

# Mesure de la diversité selon Weitzman : principes et applications

L. FOULLEY, C. THAON d'ARNOLDI, L. OLLIVIER  
INRA-Station de Génétique quantitative et appliquée, 78352 Jouy-en-Josas Cedex

**RESUME** - Cette étude rappelle les principes de la méthode de Weitzman de mesure de la diversité entre populations connaissant les distances deux à deux entre celles-ci. Elle souligne l'originalité de la méthode qui procure à la fois une mesure de la diversité d'un ensemble de populations (y compris la contribution d'une ou plusieurs d'entre elles à celle-ci) et une méthode de classification des populations. La théorie est illustrée sur un ensemble de 19 races bovines européennes caractérisée pour 13 loci de polymorphisme biochimique. La méthode offre ainsi une base rationnelle à la définition de priorités en matière de préservation des ressources génétiques ; elle devrait aussi éclairer les choix à opérer dans la création de lignées synthétiques.

## Measuring diversity according to Weitzman : principles and applications

L. FOULLEY, C. THAON d'ARNOLDI, L. OLLIVIER  
INRA-Station de Génétique quantitative et appliquée, 78352 Jouy-en-Josas Cedex

**SUMMARY** - This study reviews the foundations of Weitzman's method for measuring diversity among populations characterized by pairwise distances. It emphasizes the originality of the method which provides not only a measure of the diversity of a set of populations (including the contribution of one or several of them to the total diversity) but also a clustering procedure of the populations. The theory is illustrated on a set of 19 European cattle breeds characterized at 13 loci of biochemical polymorphism. This method provides a rational basis to define preservation strategies of genetic resources ; it should also help in creating synthetic lines.

## INTRODUCTION

La préservation des ressources génétiques est actuellement une préoccupation majeure des biologistes et des généticiens en particulier. La gestion de cette diversité doit s'appuyer sur une définition claire, rigoureuse et opérationnelle de sa mesure. La méthode proposée par l'économiste Martin Weitzman (1991, 1992, 1993) mérite de retenir l'attention de par son originalité et la rigueur du raisonnement qui y conduit. Elle procure à la fois une mesure de la diversité d'un ensemble de populations et une méthode de classification de celles-ci. Comme cette méthode est fort peu connue des biologistes, il nous est apparu utile d'en proposer un bref exposé pédagogique illustré par une application aux races bovines européennes.

## METHODE

L'objectif est de définir une fonction de diversité (symbolisée par  $V$ ) à partir de distances génétiques. La méthode s'appuie sur le principe dit de monotonie en espèces selon lequel toute adjonction d'un élément  $j$  (ici la race) à un ensemble  $Q$  se traduit par un accroissement de diversité au moins égal à la distance  $d(j, Q)$  entre  $j$  et  $Q$ ,

$$[1] \quad V(Q \cup j) \geq V(Q) + d(j, Q).$$

Ici, la distance  $d(j, Q)$  du point  $j$  à l'ensemble  $Q$  est mesurée classiquement par la distance "mathématique" entre  $j$  et son plus proche voisin de  $Q$ . En fait, une mesure de dissimilarité suffit et nous choisirons ici une distance génétique classique : Nei-standard, Nei-DA, Cavalli-Sforza, Reynolds, Sanghvi (Foulley and Hill, 1999 ; Ollivier et al, 2000).

En fait, on va considérer [1] comme une contrainte générale devant être satisfaite par tout élément  $i$  de l'ensemble complet  $S$  des populations considérées

$$[2] \quad V(S) \geq V(S \setminus i) + d(i, S \setminus i), \quad \forall i \in S,$$

( $S \setminus i$  désignant l'ensemble  $S$  privé de son élément  $i$ )

Posons  $V_i^\# = V(S \setminus i) + d(i, S \setminus i)$ . La valeur de  $V_i^\#$  dépend de l'élément  $i$  choisi dans l'ensemble  $S$  et la fonction de diversité  $V(S)$  doit donc satisfaire :

$$V(S) \geq V_1^\#, V(S) \geq V_2^\#, \dots, V(S) \geq V_i^\#, \quad \forall i$$

Une condition nécessaire et suffisante pour satisfaire l'ensemble des contraintes générées en [2] sur la mesure de diversité est que celle-ci soit supérieure ou égale à tous les quantités  $V_i^\#$  donc à la plus grande d'entre elles :

$$V(S) \geq \max_{i \in S} V_i^\#$$

Pour garantir l'unicité de la fonction, on va se restreindre au signe égalité avec la condition initiale  $V[\{\emptyset\}] = K$  (constante), d'où on déduit la définition récursive de  $V(S)$

$$[3] \quad V(S) = \max_{i \in S} [V(S \setminus i) + d(i, S \setminus i)],$$

qui s'énonce ainsi : la diversité d'un ensemble est le maximum par rapport à l'un quelconque de ses éléments de la distance entre celui-ci et son plus proche voisin augmentée de la diversité de l'ensemble privé de cet élément.

## RÈGLE FONDAMENTALE DE REPRÉSENTATION

Le calcul de [3] nécessite  $n!$  opérations ce qui est très rapidement prohibitif,  $n$  étant le nombre d'éléments de  $S$ . Par chance, ce calcul peut s'effectuer à un degré de complexité inférieur ( $2^n$ ) grâce à un algorithme qui appliqué récursivement génère un arbre de classification ascendante.

## DÉFINITION DU LIEN ET DU REPRÉSENTANT

Il existe un élément  $i$  de  $S$  pour lequel le maximum de [3] est atteint. Weitzman a montré que cet élément  $i$  de  $d(i, S \setminus i)$  est l'un des deux plus proches voisins existant dans l'ensemble  $S$  soit

$$[4] \quad d(i, S \setminus i) = \min_{u, v \in S} d(u, v).$$

Comme  $d(i, S \setminus i) = V(S) - V(S \setminus i)$ , la propriété énoncée en [4] veut dire aussi qu'il existe un élément  $i$  de  $S$  dont le retrait conduit à une perte minimum de diversité. Cet élément constitue précisément le lien.

Cela nous permet de répondre à la question suivante : ayant identifié le couple des plus proches parents, comment trouve-t-on le lien? Si l'on désigne ce couple par  $[g(S), h(S)]$ ,  $V$  doit satisfaire :

$$[5] \quad V(S) = d[g(S), h(S)] + \max\{V[S \setminus g(S)], V[S \setminus h(S)]\}.$$

Selon les notations et la terminologie de Weitzman,  $g(S)$  satisfaisant  $\max[V(S \setminus g), V(S \setminus h)]$  est donc le lien et  $h(S)$ , le représentant.

## ALGORITHME ET REPRÉSENTATION DE L'ARBRE

Par application récursive de [5], on génère un arbre enraciné selon une classification hiérarchique ascendante ; les terminaisons de cet arbre sont les éléments de  $S$  et, les nœuds, les "ancêtres" inconnus de ces éléments.

Le différentes étapes de l'algorithme sont les suivantes :

-1) rechercher les deux éléments  $i$  et  $j$  plus proches voisins et incrémenter la diversité de  $d(i, j)$  ;

-2) trouver le lien  $g$  et le représentant  $h$  à partir de la recherche de  $g$  maximisant  $[V(S \setminus g), V(S \setminus h)]$  ;

-3) fusionner  $g$  et  $h$  et

représenter le clade ainsi formé par le "représentant" ;

-4) former le nouvel ensemble privé de  $g$ , soit  $S \setminus g$  ;

-5) revenir en 1) jusqu'à  $n=1$ , ajouter la constante (souvent mise à zéro) et s'arrêter.

Le résultat est un arbre de taxonomie parfaite qui peut être représentée par la matrice correspondante de distances ultramétriques. Lors de la construction de l'arbre, on peut avantageusement placer le lien  $g$  entre le représentant et le plus proche voisin de  $h$  dans  $Q \setminus g$ ,  $Q$  étant le sous-ensemble dont on calcule alors la diversité. Intuitivement, cela signifie que le lien est plus près du reste de l'ensemble que le représentant et qu'en conséquence sa perte est moins dommageable. Cela a aussi l'avantage de ne fournir que deux représentations possibles de la taxonomie, symétriques l'une de l'autre autour de l'axe de la racine.

Enfin, la diversité peut se lire directement sur l'arbre comme la somme des longueurs des branches qui le constitue. Weitzman a par ailleurs donné une interprétation de cet arbre comme la représentation de la phylogénie maximisant la probabilité d'existence des espèces le constituant à un instant donné.

## RESULTATS

La mise en œuvre de cet algorithme peut être illustrée à partir d'un exemple très simple. Celui-ci concerne 4 races bovines allaitantes (Blonde d'Aquitaine, Charolaise, Limousine et Maine Anjou) parmi les 19 étudiées par Grosclaude et al (1990) sur le polymorphisme biochimique à 13 loci (11 de groupes sanguins, le locus de la transferrine sérique et celui de la caséine bêta). Les distances standard de Nei entre ces quatre races prises deux à deux sont reportées au tableau 1.

Tableau 1  
Distances (x10000) de Nei entre les 4 races bovines

|    | BL  | CH  | LI  | MA  |
|----|-----|-----|-----|-----|
| BL | 0   | 166 | 289 | 826 |
| CH | 166 | 0   | 309 | 739 |
| LI | 289 | 309 | 0   | 838 |
| MA | 826 | 739 | 838 | 0   |

Les deux races les plus voisines sont la Blonde d'Aquitaine (BL) et la Charolaise (CH) si bien que la diversité totale s'établit à :

$$V = d(BL, CH) + \max[V(S \setminus BL), V(S \setminus CH)]$$

$$V = 166 + \max[V\{CH, LI, MA\}, V\{BL, LI, MA\}]$$

La question est maintenant de savoir quel est le lien et quel est le représentant de ces deux races. Pour ce faire, nous allons calculer la diversité de l'ensemble S privé de BL soit  $V\{CH, LI, MA\}$  et celle de l'ensemble S privé de CH soit  $V\{BL, LI, MA\}$ . Considérons le sous-tableau de distances entre CH, LI et MA :

|    | CH  | LI  | MA  |
|----|-----|-----|-----|
| CH | 0   | 309 | 739 |
| LI | 309 | 0   | 838 |
| MA | 739 | 838 | 0   |

Les races les plus proches sont CH et LI distantes de 309 et la diversité de l'ensemble s'obtient par :

$$V\{CH, LI, MA\} = 309 + \max[V\{LI, MA\}, V\{CH, MA\}]$$

soit, sachant que la diversité de deux races se réduit à la distance entre elles :

$$V\{CH, LI, MA\} = 309 + \max(838, 739),$$

$$V\{CH, LI, MA\} = 309 + 838 = 1147.$$

Dans cet ensemble de 3 races, le lien (g) est CH et le représentant (h), LI.

Les mêmes raisonnements et calcul s'appliquent à l'ensemble  $\{BL, LI, MA\}$  qui aboutissent à :

$$V\{BL, LI, MA\} = d(BL, LI) + \max[V\{LI, MA\}, V\{BL, MA\}]$$

$$V\{BL, LI, MA\} = 229 + 838 = 1127 \text{ avec}$$

$$g = BL \text{ et } h = LI.$$

Comme  $V\{CH, LI, MA\} > V\{BL, LI, MA\}$ ,  $g = BL$  et  $h = CH$  dans l'ensemble initial des 4 races et

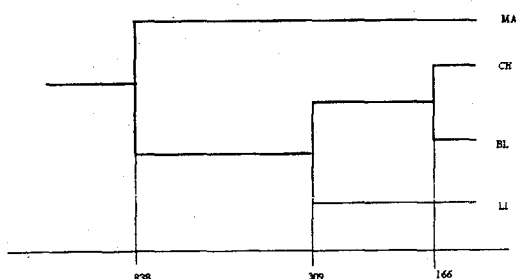
$$V(S) = d(BL, CH) + V\{CH, LI, MA\} \text{ soit :}$$

$$V(S) = d(BL, CH) + d(CH, LI) + d(LI, MA), \text{ ou encore}$$

$$V(S) = 166 + 309 + 838 = 1313.$$

Le lien BL dans  $\{BL, CH, LI, MA\}$  est placé entre le représentant CH et le plus proche voisin LI de ce même représentant dans l'ensemble S privé du lien BL ; le lien CH dans  $\{CH, LI, MA\}$  est placé entre le représentant LI et le plus proche voisin MA de LI dans  $\{LI, MA\}$  (figure 1).

Figure 1  
Représentation graphique des 4 races



Quand on traite l'ensemble des 19 races selon la méthode de Weitzman, on obtient une représentation graphique (figure 2) qui discrimine bien un groupe laitier d'un groupe mixte, à viande et rustique dont les éléments sont bien rangés dans l'ordre attendu du gradient zootechnique : rustique, viande, mixte et laitier. On notera à cet égard la place originale de la race Normande entre les deux groupes. De ce point de vue, l'arbre de Weitzman écarte les ambiguïtés ou les a priori de positionne-

ment des races liés à la rotation possible des branches autour des axes.

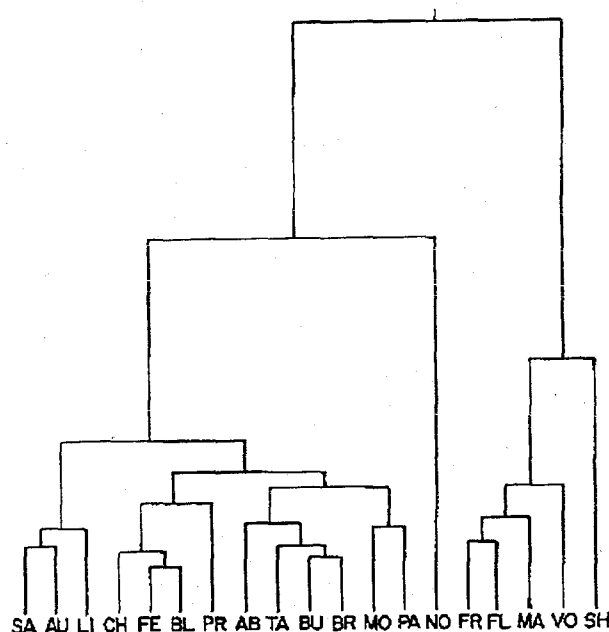
La mesure de diversité permet également d'évaluer l'importance de races régionales et de races à effectifs limités (Thaon et al, 1998). Dans le cas précédent, on s'est par exemple intéressé à la perte relative de diversité résultant du passage d'un état où les 18 races françaises de l'échantillon sont présentes à un état où ne subsistent plus que les 6 principales races laitières (Frisonne, Montbéliarde, Normande) et à viande (Blonde d'Aquitaine, Charolaise, Limousine) ainsi qu'une seule race, objet supposé de mesures de conservation, prise parmi les 12 restantes. Les résultats (en %) obtenus à partir de la distance Nei standard sont les suivants :

|                 |      |                  |      |
|-----------------|------|------------------|------|
| Salers (SA)     | 43,6 | Brune (BU)       | 52,1 |
| Tarentaise (TA) | 50,9 | Parthenaise (PA) | 52,3 |
| Aubrac (AU)     | 51,0 | Pie Rouge (PR)   | 52,8 |
| Abondance (AB)  | 51,0 | Maine Anjou (MA) | 52,3 |
| Vosgienne (VO)  | 51,7 | Flamande (FL)    | 54,1 |
| Bretonne (BR)   | 51,8 | Ferrandaise (FE) | 55,1 |

La race qui, ajoutée aux six autres, conduit avec ce critère à la plus faible perte de diversité est la Salers suivie par la Tarentaise, l'Aubrac et l'Abondance. Ce particularisme de la Salers demeure en considérant une autre distance comme celle de Cavalli-Sforza.

Bien qu'il ne s'agisse là que d'une illustration de la méthode qui mériterait plus ample développement notamment vis-à-vis des loci étudiés, on notera que la race ainsi singularisée est une race rustique dont l'intérêt zootechnique a été mis en avant depuis longtemps.

Figure 2  
Arbre de Weitzman relatif aux 19 races



## CONCLUSION

Cette méthode offre une base rationnelle à la définition des priorités en matière de préservation des populations naturelles et domestiques (Ollivier et al, 2000).

Son objectif premier est la mesure de la diversité ; secondairement, elle fournit un arbre de classification des populations comme représentation de l'algorithme récursif de calcul de la diversité et qui constitue une typologie bidimensionnelle de ces populations. Il serait probablement assez dangereux d'interpréter cet arbre en terme de phylogénie dans le cas des races exploitées en élevage dont l'origine et l'évolution sont complexes compte-tenu de l'intervention de l'homme.

La méthode de Weitzman devrait permettre également d'éclairer les choix à opérer dans la création de lignées synthétiques à partir d'un éventail important de lignées parentales a priori intéressantes. Pour ce faire, il importerait de compléter l'information sur le polymorphisme biochimique par celui de type "microsatellites" ainsi que par une information ayant trait aux caractères fonctionnels et de production.

D'un point de vue pratique, le calcul exact de la diversité devient prohibitif pour des tailles supérieures à 25-30. Dans ce cas, une approximation par simulation a été proposée par Thaon et al (1998); celle-ci est basée sur un choix aléatoire du lien à chaque étape de l'algorithme récursif décrit précédemment; cela revient à échantillonner un certain nombre d'arbres parmi les  $2^{n-1}$  arbres possibles et à retenir celui de diversité maximum.

- Foulley, J.L., Hill, W.G., 1999.** Genet. Sel. Evol., 31, 457-464
- Grosclaude, F., Aupetit, R.Y., Lefebvre, J., Mériaux, J.C., 1990.** Genet. Sel. Evol., 22, 317-338
- Ollivier, L., Chevalet C., Foulley, J.L., 2000.** INRA Prod. Anim., numéro hors série " Génétique moléculaire " (sous presse).
- Thaon d'Arnoldi, C., Foulley, J.L., Ollivier, L., 1998.** Genet. Sel. Evol., 30, 149-161
- Weitzman, M., 1991.** Harvard Inst. Econ. Research Paper No 1569.
- