JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

AGNÈS GRIMAUD
SYLVIE HUET
HERVÉ MONOD
ERIC JENCZEWSKI
FRÉDÉRIQUE EBER

Mélange de modèles mixtes : application à l'analyse des appariements de chromosomes chez des haploïdes de colza

Journal de la société française de statistique, tome 143, n° 1-2 (2002), p. 147-153

<http://www.numdam.org/item?id=JSFS_2002__143_1-2_147_0>

MÉLANGE DE MODÈLES MIXTES : APPLICATION À L'ANALYSE DES APPARIEMENTS DE CHROMOSOMES CHEZ DES HAPLOÏDES DE COLZA

Agnès GRIMAUD [1], Sylvie HUET [1], Hervé MONOD [1,*], Eric JENCZEWSKI [2], Frédérique EBER [2]

RÉSUMÉ

Nous décrivons une expérimentation dont l'objectif est d'étudier le déterminisme génétique des appariements de chromosomes chez des haploïdes de colza, au cours de la méiose. Le modèle adapté aux données observées est un modèle de mélange, dont chacun des deux composants suit un modèle mixte. Les paramètres sont estimés par maximum de vraisemblance, calculé avec un algorithme ECM. Le test du rapport de vraisemblance est présenté pour deux hypothèses sur les paramètres, dont l'une inclut la nullité d'un paramètre de variance.

ABSTRACT

The experiment described in this paper aimed at studying the genetic determinism that controls the pairing of chromosomes in oilseed rape haploids during meiosis. The oberved data are modelled using a mixture model with two components, each component being a mixed linear model. The parameters are estimated by maximising the likelihood using an algorithm based on an ECM method. Two hypotheses on the parameters are tested using the likelihood ratio testing procedure. One of these testing hypotheses is non standard because it assumes that one of the variance parameters equals zero.

1. Introduction

Des modèles de mélange apparaissent fréquemment en génétique, lorsque l'on analyse des données issues de croisements entre des parents différents génétiquement (Loisel *et al.*, 1994). Les données issues de l'expérience décrite ci-dessous présentent une structure de covariance qui est prise en compte à l'aide d'un modèle de mélange de deux modèles mixtes. Nous montrons comment les méthodes d'analyse du modèle mixte s'adaptent au cas d'un modèle de mélange.

1. INRA, Unité de Biométrie, 78352 Jouy en Josas Cedex, France
* Adresse de l'auteur à contacter : Herve.Monod@jouy.inra.fr
2. INRA, Unité AdPBV, BP35327, 35653 Le Rheu Cedex, France

2. Présentation des données

Lors de la méiose de colzas haploïdes (ne contenant qu'une copie du génôme), un certain nombre de chromosomes s'apparient et les autres, dits univalents, restent non appariés. Le nombre d'univalents est variable et dépend en particulier de la variété de colza. L'objectif est de savoir si le contrôle des appariements de chromosomes est dû à l'action d'un gène unique ou non.

Les haploïdes issus de la variété Darmor ont un fort taux d'appariement (donc peu d'univalents), alors que ceux issus de la variété Yudal ont un faible taux d'appariement. Ces deux variétés de colza ont été croisées et des haploïdes ont été produits à partir des variétés parents et à partir des hybrides de première génération, dits F1. Des comptages de chromosomes univalents ont été réalisés, en quatre lots d'observations, sur des cellules de plantes haploïdes issues des variétés parents (de 15 à 50 cellules prélevées par individu, en moyenne 20) et sur des cellules de plantes haploïdes produites à partir des F1 (de 14 à 149 cellules prélevées par individu, en moyenne 20). La répartition des données entre les différents lots d'observation est résumée dans le Tableau 1.

TABLEAU 1. – Résumé des données disponibles.

lot	Nombre d'haploïdes (nombre de cellules)		
	Yudal	Darmor	F1
1	0 (0)	0 (0)	55 (1611)
2	10 (193)	20 (411)	109 (2208)
3	0 (0)	0 (0)	35 (688)
4	3 (124)	7 (182)	45 (1011)

La distribution bi-modale du nombre moyen d'univalents, observée chez les haploïdes issus des hybrides F1 (Fig.1), suggère l'existence d'un gène majeur ayant une influence prépondérante sur le taux d'appariement des chromosomes et présentant des formes alléliques différentes dans les variétés Darmor et Yudal. En effet, sous l'hypothèse d'un gène majeur unique, et dans la mesure où l'on observe des individus haploïdes (ne portant donc qu'un seul allèle au gène majeur), on attend bien un mélange de deux populations dans la descendance, l'une constituée d'haploïdes présentant l'allèle Darmor au gène majeur et l'autre l'allèle Yudal.

