Coalescence : bases théoriques et applications à la simulation de données génétiques et à la cartographie génétique.

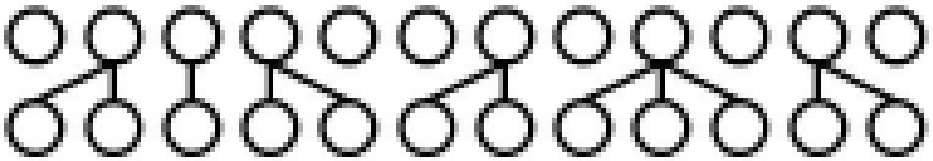
Simon Boitard

LGC

Réunion Rules and Tools, 23 sept 2010

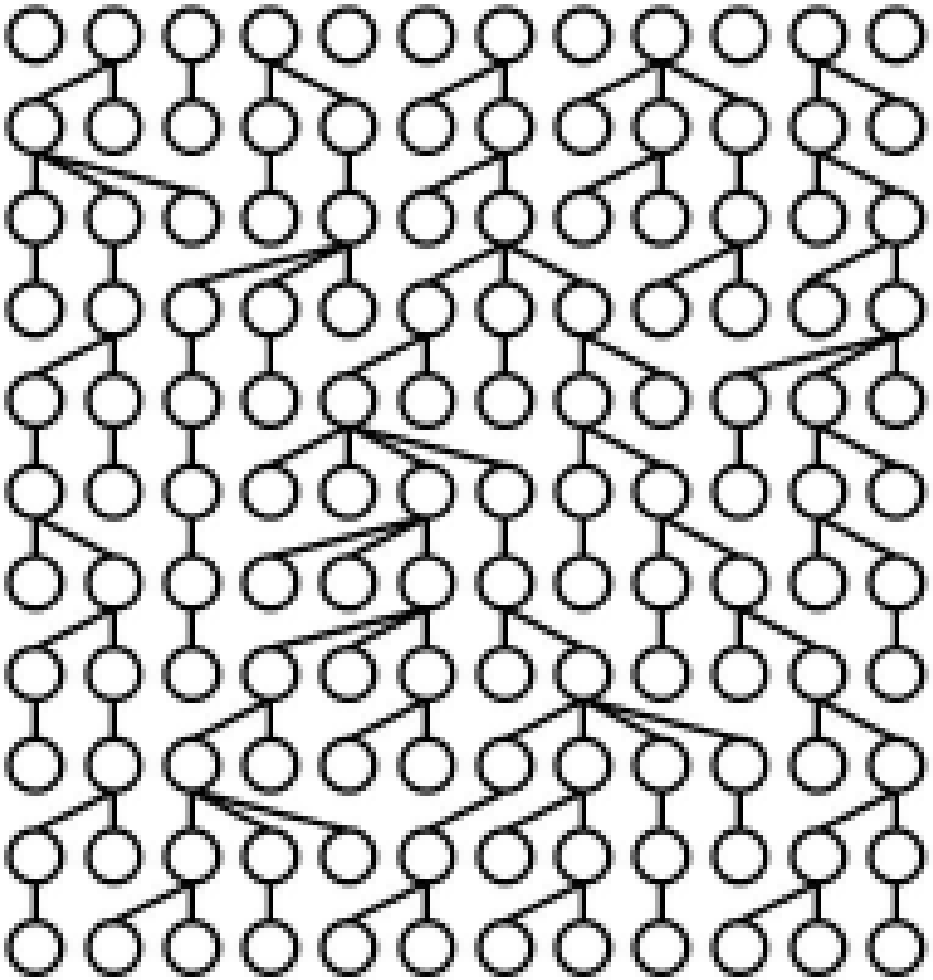
Introduction

Introduction



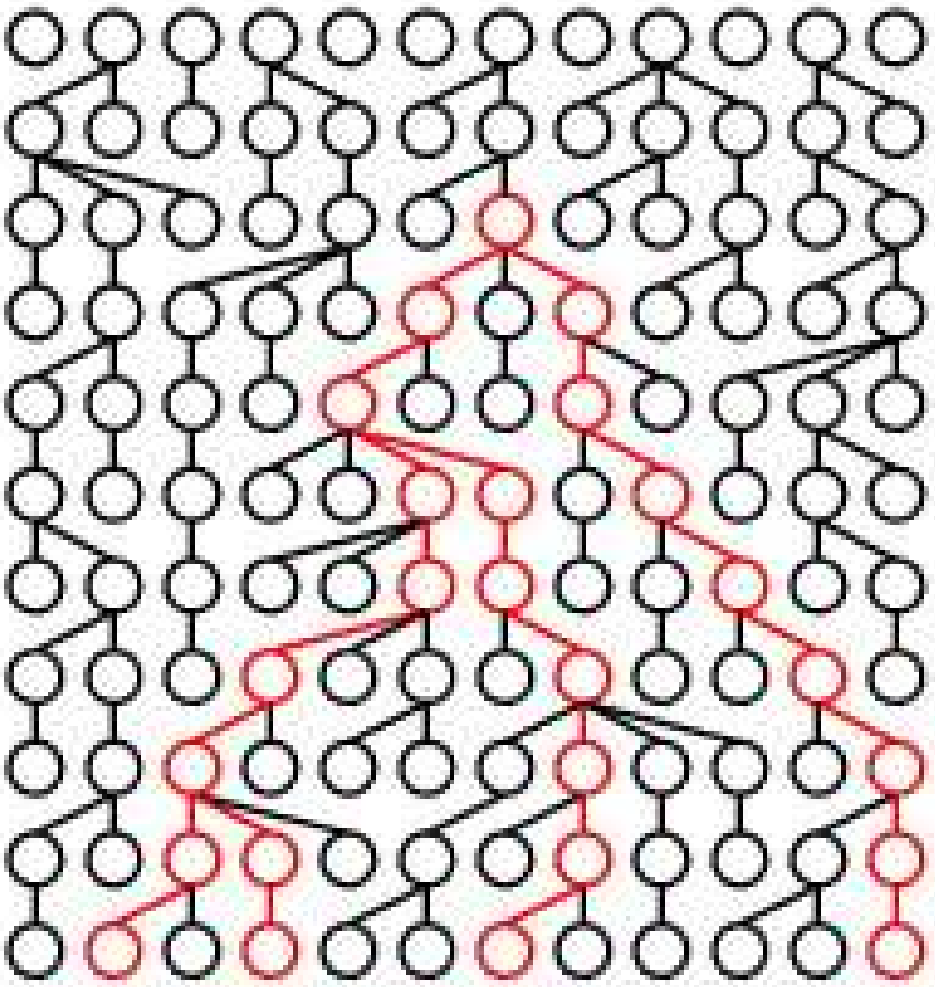
population de taille N

Introduction

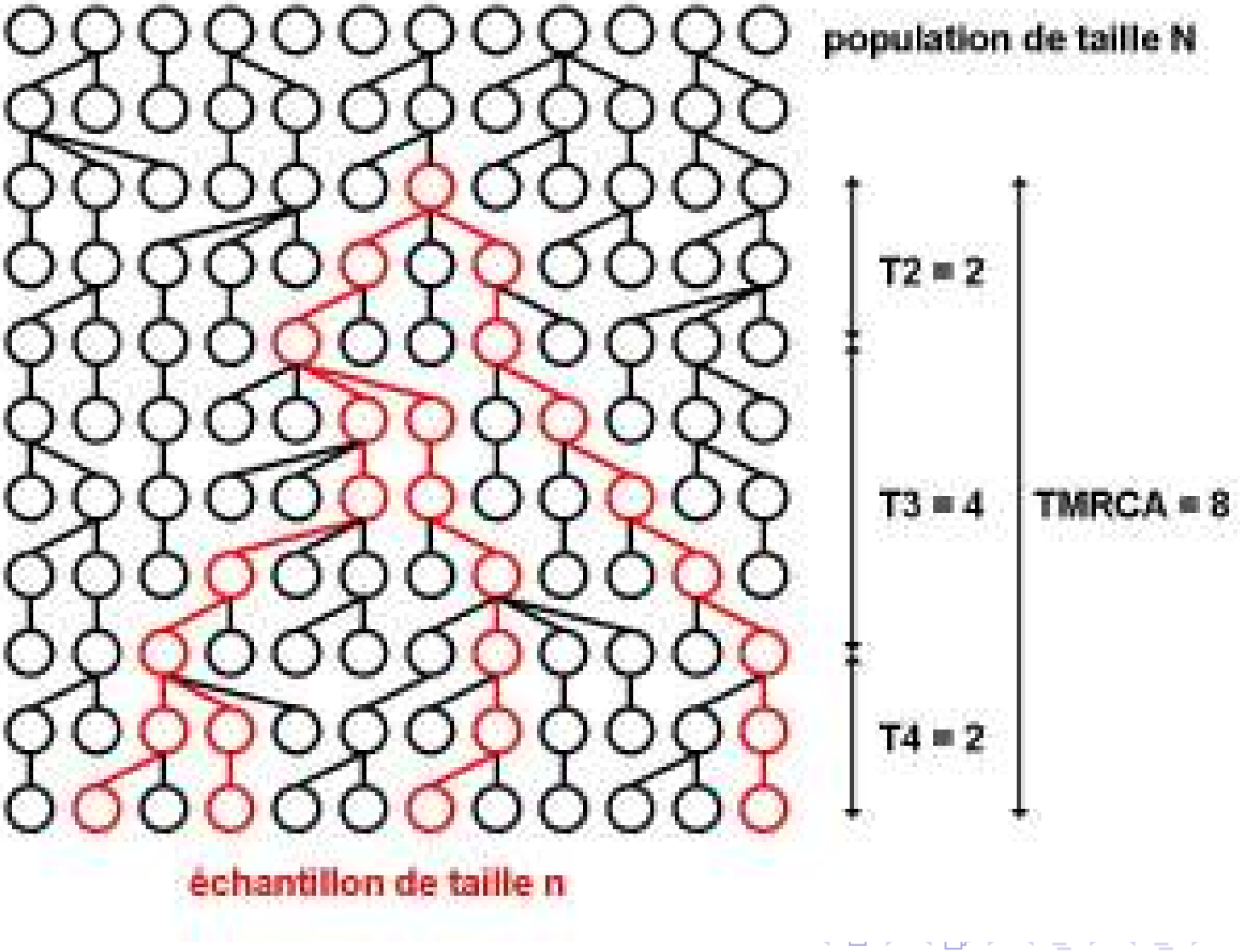


population de taille N

Introduction



Introduction



Outline

Temps de coalescence

- Proba qu'il n'y ait pas de coalescence à une génération donnée

$$q^N(n) = \prod_{i=1}^{n-1}(1 - \frac{i}{N}) = 1 - \frac{n(n-1)}{2N} + O(\frac{1}{N^2})$$

- T_n^N suit une loi géométrique

$$\mathbb{P}(T_n^N > t) = (q^N(n))^t$$

- Changement d'échelle $\tau = \frac{t}{N}$

$$\mathbb{P}(T_n^N > N\tau) = (q^N(n))^{N\tau} \approx (1 - \frac{n(n-1)}{2N})^{N\tau} \to e^{-\frac{n(n-1)}{2}\tau}$$

quand $N \to +\infty$.
$\to$ T_n^N tend vers T_n, de loi exponentielle de paramètre $\frac{n(n-1)}{2}$.

Evènements de coalescence

- La proba que trois individus ou plus coalescent à la même génération est en $O(\frac{1}{N^2})$
  $\rightarrow$ négligeable quand $N \rightarrow +\infty$.
- En pratique, une coalescence consiste toujours à regrouper exactement deux lignées.

Mutations

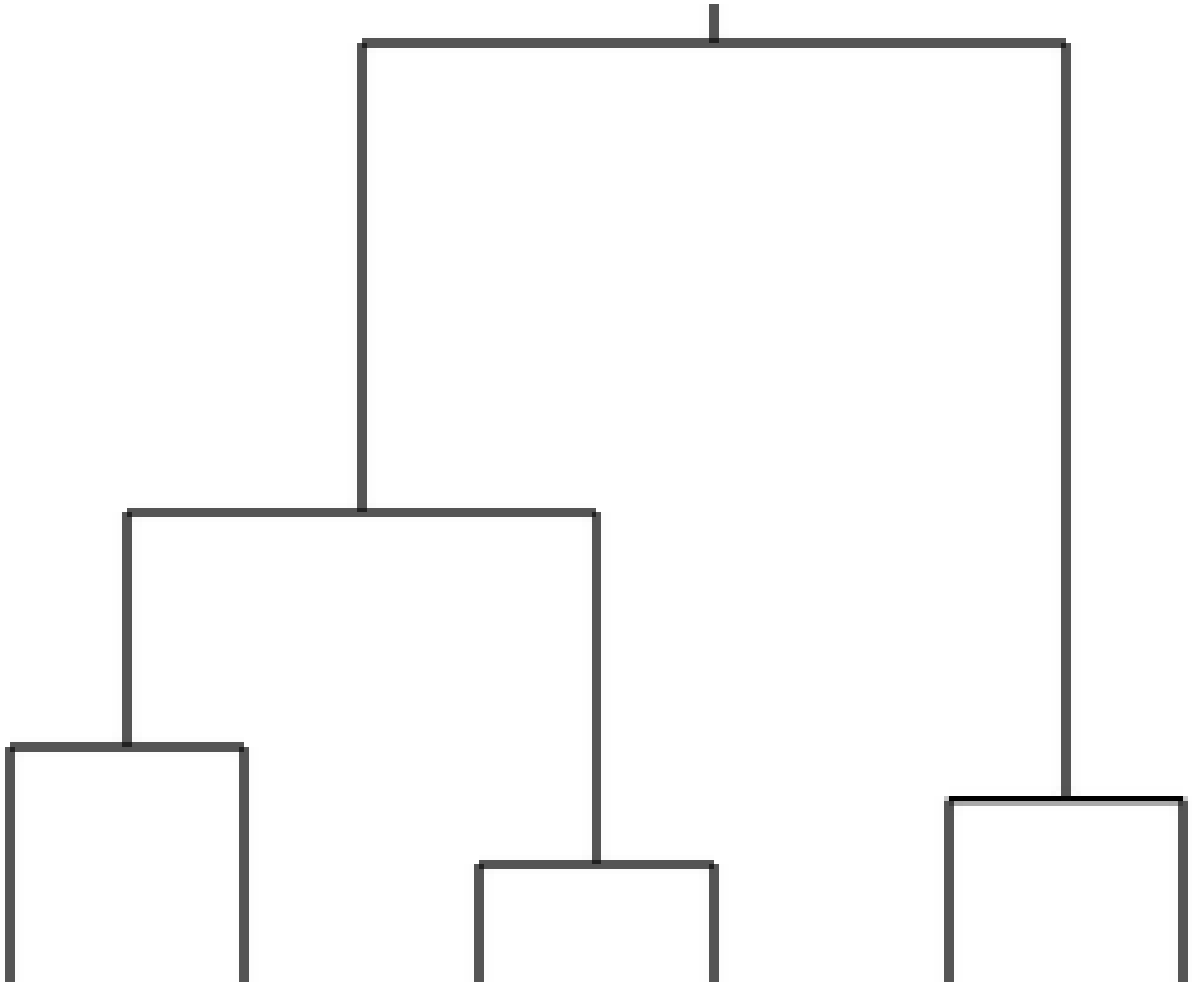
- Proba μ de mutation par méiose.
- $M(t)$ nombre de mutations pour une branche de longueur $t = N\tau$

