Estimation de l'apparentement entre plusieurs individus à l'aide d'un algorithme EM

Fabien Laporte [1] & Alain Charcosset [2] & Tristan Mary-Huard [3]

[1] *UMR de Génétique Végétale INRA/Univ. Paris Sud/CNRS, Ferme du Moulon 91190 Gif-sur-Yvette, France, fabien.laporte@moulon.inra.fr*

[2] *UMR de Génétique Végétale INRA/Univ. Paris Sud/CNRS, Ferme du Moulon 91190 Gif-sur-Yvette, France, alain.charcosset@moulon.inra.fr*

[3] *INRA/AgroParisTech, UMR 518, 75231, Paris, France, tristan.mary-huard@agroparistech.fr*

Résumé. L'apparentement entre deux individus se définit comme le nombre d'allèles hérités d'un ou plusieurs ancêtres communs sur l'ensemble du génome. Cet apparentement peut être modélisé à l'aide d'un modèle de mélange, où les variables observées sont les allèles observés à chaque marqueur qui définissent un mode IBS (Identity by State), et les variables cachées sont les origines ancestrales de ces allèles, qui définissent un mode IBD (Identity by Descent) du marqueur. L'objectif est d'estimer les proportions des modes IBD sur l'ensemble des marqueurs. Une approche usuelle consiste alors à estimer ces proportions par maximum de vraisemblance. Toutefois, l'identifiabilité du modèle n'est actuellement garantie que lorsque les marqueurs sont multi-alléliques (plus de 3 allèles observés). Par ailleurs les individus étudiés sont supposés tous appartenir à une seule population. Le présent article considère la question de l'identifiabilité pour des marqueurs bialléliques, et généralise les résultats classiques au cas où les individus sont issus de croisements entre parents issus de différentes populations.

Mots-clés. Apparentement, Modèle de Mélange, Identifiabilité, Marqueurs SNP

Abstract. The relatedness between two individuals is defined as the number of alleles inherited from one or several common ancestors along all the genome. The relatedness can be modeled with a mixture model, where observed variables are the alleles which define the Identity by State (IBS) mode of the marker and, hidden variables are the ancestral origins of these alleles which define Identity by Descent (IBD) modes of the marker. The objective is then to estimate the proportions of the different IBD modes over all available markers. The usual approach is the estimation of relatedness parameters with Maximum-Likelihood method. Some extensions of previous work have to be done. First, the identifiability of the model has been proofed only with multi-allelic markers (more than 3 alleles). Secondly, studied individuals are supposed to belong to one unique population. First, we study the identifiability of the model with biallelic markers. Then, we generalize results to multi-population data.

Keywords. Relatedness, Mixture Model, Identifiability, SNP Markers

1 Introduction

En génétique, l'apparentement est une notion qui vise à résumer l'information sur la proportion de génome que deux individus ont héritée des mêmes ancêtres. On suppose que les individus considérés font partie d'un pedigree, c'est-à-dire possèdent un arbre généalogique commun d'une profondeur donnée, dont les individus les plus anciens sont appelés fondateurs. Tout individu du pedigree est par définition le descendant d'un ou plusieurs fondateurs. Le génome d'un tel individu sera donc entièrement composé de portions de génomes issues des différents fondateurs et héritées au gré des appariements (mariages chez les humains, croisements chez les plantes ou les animaux) entre individus des générations antérieures. L'identité par descendance (IBD) se définit comme le fait que deux allèles homologues (c'est-à-dire correspondant à une même position du génome) soient la copie exacte d'un allèle ancestral apporté par un fondateur. Dans sa définition la plus simple, l'apparentement entre deux individus est la probabilité que deux allèles homologues issus de chacun de ces individus soient IBD pour une position du génome tirée au hasard.
D'autres définitions plus générales sont toutefois nécessaires en génétique quantitative. Si l'on suppose que les individus sont diploïdes, c'est-à-dire possèdant chacun deux copies de chaque chromosome, un couple d'individu présente un total de quatre allèles. Il existe 15 statuts possibles pour l'IBD entre ces 4 allèles, décrits par Gillois (1964). On définit alors l'apparentement entre deux individus comme les proportions de locus associées à chacun des 15 statuts IBD. Le statut IBD des locus n'étant en pratique pas observés, les proportions d'IBD constituant l'apparentement doivent être inférées.

En pratique, le statut IBD à un marqueur n'est pas observable, la seul information disponible est la lecture des allèles aux marqueurs, qui définissent des statuts IBS (Identical by State). On utilise ici l'approche proposée par Milligan (2002) qui explique les statuts IBS en fonction des statuts IBD. Ce dernier utilise un modèle de mélange, et l'inférence se fait donc généralement à l'aide de l'estimateur du maximum de vraisemblance.

