Estimation bayésienne de fréquences alléliques dans un modèle de métapopulation

Emily Walker, Julien Papaix, Etienne Klein, Olivier Bonnefon, Pierre Franck

Estimation bayésienne de fréquences alléliques dans un modèle de métapopulation

Emily Walker[1], Julien Papaïx[2], Pierre Franck[3] & Etienne Klein[4]

[1] *INRA BioSP, 228 route de l'aérodrome, 84914 Avignon, France, Emily.Walker@inra.fr*
[2] *INRA BioSP, 228 route de l'aérodrome, 84914 Avignon, France, Julien.Papaix@inra.fr*
[3] *INRA PSH, 228 route de l'aérodrome, 84914 Avignon, France, Pierre.Franck@inra.fr*
[4] *INRA BioSP, 228 route de l'aérodrome, 84914 Avignon, France, Etienne.Klein@inra.fr*

Résumé. Ce travail consiste à étudier la structure génétique d'une métapopulation et la dispersion d'individus entre les sous-populations en estimant l'hétérogénéité des fréquences alléliques au sein de la métapopulation à partir des génotypes collectés dans plusieurs sous-populations. Un modèle bayésien "multinomial-Dirichlet" a été construit et implémenté sur JAGS. Ainsi, le nombre de copies d'allèles calculé à partir des génotypes suit une loi multinomiale dont le vecteur de probabilités est les fréquences alléliques des sous-populations. Ces fréquences alléliques sont tirées dans une loi de Dirichlet dont les paramètres sont le produit des fréquences alléliques d'une population consensus (de la métapopulation) et d'un paramètre lié à un indice de différenciation génétique de la sous-population (*F-model*, Gaggiotti and Foll 2010). Ce modèle a été appliqué à des données de marqueurs microsatellites d'un insecte ravageur des pommiers (*Cydia pomonella*) échantillonnées dans plusieurs vergers de Basse Vallée de la Durance.

Mots-clés. estimation bayésienne, modèle hiérarchique, multinomial-Dirichlet, métapopulation, structure génétique, dispersion d'insectes

Abstract. This work aims at studying the meta-population genetic structure along with the dispersal ability of individuals among local populations by estimating the heterogeneity of allelic frequencies based on "multinomial-Dirichlet" distributions, and implemented on the JAGS software (R package rjags). The number of copies of the different alleles are sampled in a multinomial distribution with parameters depending on population allelic frequencies. The population allelic frequencies are sampled in a Dirichlet distribution with parameter equal to the product of the metapopulation allelic frequencies and a F–parameter measuring differentiation (*F-model*, Gaggiotti and Foll 2010). This model was applied to an insect pest (*Cydia pomonella*), collected in several orchards in South of France at landscape scale.

Keywords. Bayesian statistics, hierarchical model, multinomial-Dirichlet, metapopulation, genetic structure, insect pests dispersion

1 Structure génétique d'une métapopulation

L'étude de la dispersion d'individus peut être appréhendée à l'échelle d'un paysage par les dynamiques démographiques et la structure génétique des populations. Nous étudions ici la structure génétique et la dispersion d'un insecte ravageur des pommiers (carpocapse du pommier – *Cydia Pomonella*).

Les carpocapses ont été collectés sur un ensemble de vergers de la basse vallée de la Durance, dont le suivi est assuré par l'INRA depuis plusieurs années. Dans ce travail, 995 individus collectés en 2006 ont été analysés sur 20 marqueurs microsatellites. Les vergers sont considérés comme des populations sur lesquelles les fréquences alléliques ont été calculées. L'idée est ensuite d'estimer les fréquences alléliques d'une population dite consensus dont les sous-populations (vergers) seraient issues, à un paramètre de variabilité près. Ce paramètre dépend d'une mesure de différenciation génétique (Fst, Holsinger et Weir (1999)).

La différenciation génétique à l'échelle de la métapopulation (Wright's F-statistics Fst global calculé) est de 5 %. Nous disposons de données provenant de 51 vergers mais près de la moitié des vergers comporte peu d'individus (≤ 10).

2 Modèle d'estimation bayésienne : *F-model*

Le modèle mis en œuvre pour l'estimation des fréquences alléliques de la population consensus est un modèle multinomial-Dirichlet. Ce type de modèle est communément utilisé en statistique bayésienne, mais pose un certain nombre de problèmes lors de l'implémentation que nous aborderons en discussion. Dans le contexte de modèle de métapopulation (Holsinger 1999), un modèle nommé *F-model* a été proposé par Balding (2003), Falush et al. (2003) et Gaggiotti et Foll (2010). Dans le modèle, les Fst spécifiques à chaque sous-population, c'est-à-dire les probabilités que deux gènes choisis au hasard dans la sous-population aient un ancêtre commun en excluant toute immigration ou colonisation, sont estimés. Cette définition permet de considérer le cas de tailles de population et de taux de migration différents selon les populations (Balding 2003).

Soit $v = 1, .., V$ les vergers (V=51), $l = 1, .., L$ les locus (avec L=20), $a_l = 1, .., A_l$ les allèles au locus l. Soit $N_{v,l}$ le nombre total de copies d'allèles échantillonnés dans le verger v au locus l. Les fréquences allèliques consensus sont notées $FAcons_{l,a}$, les fréquences alléliques de chaque verger $FA_{v,l,a}$, et le nombre de copies de l'allèle a calculé sur les génotypes observés $Ncopies_{v,l,a}$.

On pose :

$$Ncopies_{v,l,a} \sim Multinomial(FA_{v,l,a}, N_{v,l})$$

$$[FA_{v,l,a}] \sim Dirichlet([\theta_v\, FAcons_{l,a}])$$

$$[FAcons_{l,a}] \sim Dirichlet([\alpha_l])$$

$$\theta_v = \frac{1}{Fst_v} - 1$$

Les lois a priori sont définies comme :

$$Fst_v \sim Normale(\mu, \tau)$$

Les Fst suivent une loi normale tronquée sur $[0, 0.1]$ dont le paramètre μ suit une loi normale de paramètres 0 et 0.1, Dans un premier temps les α_l sont fixés à 1.

La vraisemblance est de la forme :

$$L(FAcons_{l,a}, Fst_v) = \prod_{v=1}^{V} \prod_{l=1}^{L} P(FAcons_{v,l,a} | FA_{v,l,a}, Fst_v)$$

Et la loi a posteriori s'écrit :

$$\pi(FAcons, Fst | Ncopies) \propto L(FAcons, Fst)\pi(FAcons)\pi(Fst)$$

Le modèle a tourné sur JAGS (Just Another Gibbs Sampler), avec le package R rjags (Plummer 2003). Plusieurs chaînes ont convergé, mais ce type de modèle pose des problèmes d'instabilité. Des modifications dans le choix des *a priori* peuvent aider à résoudre ces problèmes. Il est aussi envisagé de tester ce modèle avec d'autres algorithmes d'estimation (Hamiltonien Monte Carlo par exemple).

Bibliographie

[1] Balding, D. J. (2003), Likelihood-based inference for genetic correlation coefficients, Theoretical population biology, 63(3), 221–230.
[2] Gaggiotti, O. E., and Foll, M. (2010), Quantifying population structure using the F–mode, Molecular Ecology Resources, 10(5), 821–830.
[3] Holsinger, K. E. (1999), Analysis of genetic diversity in geographically structured populations : a Bayesian perspective, Hereditas, 130(3), 245–255.
[4] Plummer, M. (2003), JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing, 124,125.