L'objectif est de savoir si le taux des appariements est sous le contrôle d'un gène majeur unique ou s'il dépend également d'autres gènes. Sous l'hypothèse d'un gène majeur unique, on s'attend à observer les deux phénomènes suivants : l'équiprobabilité des deux types de comportement méiotique dans la descendance, et l'identité des distributions d'univalents des deux populations d'haploïdes issus de F1 aux deux distributions d'haploïdes issues des variétés parents. Sous l'hypothèse où le taux d'appariement est contrôlé par plusieurs

gènes, de nombreuses configurations génétiques peuvent être envisagées selon le nombre de gènes impliqués, leurs distances sur le génôme, leurs effets relatifs et leurs interactions éventuelles. Quelle que soit l'hypothèse génétique envisagée, on s'attend à ce que la présence d'autres gènes (d'effet inférieur et ségrégeant indépendamment du gène majeur), d'une part modifie la moyenne des distributions observées chez les haploïdes issus de plantes F1 par rapport aux variétés parentales, et d'autre part engendre une variabilité entre plantes haploïdes issues de plantes F1.

Nous présentons dans la section suivante un modèle qui s'appuie sur cette interprétation des données, et dont l'objectif est de préciser l'influence du gène majeur, et de tester s'il existe un effet lié à la présence d'autres gènes mineurs.

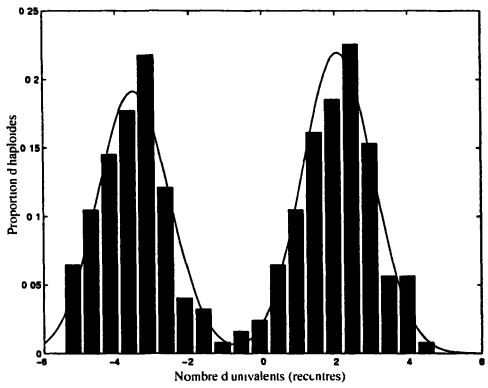

FIG 1. – Histogramme des nombres moyens d'univalents et densité estimée (moyennes par plantes haploïdes issues de F1)

3. Modélisation

3.1. Modèle

Toutes les cellules haploïdes issues d'une même variété parent sont supposées homogènes génétiquement. On note z_{gli_j} la réponse pour la cellule j de l'haploïde i de génotype g (Yudal ou Darmor) et provenant du lot l. Pour ces observations, le modèle est

$$z_{gli_j} = \gamma_l + \mu_g + \varepsilon_{gli_j}$$

où γ_l ($l \in \{2,4\}$) est la moyenne (fixe) du lot l; $\mu_g (g \in \{D,Y\})$ est l'effet variété (fixe); ε_{gli_j} est l'erreur résiduelle, supposée gaussienne centrée et de variance σ_E^2; $i \in \{1,...,n_{gl}\}$ où n_{gl} est le nombre d'haploïdes issus des parents; $j \in \{1,...,m_{gli}\}$ où m_{gli} est le nombre de cellules observées. Les variables ε_{gli_j} sont supposées indépendantes. Pour éviter la surparamétrisation, on définit $\mu = \mu_D = -\mu_Y$.

Les haploïdes issus de plantes F1 se répartissent comme décrit précédemment en deux populations, que nous noterons Pd et Py, associées chacune à l'un des deux allèles du gène majeur. Pour les cellules issues de ces haploïdes, on note y_{lij} la réponse pour la cellule j de la plante i provenant du lot l. Pour ces observations, le modèle est

$$y_{lij} = \gamma_l + a_{li} + b_{li} + \varepsilon_{lij}$$

où γ_l ($l \in \{1,2,3,4\}$) est la moyenne (fixe) du lot l ; a_{li} est l'effet population ; b_{li} est un effet plante lié à l'éventuel « fond génétique » et supposé aléatoire, gaussien, centré de variance σ_H^2, indépendant entre haploïdes mais commun à toutes les cellules issues d'un même haploïde ; ε_{lij} est l'erreur résiduelle, supposée gaussienne centrée et de variance σ_E^2 ; $i \in \{1,...,n_l\}$ où n_l est le nombre d'haploïdes issus de F1 ; $j \in \{1,...,m_{li}\}$ où m_{li} est le nombre de cellules observées. Les $\varepsilon_{lij}, a_{li}$ et b_{li} sont supposés indépendants entre eux et indépendants des ε_{glij}.