$$\mathbb{E}[M(t)] = \mu N\tau = \frac{\theta}{2}\tau$$

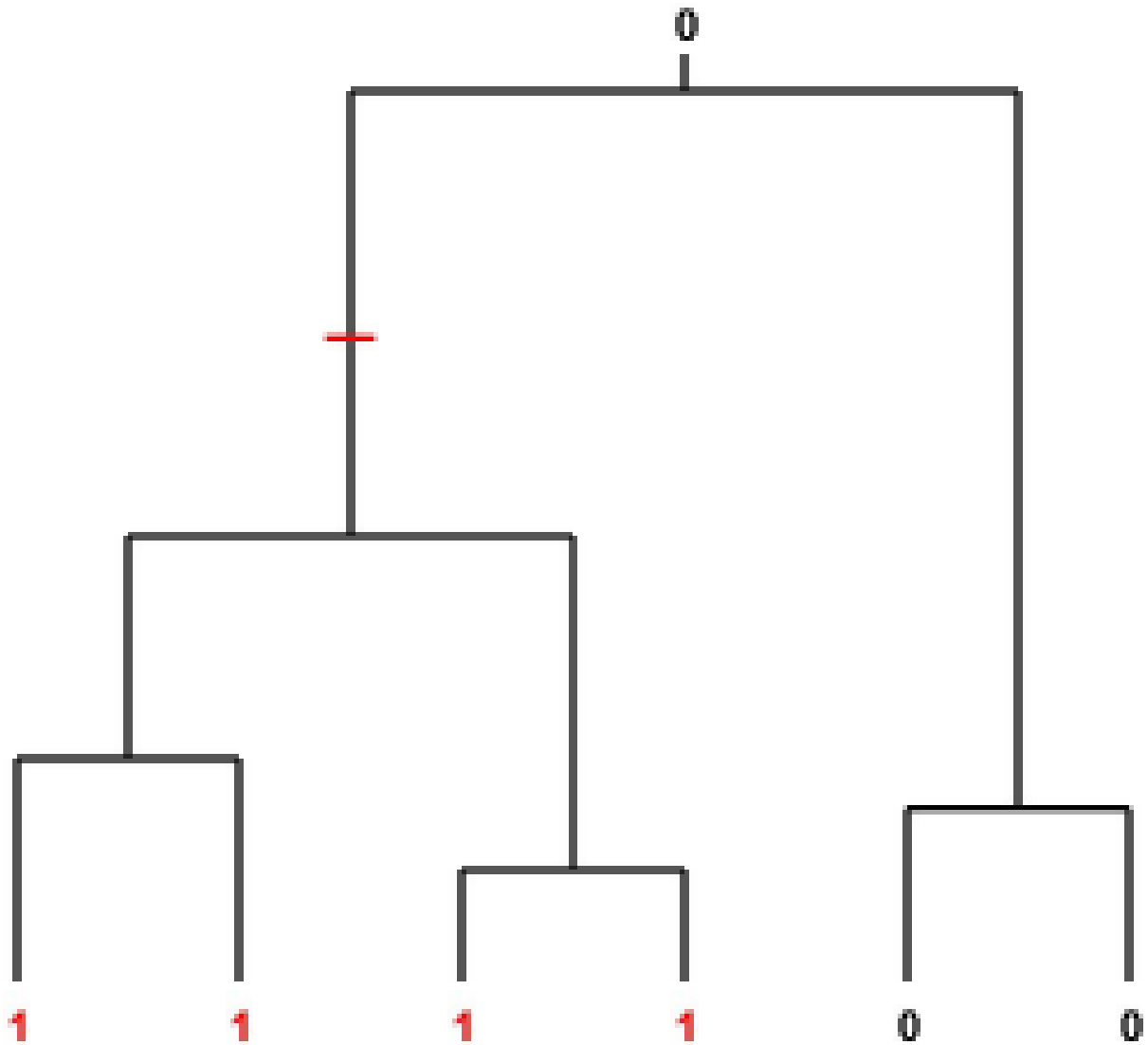
en posant $\theta = 2N\mu$.

- $M(\tau)$ processus de Poisson d'intensité $\frac{\theta}{2}$.
- On peut choisir ensuite le modèle qu'on veut pour décrire ce qui se passe quand une mutation se produit.

