université PARIS-SACLAY

TP3

Exercice n°1 : Modélisation pour la génétique des populations

La génétique des populations est l'étude de la distribution et des changements des fréquence allèliques dans des populations. Ces données se présentent sous la forme d'une matrice notée Y indiquant le nombre de fois où une allèle donnée a été observée pour un individu donné $i = 1, \dots, n$ et un locus donné $j = 1, \dots, L$ où n est le nombre d'individus étudiés, et L le nombre de locus étudiés.

On propose deux modèles : un modèle "simple" (M1) qui suppose que chaque individu appartient à une population parmi K populations distinctes possibles, où K est un entier non nul et un second modèle (M2) qui suppose que le génotype de chaque individu provient d'une population distincte pour chaque locus du génome.

Modèle M1 : On suppose d'abord que chaque individu appartient à une seule population et on pose le modèle suivant :

$$(Z_i)_i \text{ i.i.d. } \sim \mathcal{M}(1, (\pi_1, \dots, \pi_K))$$

$$(Y_{ij})_{ij} \text{ indép. } | (Z_i = k) \sim \mathcal{M}(1, G_{kj})$$

où G_{kj} est le vecteur des fréquences alléliques pour le locus j dans la population k.

Modèle M2 : On suppose que le génotype de chaque individu provient d'une population différente à chaque locus. Cependant, chaque individu est associé à une tendance préférentielle caractérisée par une variable latente Q_i. Cette variable peut être interprétée comme la position de l'individu i dans le simplexe de $\mathbb{R}^K$, où les sommets du simplexe correspondraient à des individus purs issus de chaque population. On note $\mathcal{D}$ la loi de Dirichlet (voir la fonction `rdirichlet` sous R). Le modèle s'écrit sous la forme suivante :

$$(Q_i)_i \text{ i.i.d. } \sim \mathcal{D}(1, \alpha)$$

$$(S_{ij})_{ij} \text{ indép. } | (Q_i) \sim \mathcal{M}(1, Q_i)$$

$$(Y_{ij})_{ij} \text{ indép. } | (S_{ij}) \sim \mathcal{M}(1, G_{S_{ij}})$$

1. Pour une espèce diploïde, quelles sont les valeurs possibles prises par Y_{ij} pour un individu i donné et un locus j donné ?
2. Dans le **modèle M1**, donner le type (vecteur, matrice, ...) et la taille des objets **Z**, **Y**, π et **G** en fonction de n, K, L et en supposant que l'espèce est diploïde.
3. Simuler une matrice de marqueurs Y à l'aide du **modèle M1**. Recopier le code permettant d'effectuer cette simulation de données sur votre copie. On pourra prendre 100 individus, 5 locus, 2 populations et supposer qu'il s'agit d'une espèce diploïde.
4. Dans le **modèle M2**, donner le type (vecteur, matrice, ...) et la taille des objets **Q**, **S**, **Y**, α et **G** en fonction de n, K, L et en supposant que l'espèce est diploïde.
5. Simuler une matrice de marqueurs Y à l'aide du **modèle M2**. Recopier sur votre copie le code permettant d'effectuer cette simulation de données. On pourra prendre 100 individus, 5 locus, 2 populations et supposer qu'il s'agit d'une espèce diploïde.
6. Dans les modèles M1 et M2, chaque locus est supposé indépendant des autres locus. Cette hypothèse est-elle réaliste ? Justifier votre réponse.