Projet 1: Génétique des populations (Wright-Fisher)

1 But du programme

Simuler l'évolution temporelle d'une population selon le modèle de Wright-Fisher. Une population est représentée par N individus, chacun possède un génome. Dans la version la plus simple, il existe un nombre fixe A d'allèles (nombre de génotypes différents) et la population se résume donc au nombre de copies de chaque allèle.

A chaque génération, une nouvelle population de taille N est tirée au sort parmi les individus de la génération précédente (tirage aléatoire avec remise): si au temps t les effectifs des allèles $1, \ldots, A$ sont $n_1, \ldots, n_A$ alors au temps $t+1$ ils seront donnés par les nombres aléatoires $k_1, \ldots, k_A$ distribués selon la distribution multinomiale définie par

$$\begin{aligned} P(k_1, \ldots, k_A \,|\, n_1, \ldots, n_A, N) &= \frac{N!}{k_1! \cdots k_A!} \left(\frac{n_1}{N}\right)^{k_1} \cdots \left(\frac{n_A}{N}\right)^{k_A} , \\ N &= \sum_{i=1}^{A} n_i = \sum_{i=1}^{A} k_i . \end{aligned}$$

Cette distribution peut être générée par une succession de distributions binomiales:

$$\begin{aligned} k_1 &\sim \text{Binom}\left(N, \frac{n_1}{N}\right) , \\ k_2 &\sim \text{Binom}\left(N - k_1, \frac{n_2}{N - n_1}\right) , \\ k_3 &\sim \text{Binom}\left(N - k_1 - k_2, \frac{n_3}{N - n_1 - n_2}\right) , \\ &\cdots \\ k_A &= N - \sum_{i=1}^{A-1} k_i , \end{aligned}$$

où $\sim$ signifie "est un nombre aléatoire de distribution...".

2 Paramètres, formats de fichiers

Les paramètres suivants sont librements choisis par l'utilisateur:

- N: la taille de la population
- T: la durée (en nombre de générations) de la simulation
- A: le nombre d'allèles dans la population

- $f_1, \ldots, f_A$: les fréquences initiales des allèles (en fraction de la population)
- R: la simulation sera répétée R fois (avec les mêmes paramètres mais des nombres aléatoires différents) pour produire des statistiques.

Des valeurs par défaut peuvent être prévues (correspondant à une simulation rapide), par ex. $N = 100$, $A = 2$, $T = 10$, $R = 2$ et $f_1 = 0.8$.

A la place de $N, A, f_1, \ldots f_A$ on peut spécifier les données suivantes:

- un fichier fasta contenant une séquence génomique pour chaque individu
- une liste de "marqueurs" $m_1, \ldots, m_L$: des positions le long de la séquence qui déterminent les allèles

On en déduit N (nombre de séquences dans le fichier), A (nombre de combinaisons de nucléotides observées aux positions spécifiées) et les fréquences observées. Par exemple, si les séquences sont

```
ACCTAGTG
ACGTAGTG
TCGTACTG
TCGTACTG
```

et les marqueurs sont $1, 3, 6$, on en déduit que $N = 4$, les allèles sont `ACG`, `AGG`, `TGC` et les fréquences initiales $1/4, 1/4, 1/2$.

Les valeurs typiques pour une simulation intéressante sont

$$N = 100, 1000, 10000\ , \quad T = 3000\ , \quad A = 2, 4\ , \quad f_1 = 0.5, 0.6, \ldots, 0.9\ , \quad R = 500\ .$$

Il faut éviter de stocker en mémoire toutes les générations jusqu'à la fin de la simulation et donc imprimer les résultats au fur et à mesure de l'avancement de la simulation.

3 Résultats

Le résultat standard (dans un fichier texte et/ou à l'écran) doit être formaté ainsi:

0	0.1|0.2|0.3|0.4	0.1|0.2|0.3|0.4	0.1|0.2|0.3|0.4
1	0.096|0.21|0.274|0.42	0.087|0.206|0.298|0.409	0.098|0.207|0.311|0.384
2	0.075|0.208|0.279|0.438	0.086|0.208|0.307|0.399	0.087|0.214|0.307|0.392
3	0.067|0.2|0.284|0.449	0.074|0.193|0.333|0.4	0.096|0.215|0.288|0.401
4	0.067|0.208|0.287|0.438	0.088|0.196|0.311|0.405	0.106|0.222|0.279|0.393
5	0.074|0.219|0.276|0.431	0.099|0.214|0.31|0.377	0.089|0.24|0.27|0.401
6	0.079|0.204|0.296|0.421	0.107|0.236|0.283|0.374	0.087|0.22|0.283|0.41
7	0.071|0.2|0.291|0.438	0.108|0.236|0.293|0.363	0.082|0.237|0.262|0.419
8	0.071|0.224|0.253|0.452	0.103|0.207|0.299|0.391	0.078|0.259|0.236|0.427
9	0.073|0.208|0.275|0.444	0.103|0.197|0.299|0.401	0.072|0.24|0.239|0.449
10	0.068|0.223|0.243|0.466	0.111|0.198|0.298|0.393	0.075|0.23|0.244|0.451
	ACG|ACA|ATA|ATG	ACT|ACA|AGA|GCA	ACG|AGA|AGT|AAA

Chaque ligne correspond à une génération, la première colonne est le numéro de la génération (le temps), chaque colonne suivante représente une réplique de la simulation où chacune est représentée par les fréquences des allèles dans la population (séparées par '|'). Si un fichier fasta a été fourni en entrée, la dernière ligne indique les allèles calculés. Les colonnes sont séparées par des tabulations '$\backslash t$'.

4 Tests

Un propriété importante que l'on peut tester est que sur R réplicats de la même simulation, les fréquences seront en moyenne conservées d'une génération à la suivante: la moyenne (sur les différentes simulations) de $f_i(t)$ est égale à $f_i(0)$ pour tous les temps $t > 0$.