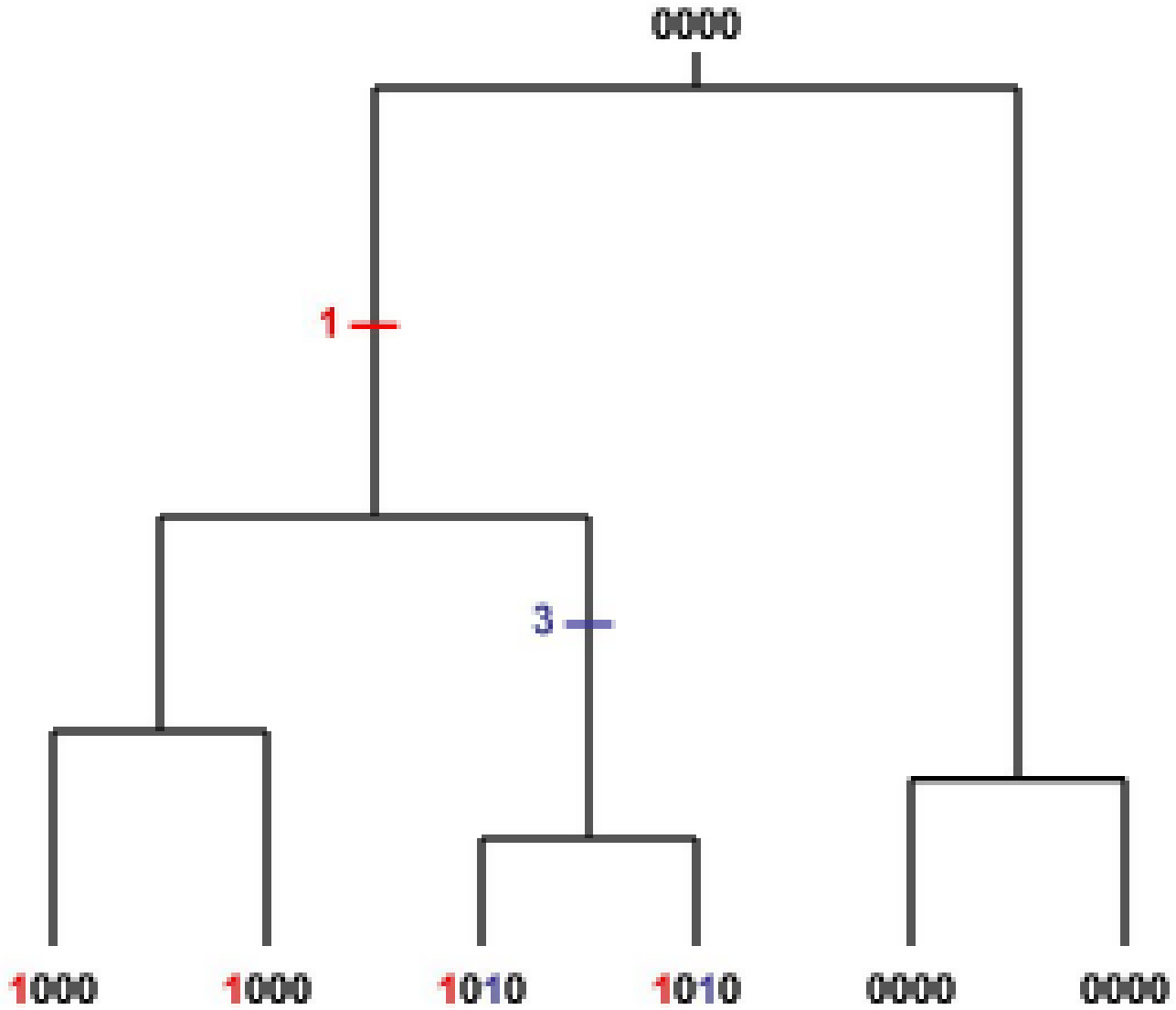
Mutations



Mutations



Infinite site model



Simulation

1. Pour k allant de n à 2 :
   1. Simuler une loi exponentielle de paramètre $\frac{k(k-1)}{2}$.
   2. Choisir uniformément celui des $\frac{k(k-1)}{2}$ couples d'haplotypes qui coalesce.
2. Placer les mutations indépendamment sur chaque branche selon un processus de Poisson d'intensité $\frac{\theta}{2}$.

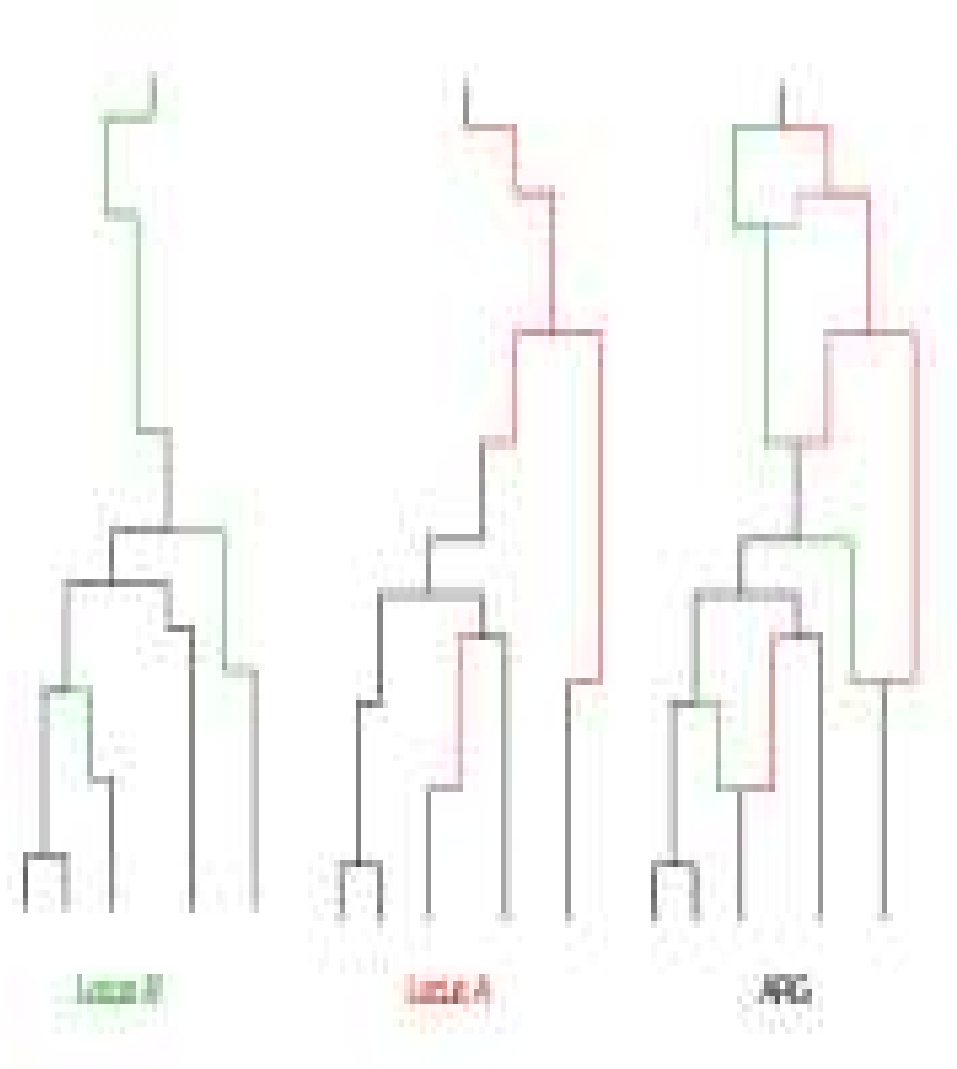
Beaucoup plus rapide que de simuler la population en forward!

Outline

Principes

- Un haplotype créé par recombinaison n'a plus un seul mais deux parents.
- En remontant dans le temps, une branche peut donc se diviser en deux. Cela se produit avec une proba c par génération, c taux de recombinaison entre les loci.
- A l'échelle $\tau = \frac{t}{N}$, et en posant $\rho = 2Nc$, le temps au bout duquel une branche se divise en deux suit une loi exponentielle de paramètre $\frac{\rho}{2}$.
- La généalogie joite des deux locus est appellée "Ancestral Recombination Graph". On peut en déduire les arbres de coalescence marginaux pour chacun des loci (qui sont corrélés).

Exemple



Simulation

$k = n$. Tant que $k > 1$:

1. Simuler U, une loi exponentielle de paramètre $\frac{k(k-1)}{2}$.
2. Simuler V, une loi exponentielle de paramètre $\frac{k\rho}{2}$.
3.
   - Si $U \leq V$, choisir au hasard celui des $\frac{k(k-1)}{2}$ couples d'haplotypes qui coalesce.
   - Si $U > V$, choisir au hasard celle des k lignées qui se sépare.

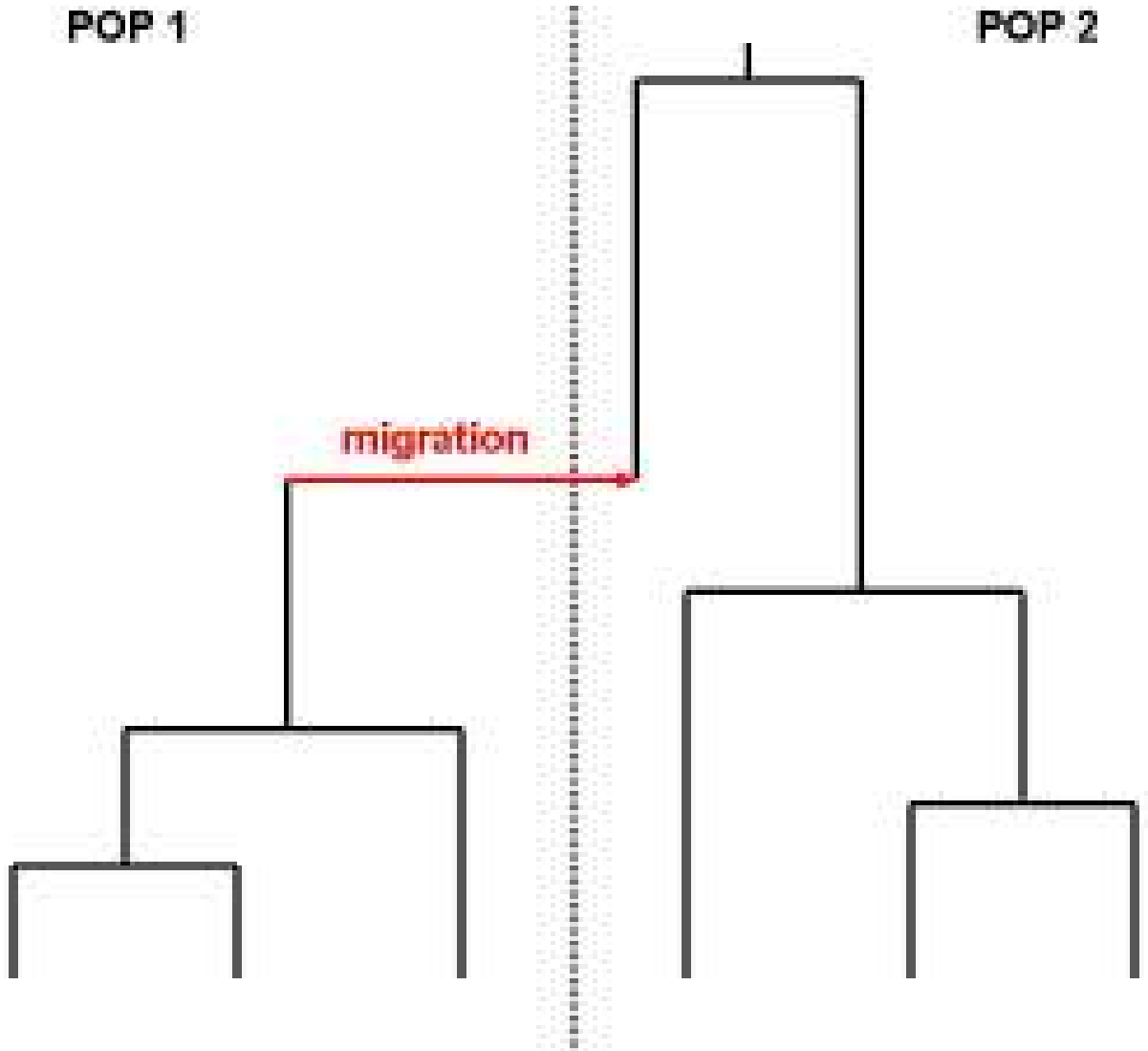
Remarque : le stade $k = 1$ finit toujours par être atteint, mais cela peut être long.

Outline

Population structurée

- Pour chaque noeud de l'arbre, une étiquette indique sa population d'origine.
- Coalescences et recombinaisons intra population. Par ex, pour deux pops 1 et 2, 4 évènements possibles : coalescence dans 1, recombinaison dans 1, coalescence dans 2, recombinaison dans 2.
- Taille relative des pops importante car détermine la vitesse des évènements.
- Si migrations entre populations avec une proba par génération en $O(\frac{1}{N})$, évènements "migration de 1 vers 2" et "migration de 2 vers 1" possibles.

Population structurée



Sélection

- La distribution du nombre d'enfants dépend du type allélique (fitness) $\rightarrow$ séparation généalogie / type allélique impossible.
- Solution classique : utiliser un coalescent structuré, où une pop = un type allélique.
- "Migration" d'un allèle à un autre par mutation ou recombinaison.
- Nécessite de simuler au préalable l'évolution de la fréquence pour les allèles sous sélection.
- logiciels *Selsim* (Spencer and Coop, 2004), *MSMS* (Ewing and Hermisson, 2010).

Outline

Principe général

- θ paramètre(s) à estimer, $\mathcal{D}$ données observées.
- Calcul de la vraisemblance

$$\mathcal{L}(\theta) = \mathbb{P}(\mathcal{D} \mid \theta) = \mathbb{E}[\mathbb{P}(\mathcal{D} \mid \theta, \mathcal{A})] = \int \mathbb{P}(\mathcal{D} \mid \theta, \mathcal{A}) \mathbb{P}(\mathcal{A} \mid \theta) \, d\mathcal{A}$$

  $\mathcal{A}$ généalogie (coalescence ou ARG) de l'échantillon.
- $\mathbb{P}(\mathcal{D} \mid \theta, \mathcal{A})$ généralement facile à calculer.
- $\mathbb{P}(\mathcal{A} \mid \theta) \, d\mathcal{A}$ non calculable
  $\rightarrow$ évaluation de l'intégrale par simulation : EM, MCMC, importance sampling...
- $\mathcal{A}$ dans un espace de dimension très large $\rightarrow$ **problème numérique difficile**.

Exemple en cartographie génétique

- $\mathcal{D}$ = phénotypes Z + génotypes aux marqueurs G. θ effet du QTL. Vraisemblance sachant l'arbre

$$\mathcal{L}(\theta) = \mathbb{P}(Z \mid \mathcal{A}, \theta, G) = \mathbb{P}(Z \mid \mathcal{A}, \theta)$$

- Modèle à effets fixes. X génotypes au QTL.

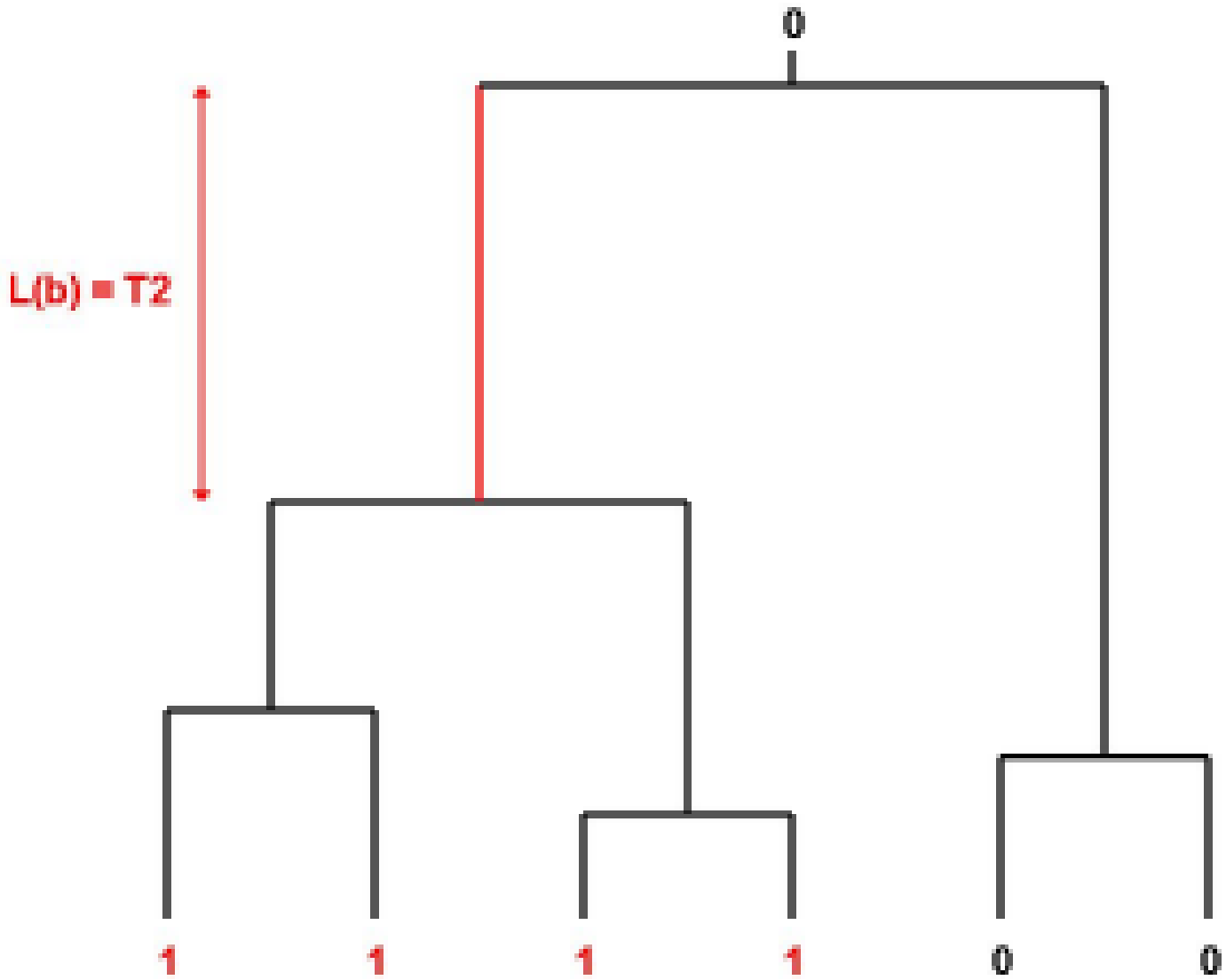
$$\begin{aligned} \mathbb{P}(Z \mid \mathcal{A}, \theta) &= \mathbb{P}(Z \mid X, \theta)\mathbb{P}(X \mid \mathcal{A}) \\ &= (\prod_i \mathbb{P}(Z_i \mid X_i, \theta))\mathbb{P}(X \mid \mathcal{A}) \end{aligned}$$

- $\mathbb{P}(X \mid \mathcal{A})$ déterminée par la position des mutations sur l'arbre. Si mutation unique :

$$\mathbb{P}(X \mid \mathcal{A}) = \sum_{b \in \mathcal{A}} \mathbb{P}(X \mid \text{mutation sur b})(1 - e^{-\frac{\theta L(b)}{2}})$$

$L(b)$ longueur de la branche b.

Exemple en cartographie génétique



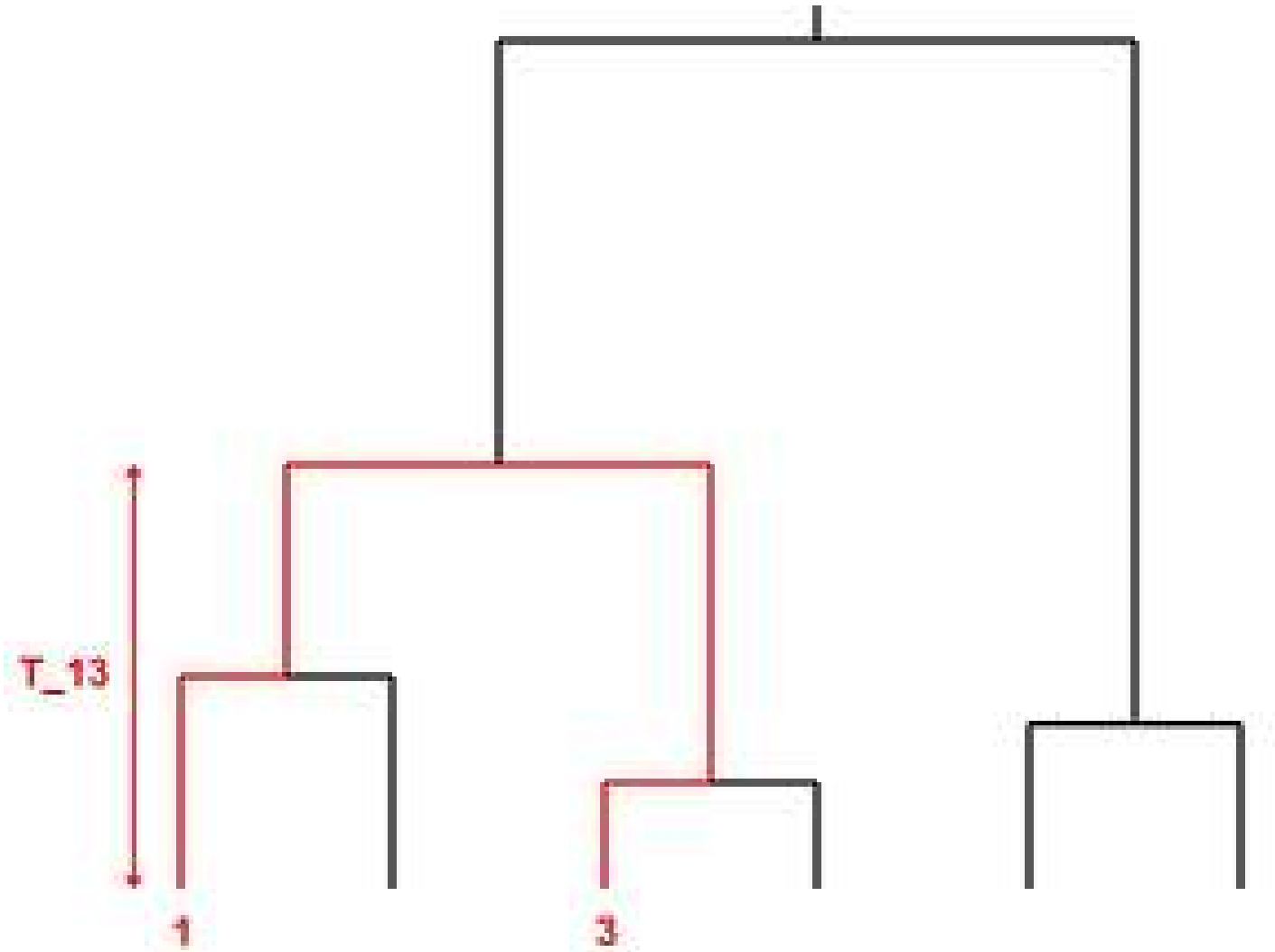
Exemple en cartographie génétique

Modèle mixte. X matrice IBD au QTL (haplotypes)

$$\begin{aligned}\mathbb{P}(X_{i,j}=1 \mid \mathcal{A}) &= \mathbb{P}(\text{pas de mutation sur } b_{i,j}) \\ &= e^{-\frac{\theta L(b_{i,j})}{2}} \\ &= e^{-\theta T_{i,j}}\end{aligned}$$

$b_{i,j}$ arc reliant i et j, $T_{i,j}$ temps de coalescence entre i et j.

Exemple en cartographie génétique



Conclusions

- Outil très performant pour simuler des données génétiques neutres, incluant toutes sortes d'évènements démographiques et de la recombinaison.
- Simuler la sélection est possible, mais pas très naturel. Logiciels limités à un locus sous sélection.
- Utilisation pour estimation de paramètres (en carto ou autre) tentante mais difficultés numériques. Pas très utilisé en carto.

Quelques refs

- Simon Tavaré (2001), *Ancestral inference in molecular biology*. Ecole de Probabilités de Saint-Flour.
- Handbook of statistical genetics, Wiley. Chapitres 22.6, 25, 26, 27.3.
- Zöllner et Pritchard (2004). Coalescent-Based Association Mapping and Fine Mapping of Complex Trait Loci.