Les différents lots correspondent à des conditions expérimentales différentes : les haploïdes ont été produits à des dates différentes et on sait que des variations environnementales, comme la température par exemple, influent sur le taux d'appariements.

L'allèle du gène majeur porté par les haploïdes issus de plantes F1 étant inconnu, l'effet population a_{li} est une variable aléatoire de Bernoulli qui prend deux valeurs possibles μ_{Pd} et μ_{Py} avec probabilités p et $1-p$. Afin d'assurer l'identifiabilité des paramètres du modèle, on pose $\mu_{\mathrm{Py}} = \mu_{\mathrm{Pd}} + \alpha$ et on suppose que α est strictement positif.

Ainsi, conditionnellement à la variable a_{li}, le modèle comprend des facteurs à effets fixes (lot, population, variété) et un facteur à effets aléatoires (plante issue de F1). On a donc un modèle mixte avec mélange, le mélange portant ici sur les effets du facteur population.

3.2. Expression de la log-vraisemblance

On note $\theta = (p, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \mu, \mu_{\mathrm{Pd}}, \alpha, \sigma_E^2, \sigma_H^2)$ le vecteur des paramètres.

Les données sont indépendantes entre haploïdes, dépendantes entre cellules d'un même haploïde. La log-vraisemblance, notée $L_n(\theta)$ s'exprime donc comme la somme des log-vraisemblances associées aux vecteurs d'observations de l'ensemble des cellules d'un même haploïde, notés Z_{gli} et Y_{li}.

$$L_n(\theta) = \sum_{l=1}^{4} \sum_{i=1}^{n_l} \log\left[p f(Y_{li}, \gamma_l + \mu_{\mathrm{Pd}}, V_{li}) + (1-p) f(Y_{li}, \gamma_l + \mu_{\mathrm{Pd}} + \alpha, V_{li})\right]$$
$$+ \sum_{g \in \{D,Y\}} \sum_{l \in \{2,4\}} \sum_{i=1}^{n_{gl}} \log\left[f(Z_{gli}, \gamma_l + \mu_g, \sigma_E^2 I_{m_{gli}})\right]$$

où $V_{li} = \sigma_H^2 J_{m_{li}} + \sigma_E^2 I_{m_{li}}$; I_k désigne la matrice identité d'ordre k ; J_k désigne la matrice $k \times k$ dont tous les éléments sont égaux à 1 ; $f(Y, \varphi, V)$ représente la densité d'un vecteur gaussien d'espérance φ et de matrice de variance V.

4. Estimation des paramètres par maximum de vraisemblance

Dans notre modèle, on peut considérer que le vecteur des données est constitué de deux composantes : le vecteur des données observées (les nombres d'univalents par cellule) et le vecteur des données manquantes (l'allèle du gène majeur porté par les haploïdes issus de F1).

L'estimateur du maximum de vraisemblance dans le cas d'un modèle de mélange se calcule en général à l'aide d'un algorithme EM (Expectation-Maximisation, Dempster *et al.*, 1977), en exploitant la décomposition en données observées et données manquantes que nous venons de décrire. Il s'agit d'un algorithme itératif, chaque itération comportant deux étapes : une étape, dite étape E, où l'on calcule l'espérance de la log-vraisemblance conditionnellement aux observations et aux valeurs courantes des paramètres, notée $Q(\theta)$, et une étape, dite étape M, de maximisation de la fonction Q.

Dans notre cas, la structure de modèle mixte nous a conduit à utiliser un algorithme ECM (Expectation-Conditional Maximisation, Meng et Rubin, 1993). Il s'agit également d'un algorithme itératif avec, à chaque itération, une étape E comme pour l'algorithme EM suivie de plusieurs étapes CM à la place de l'étape M de l'algorithme EM.

Pratiquement, nous considérons trois étapes CM. À l'itération t, la première étape CM consiste à calculer p et η, où η désigne le vecteur des paramètres associés à la partie fixe du modèle, $\eta = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \mu, \mu_{\mathrm{Pd}}, \alpha)$. L'estimation de p s'obtient par un calcul direct et η comme solution d'un système linéaire, similaire à celui obtenu pour la partie fixe d'un modèle mixte classique en supposant connues les composantes de la variance. On recherche ensuite les valeurs de σ_E^2, puis de ρ, où $\rho = \dfrac{\sigma_H^2}{\sigma_E^2}$, qui maximisent la vraisemblance obtenue en fixant les autres paramètres à leur valeur courante.