2 Méthode

2.1 Notations

L'observation d'un marqueur pour deux individus comporte 4 allèles qui définissent un mode IBS (Identity by State). Ces mêmes 4 alleles peuvent descendre d'ancêtres communs de 15 manières différentes, qui définissent un mode IBD (Identity by Descent). Ainsi, à un marqueur fixé, deux individus $H_1 = (A, B)$ et $H_2 = (C, D)$ peuvent prendre 16 modes IBS différents et 15 modes IBD différents, résumés dans les tableaux ci-dessous.

Table 1: Table IBD

C_1	(A, B, C, D)	C_9	(A, B, C, D)
C_2	(A, B, C, D)	C_{10}	(A, B, C, D)
C_3	(A, B, C, D)	C_{11}	(A, B, C, D)
C_4	(A, B, C, D)	C_{12}	(A, B, C, D)
C_5	(A, B, C, D)	C_{13}	(A, B, C, D)
C_6	(A, B, C, D)	C_{14}	(A, B, C, D)
C_7	(A, B, C, D)	C_{15}	(A, B, C, D)
C_8	(A, B, C, D)		

Deux gamètes sont de la même couleur s'ils partagent un allèle provenant d'un ancêtre commun

Table 2: Table IBS

O_1	(A, B, C, D)	O_9	(A, B, C, D)
O_2	(A, B, C, D)	O_{10}	(A, B, C, D)
O_3	(A, B, C, D)	O_{11}	(A, B, C, D)
O_4	(A, B, C, D)	O_{12}	(A, B, C, D)
O_5	(A, B, C, D)	O_{13}	(A, B, C, D)
O_6	(A, B, C, D)	O_{14}	(A, B, C, D)
O_7	(A, B, C, D)	O_{15}	(A, B, C, D)
O_8	(A, B, C, D)	O_{16}	(A, B, C, D)

Une gamète est noire si son allèle est 0 et rouge s'il est 1

Par ailleurs, chacun des allèles A, B, C et D ont été hérités d'individus potentiellement issus de populations différentes, et les fréquences alléliques peuvent elle-même être différentes dans chacune de ces populations. La fréquence de l'allèle 1 dans chacune des 4 populations concernées est notée respectivement p_1, p_2, p_3 ou p_4. On note par ailleurs $q_i = 1 - p_i$ pour tout i dans $\{1, ..., 4\}$. Le nombre de marqueurs disponible est noté L.

À un marqueur ℓ, l'observation du mode IBS est notée IBS^ℓ et le mode IBD est noté IBD^ℓ. La probabilité a priori d'être dans le mode IBD C_j est notée $\Delta_j = P(IBD^\ell = C_j)$ et ne dépend pas du marqueur observé.

2.2 Modèle

Le but est d'estimer les paramètres $(\Delta_j)_{1 \leq j \leq 15}$ à partir des modes IBS. Pour cela, Milligan (2002) propose un modèle de mélange dont les variables observées sont les statuts IBS aux marqueurs et les variables cachées sont les statuts IBD à ces même marqueurs. Pour ce modèle la vraisemblance associée au marqueur ℓ est :

$$P(IBS^\ell = O_i) = \sum_{j=1}^{15} P(IBS^\ell = O_i | IBD^\ell = C_j)\Delta_j \ .$$

La log-vraisemblance complète est de la forme :

$$\sum_{\ell=1}^{L} \log(\sum_{j=1}^{15} P(IBS^\ell = O_i | IBD^\ell = C_j)\Delta_j)$$

et les estimateurs du maximum de vraisemblance des paramètres Δ_j sont définis par:

$$(\hat{\Delta}_1, ...\hat{\Delta}_{15}) = \underset{\Delta_1,...,\Delta_{15}}{\arg\max} \sum_{\ell=1}^{L} \log(\sum_{j=1}^{15} P(IBS^\ell = O_i | IBD^\ell = C_j)\Delta_j)$$

Ce modèle peut se reécrire de façon matricielle. Si :

- $P(IBS^\ell) = (P(IBS^\ell = O_1), ..., P(IBS^\ell = O_{16}))^t$
- $M^\ell = (m_{ij})_{1 \leq i \leq 16, 1 \leq j \leq 15}$ avec $m_{ij} = P(IBS^\ell = O_i | IBD^\ell = C_j)$
- $\Delta = (\Delta_1, ..., \Delta_{15})^t$

alors le modèle peut s'écrire :

$$P(IBS^\ell) = M^\ell \Delta$$

La matrice M^ℓ est entièrement déterminée par les fréquences alléliques, supposées connues.

