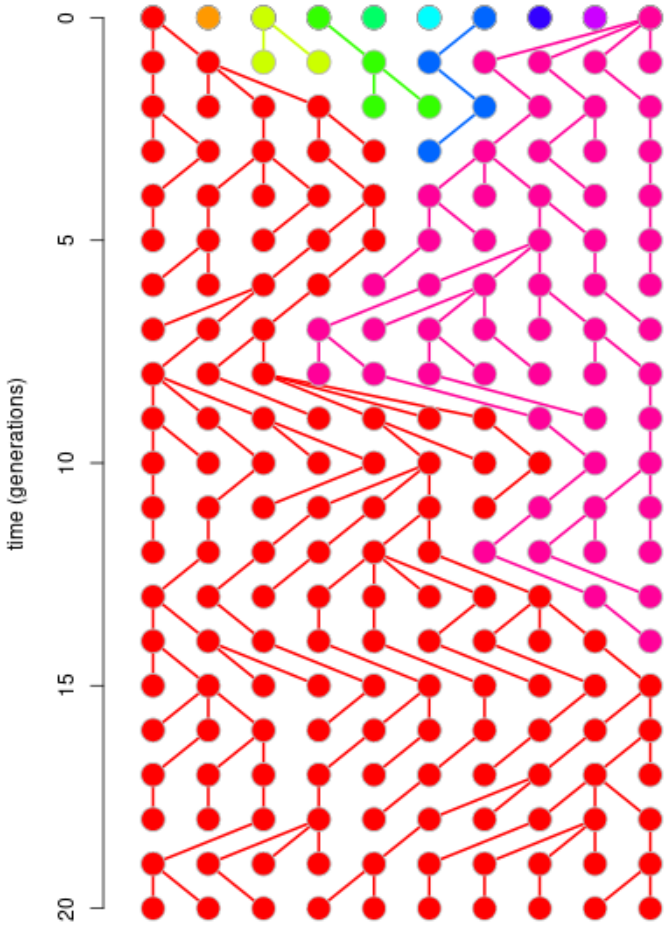
Tema 14. Coalescència

J. Ignacio Lucas Lledó

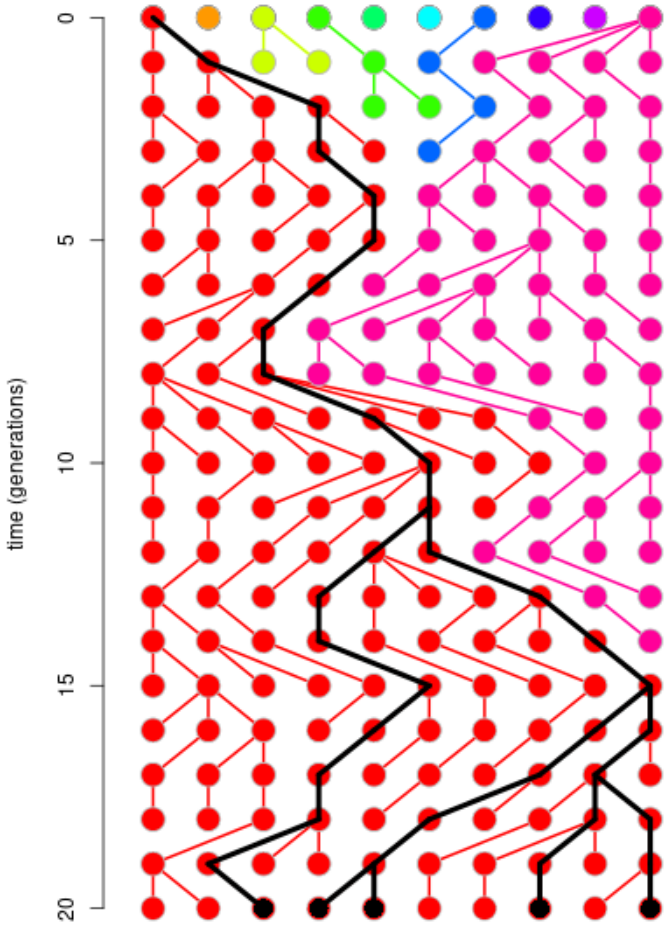
21/4/2022

Població de Fisher-Wright



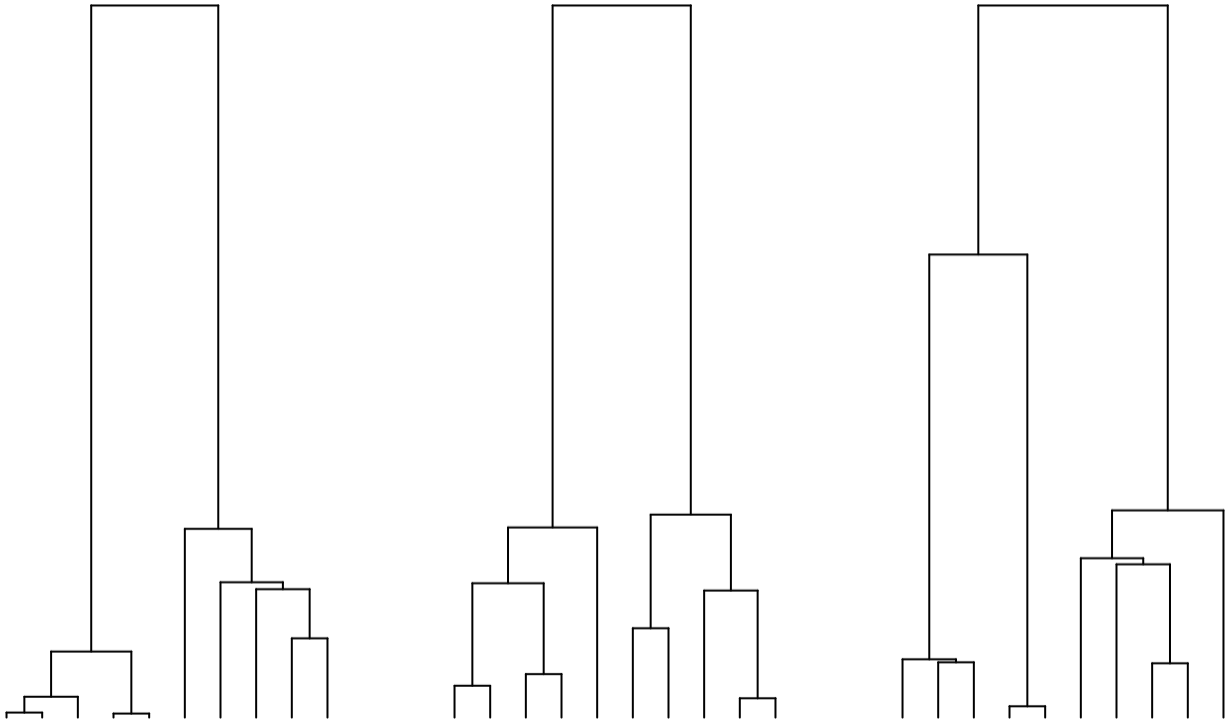
- La mida de la població, *N*, és finita i constant.
- A cada generació tots els individus (haploides) tenen la mateixa probabilitat de deixar descendència.
- Si la població és diploide, cada cercle representa una còpia del mateix gen.

Coalescència



- Seleccionem una **mostra** d'individus de la població **actual**.
- Tracem la seua genealogia *remuntant-nos en el passat*, generació per generació.
- Retrocedint en el temps, els llinatges conflueixen: **coalescència**.
- El **MRCA** és l'ancestre comú **més recent** de la mostra.

Coalescència



El procés *estocàstic* de la coalescència genera arbres amb una froma peculiar: amb una disminució de la *taxa de coalescència* cap al passat.

Coalescència de dues seqüències

El **nombre de generacions** que hem de retrocedir en el passat per trobar l'ancestre comú de **dues** còpies d'un gen en una població de Fisher-Wright, T_2, és una variable aleatòria amb una **distribució geomètrica**:

$$P(T_2 = j) = (1-p)^{j-1}p$$

$P(T_2 = j)$ és la probabilitat de què la coalescència es produïsca exactament *j* generacions en el passat. On *p* és la probabilitat de què la coalescència es produïsca en cada generació. Si la població és **haploide**:

$$p = \frac{1}{N}$$

és la probabilitat de què dues còpies d'un gen tinguen el mateix ancestre en la generació anterior. Per tant:

$$P(T_2 = j) = \left(1 - \frac{1}{N}\right)^{j-1} \frac{1}{N}$$

Coalescència de dues seqüències

Ploidia	Paràmetre	$P(T_2 = j)$	E(T_2)	Var(T_2)
haploide	$1/N$	$\left(1 - \frac{1}{N}\right)^{j-1} \frac{1}{N}$	N	$N(N-1)$
diploide	$1/2N$	$\left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}$	2N	$2N(2N-1)$