Les conditions de convergence de l'algorithme vers un point stationnaire de la log-vraisemblance données dans Meng et Rubin (1993) se vérifient facilement pour le modèle étudié.

5. Résultats

5.1. Estimation des paramètres

Les valeurs de θ estimées par maximum de vraisemblance sont présentées dans le Tableau 2.

On peut montrer que l'estimateur du maximum de vraisemblance possède les propriétés de consistance et de convergence en loi, lorsque l'on fait tendre vers l'infini les nombres d'haploïdes n_{gl} et n_l.

5.2. Test sur la ségrégation mendélienne

Une première hypothèse à tester est que la ségrégation du gène majeur suit les lois de Mendel, c'est-à-dire H_0 : « $p = \frac{1}{2}$ ». La statistique du test de rapport de vraisemblance de l'hypothèse H_0 contre l'alternative A_0 : « $p \neq \frac{1}{2}$ », converge en loi vers un chi-deux à 1 degré de liberté.

L'hypothèse H_0 n'est pas rejetée au niveau asymptotique 5% (cf. Tableau 2). Ce résultat suggère que la distribution est compatible avec la ségrégation de deux allèles d'un gène majeur.

TABLEAU 2. – Estimations et résultats des tests sous différentes hypothèses.

	Modèle		
	sans contraintes	sous H_0	sous H_0'
$\hat{p}$	0,47	0,50	0,47
$\hat{\gamma}_1$	8,17	8,16	7,53
$\hat{\gamma}_2$	7,72	7,72	7,14
$\hat{\gamma}_3$	7,77	7,77	7,11
$\hat{\gamma}_4$	9,16	9,16	8,59
$\hat{\mu}_D$	-3,72	-3,72	-2,90
$\hat{\mu}_Y$	3,72	3,72	2,90
$\hat{\mu}_{\mathrm{Pd}}$	-3,53	-3,53	-2,90
$\hat{\mu}_{\mathrm{Py}}$	2,07	2,07	2,90
$\hat{\sigma}_E^2$	3,16	3,16	3,92
$\hat{\sigma}_H^2$	0,80	0,80	-
Maximum de la log-vraisemblance	-13036	-13037	-13503
Statistique de test	-	1,12	933,46
quantile à 95%	-	3,84	7,07
seuil	-	0.30	$< 10^{-3}$

5.3. Test sur l'action d'un gène unique

Dans le modèle tel que nous l'avons présenté, l'hypothèse de l'action d'un gène unique s'exprime par H_0' : « $\sigma_H^2 = 0; \mu_D = \mu_{\mathrm{Pd}}; \mu_Y = \mu_{\mathrm{Py}}$ ». Sous cette hypothèse, le paramètre de variance σ_H^2 appartient à la frontière de son domaine de définition. Nous ne sommes pas dans un des cas référencés dans l'article de Self et Liang (Self et Liang, 1987), mais la configuration des paramètres se rapproche du cas 6, où il y a un seul paramètre qui, sous l'hypothèse à tester, appartient à la frontière du domaine de définition des paramètres. On montre, de manière similaire, que la statistique du test de rapport de vraisemblance de l'hypothèse H_0' contre l'alternative A_0' : « $\sigma_H^2 > 0$ ou $\mu_D \neq \mu_{\mathrm{Pd}}$ ou $\mu_Y \neq \mu_{\mathrm{Py}}$ » converge en loi vers un mélange 50 : 50

de chi-deux à 2 et 3 degrés de liberté. La loi de ce mélange n'étant pas tabulée, nous avons estimé le quantile à 95% par le quantile empirique obtenu en simulant un échantillon de taille 100000.

L'hypothèse est nettement rejetée à un seuil inférieur à 10^{-3} (cf. Tableau 2). Il est donc vraisemblable que le gène majeur n'agit pas seul et que son effet soit nuancé par d'autres gènes mineurs.

RÉFÉRENCES

DEMPSTER A., LAIRD N.M., RUBIN D.B. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, B 39, 1-38.

LOISEL P., GOFFINET B., MONOD H., MONTES DE OCA G. (1994), Detecting a major gene in an F2 population, *Biometrics*, 50, 512-516.

MENG X.L., RUBIN D.B. (1993), Maximum likelihood estimation via the ECM algorithm : A general framework, *Biometrika*, 80, 267-278.

SELF S.G., LIANG K. (1987), Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard condition, *Journal of the American Statistical Association*, 82, 605-610.