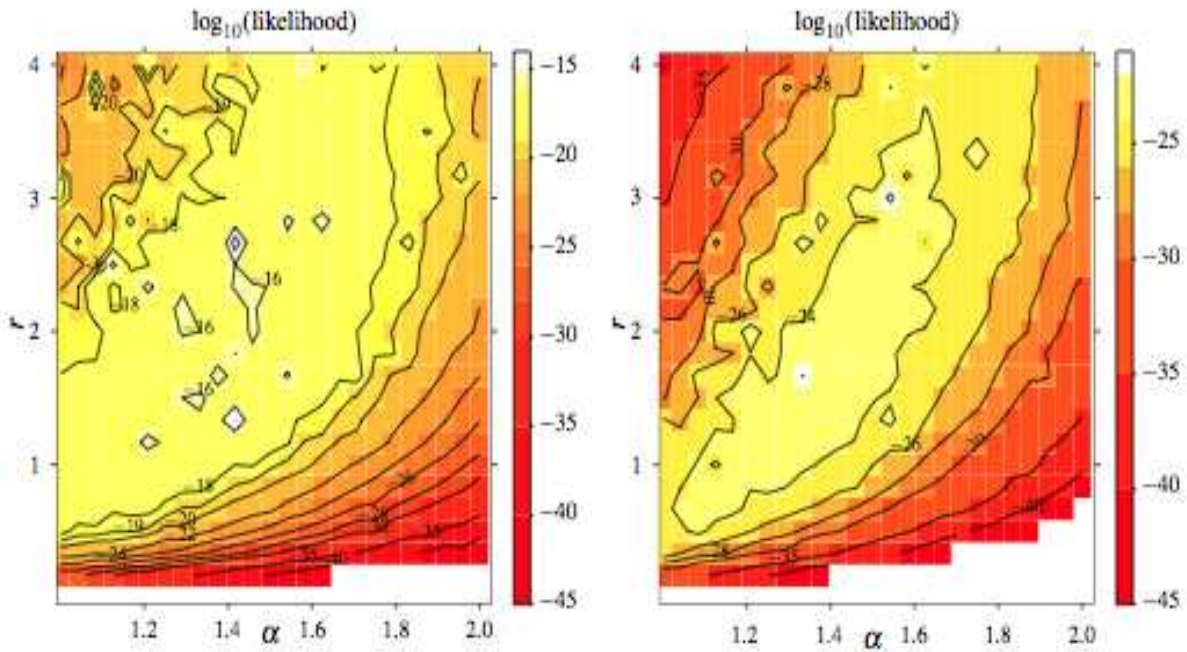
- $\forall t > 0, N(t) = \infty$
- $\forall t > 0, N(t) < \infty$ *(et dans ce cas CD∞).*

On a des critères.

Beta-coalescents

Famille à 1 paramètre, objet naturel (Galton-Watson avec ressources finies), va de Kingman au BolthausenSznitman.
B. Berestycki et Schweinsberg donnent une formule asymptotique pour la partition allélique et le spectre de mutation.
Permet de faire de l'estimation.

EMV


log10(likelihood)
log10(likelihood)
α
α
1.2
1.4
1.6
1.8
2.0
1
2
3
4
−15
−20
−25
−30
−35
−40
−45

(a) 1.25 (b) 1.5

Un peu de maths

Qu'est ce qu'un Λ-coalescent ? On veut que quand b lignées chaque k-tuple coalesce à taux $\lambda_{b,k}$.
Flash quizz Quels $\lambda_{b,k}$ pour avoir Kingman ?
Question : Quels sont les $\lambda_{b,k}$ possibles ? (Attention il faut consistence. Échantillon de taille $n+1$, avec cette dynamique. Si on oublie indiv. $n+1$ on veut la bonne loi.

Théorème

(Pitman 99) Possibl SSI $\exists \Lambda$ *mesure finie sur* $[0,1]$ *t.q.*

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

Kingman : $\Lambda = \delta_0$
Rateau : $\Lambda = \delta_1$

Cas des **Beta-coalescents**.
$\Lambda = \text{Beta}(2-\alpha, \alpha)$ $1 < \alpha < 2$.

$$\Lambda(dx) = \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1} dx$$

CD∞
(Berestycki-B.-Schweinsberg 06)

Théorème

Soit $N_t = \#$ blocs dans un Beta-coalescent.

$$t^{\frac{1}{\alpha-1}} N_t \xrightarrow[t\to 0]{a.s.} (\alpha\Gamma(\alpha))^{\frac{-1}{\alpha-1}},$$

Spectre allélique dans le cas Beta

Théorème

(BBS, BBL) Fixons $k \geq 0$. Si la généalogie est décrite par un Beta-coalescent (α).
$S_n = \#$ sites de ségrégation

$$n^{\alpha-2} S_n \to_{p.s} \theta \frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}.$$

$N_k(n) = \#$ de types portés par k indiv.

$$n^{\alpha-2} N_k(n) \to_{p.s} \theta\alpha(\alpha-1)^2 \frac{\Gamma(k+\alpha-2)}{k!}$$

Permet de fitter α sur exemple huitre $\alpha = 1.42...$
Pb : ne concerne que les petites mutations...

Cas général

Pour une mesure de proba. Λ donnée.

$$\psi(q) := \int_0^1 (e^{-qx} - 1 + qx)x^{-2}\Lambda(dx)$$

ψ est le mécanisme de branchement d'un CSBP (processus de branchement à espace d'états continu) $(Z_t, t \geq 0)$
Si

$$\int^{\infty} \frac{dq}{\psi(q)} < \infty \tag{1}$$

on pose

$$u(s) = \int_s^{\infty} \frac{dq}{\psi(q)}$$

et soit $v(t)$ son inverse cadlag.

Théorème

BBL For any fixed $x > 0$

$$S_n / \left(\int_x^n \frac{q dq}{\psi(q)} \right) \to \theta \tag{2}$$

in probability.
If furthermore Λ has (strong) regular variation at zero in the sense that $\Lambda(dx) = f(x)dx$ where $f(x) \sim Ax^{1-\alpha}$ as $x \to 0$ for some $A > 0$ and $1 < \alpha < 2$, then the above convergence holds almost surely and thus

$$\frac{S_n}{n^{2-\alpha}} \longrightarrow \theta B, a.s.$$

for some constant B depending only on A and α.

Théorème

Suppose that Λ has (strong) regular variation at zero in the sense that $\Lambda(dx) = f(x)dx$ where $f(x) \sim Ax^{1-\alpha}$ as $x \to 0$ for some $A > 0$ and $1 < \alpha < 2$.
For all $k \geq 1$, as $n \to \infty$,

$$\frac{F_{k,n}}{n^{\alpha-2}} \to \theta C \frac{\Gamma(k+\alpha-2)}{k!}, \quad a.s. \tag{3}$$

and

$$\frac{M_{k,n}}{n^{\alpha-2}} \to \theta C \frac{\Gamma(k+\alpha-2)}{k!}, \quad a.s. \tag{4}$$