En temps continu

La distribució geomètrica és *discreta*: només definida per valors enters de T_2. Es pot aproximar amb la **distribució exponencial** (que és contínua), quan *N* és gran. En haploides:

$$P(T_2 > j) = \left(1 - \frac{1}{N}\right)^j \simeq e^{-\frac{j}{N}}$$

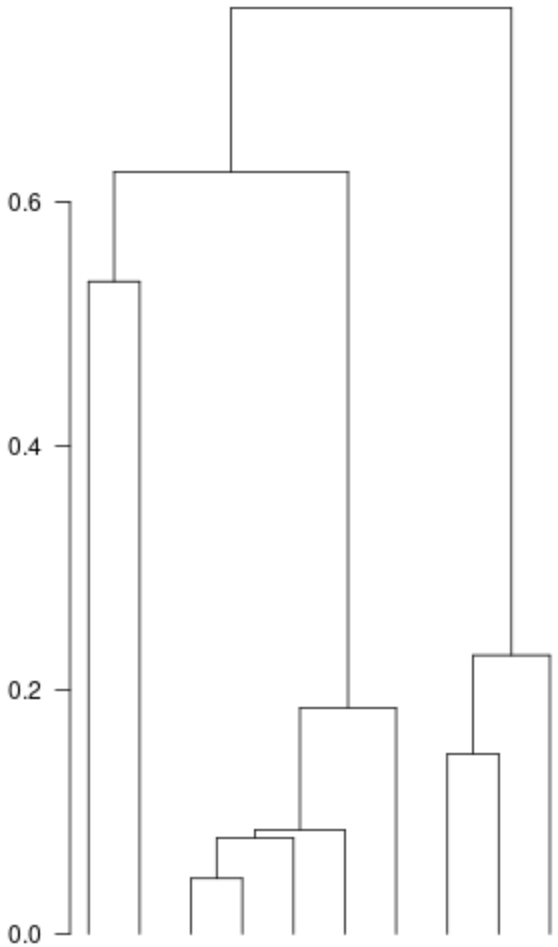
$$P(T_2 \leq j) \simeq 1 - e^{-\frac{j}{N}}$$

En diploides:

$$P(T_2 > j) \simeq e^{-\frac{j}{2N}}$$

$$P(T_2 \leq j) \simeq 1 - e^{-\frac{j}{2N}}$$

En temps continu



- Definim el **temps** continu: $t = \frac{j}{N}$ en haploides, i $t = \frac{j}{2N}$ en diploides.
- Una unitat de temps de coalescència és el nombre de generacions esperat fins la coalescència de dues seqüències: N generacions en haploides, i $2N$ generacions en diploides.

Mida poblacional efectiva

- Les poblacions reals es diferencien de les de Fisher-Wright.
- En lloc d'utilitzar la mida poblacional real, N, utilitzem la **mida poblacional efectiva**, N_e: mida d'una població de Fisher-Wright que experimentaria la mateixa quantitat de deriva que la població real.
- En avant, supose que la població és diploide amb mida efectiva N_e.

Coalescència entre *n* seqüències

Aproximació: màxim d'una coalescència entre dos llinatges per generació.

El temps (en generacions) que hem de retrocedir per trobar **la primera coalescència** entre *n* còpies d'un gen és una variable estocàstica amb distribució geomètrica.

La probabilitat de produir-se una coalescència entre dues còpies qualsevols d'entre una mostra d'*n* còpies d'un gen en la generació immediatament anterior, en una població diploide de mida efectiva N_e és:

$$p = \binom{n}{2} \frac{1}{2N_e} = \frac{n(n-1)}{2} \cdot \frac{1}{2N_e}$$

Per tant:

$$P(T_n = j) = \left(1 - \frac{n(n-1)}{4N_e}\right)^{j-1} \cdot \frac{n(n-1)}{4N_e}$$

Coalescència entre *n* seqüències

En **temps continu**, amb $t = \frac{j}{2N_e}$:

$$P(T_n \leq t) = 1 - e^{-\binom{n}{2}t} = 1 - e^{-\frac{n(n-1)t}{2}}$$

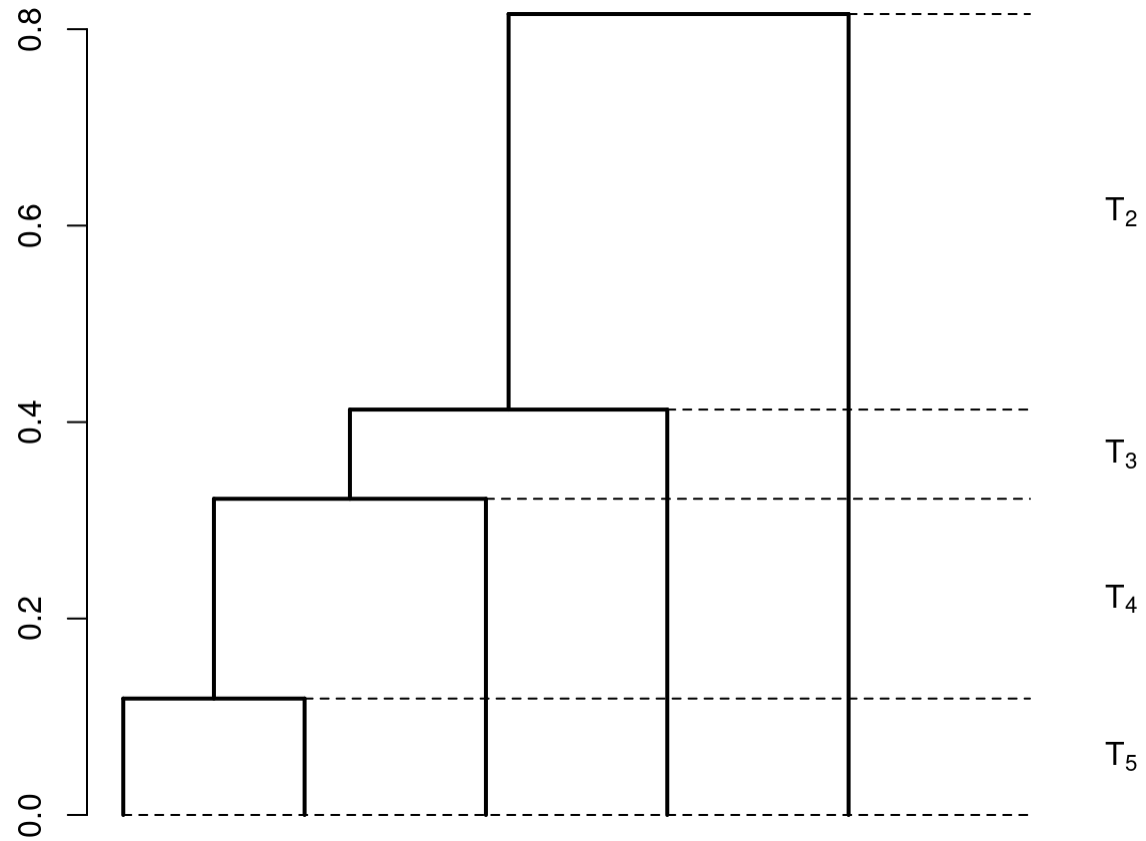
És a dir, el temps que ens hem de remuntar en el passat per tal que la genealogia de n llinatges es condense en $n-1$ llinatges és aproximadament exponencial amb paràmetre $\binom{n}{2}$.

$f(t) = \binom{n}{2} e^{-\binom{n}{2}t}$, per $t \geq 0$.

$F(t) = P(T_n \leq t) = 1 - e^{-\binom{n}{2}t}$, per $t \geq 0$.

$$\mathrm{E}(t) = \frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)} \qquad ; \qquad \mathrm{var}(t) = \frac{1}{\binom{n}{2}^2} = \left(\frac{2}{n(n-1)}\right)^2$$

Coalescència entre *n* seqüències



Coalescència entre *n* seqüències

La mitjana de l'alçada total de l'arbre de coalescència entre n seqüències homòlogues és:

$$\mathrm{E}(H_n) = \sum_{k=2}^{n} \mathrm{E}(T_k) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right) \approx 2$$

(Veure l'apèndix). En unitats de $2N_e$ generacions en poblacions diploides, $\mathrm{E}(H_n) \approx 4N_e$. Mentre que en poblacions haploides, 2 unitats de temps són $2N_e$ generacions.

$$\mathrm{var}(H_n) = \sum_{k=2}^{n} \mathrm{var}(T_n) = 4\sum_{k=2}^{n} \frac{1}{k^2(k-1)^2}$$

Apèndix

Per què $\sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right)$?

$$\sum_{k=2}^{n} \frac{2}{k(k-1)} = 2 \cdot \sum_{k=2}^{n} \frac{1}{k(k-1)}$$

$$= 2 \cdot \sum_{k=2}^{n} \frac{k-(k-1)}{k(k-1)}$$

$$= 2 \cdot \sum_{k=2}^{n} \left(\frac{k}{k(k-1)} - \frac{k-1}{k(k-1)} \right)$$

$$= 2 \cdot \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right)$$

$$= 2 \cdot \left(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} + \ldots + \frac{1}{n-1} - \frac{1}{n}\right)$$

$$= 2 \cdot \left(1 - \frac{1}{n}\right)$$