2.3 Identifiabilité

On rappelle ici que les coefficients d'apparentement sont soumis à certaines contraintes génétiques :

Propriété 1 *Tout vecteur d'apparentement Δ doit respecter les trois contraintes génétiques suivantes :*

- $\Delta_8 \times \Delta_1 = \Delta_2 \times \Delta_3$
- $\Delta_9 \times \Delta_1 = \Delta_4 \times \Delta_7$
- $\Delta_{10} \times \Delta_1 = \Delta_5 \times \Delta_6$

La propriété suivante découle naturellement :

Propriété 2 *Si les contraintes génétiques sont prises en compte alors le modèle est identifiable.*

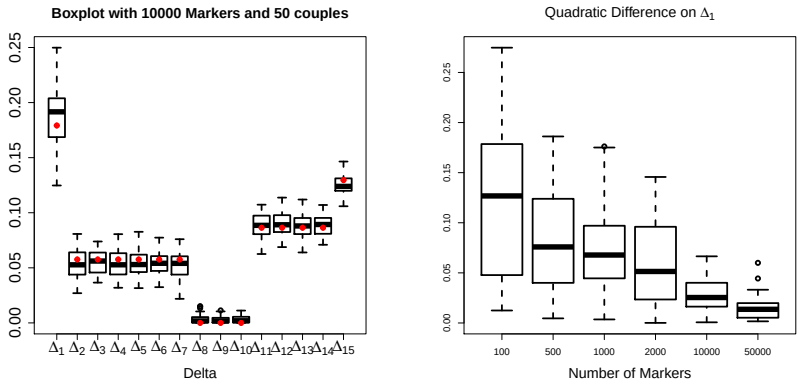
2.4 Estimation

Il n'y a pas de forme close pour les estimateurs du maximum de vraisemblance. La vraisemblance est donc maximisée algorithmiquement via un algorithme EM. A cause des possibles maximum locaux, l'algorithme EM est lancé sur plusieurs initialisations vérifiant les contraintes génétiques.

3 Résultats

3.1 Une Population

Les performances de la méthode sont étudiées sur des données simulées. Les individus appartiennent à une même population. Cette dernière est issue d'un fondateur commun et chaque lignée parentale (individu homozygote) a la même probabilité d'avoir hérité d'un allèle issu du fondateur. Ainsi l'apparentement théorique entre les individus hybrides (croisement de deux lignées) est connu.

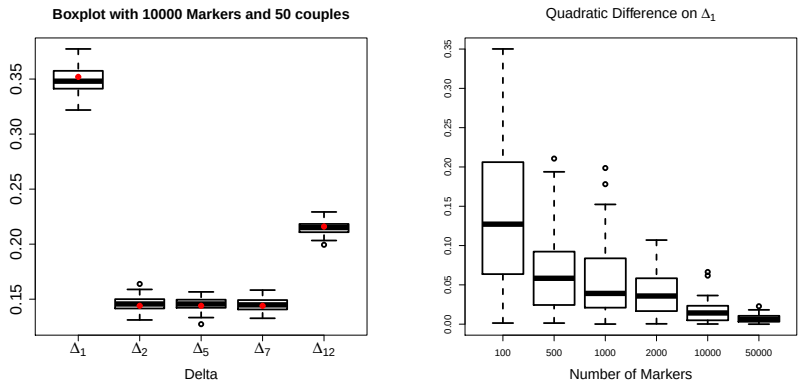


Sur la figure de gauche est réprésenté le boxplot des estimations de chaque coefficient de l'apparentement. Les points rouges représentent l'apparentement théorique entre les deux hybrides. La figure de droite représente l'évolution de la précision de l'algorithme en fonction du nombre de marqueurs utilisés.

Le coût algorithmique dépend de façon quadratique du nombre d'hybrides étudiés (en effet pour N hybrides il y a $\frac{N(N-1)}{2}$ jeux de paramètres d'apparentement à estimer), et linéairement du nombre de marqueurs. Par exemple si le panel étudié possède 500 hybrides génotypés à l'aide de 50000 marqueurs, l'algorithme calcule durant une journée. En fonction de la précision requise pour l'estimation des paramètres, on peut donc réduire le nombre de marqueurs utilisés afin de diminuer le temps d'estimation.

3.2 Multi-Populations

Dans cette partie, les parents sont simulés venant de populations différentes. Ce qui induit la nullité de certains coefficients de l'apparentement.



Les graphiques sont les même représentations que précédemment. Dans celui de gauche, il n'apparaît que 5 coefficients, les autres étant fixés à 0 par l'algorithme.
Sur la figure de droite, on remarque que le gain de précision est faible entre 10000 et 50000 marqueurs. On peut s'intéresser à réduire le nombre de marqueurs utilisés (on gagne un facteur 5 dans le temps de calcul en passant de 50000 à 10000 marqueurs).

4 Conclusion

En conclusion, l'identifiabilité du modèle est montrée tant pour le cas classique que le cas multi-population. De plus un outil est développé pour inférrer les coefficients de l'apparentement dans tous les cas de figure.

Bibliographie

[1] Milligan, B.G. (2002), *Maximum-Likelihood Estimation of Relatedness*, *Genetics 163: 1153-1167.*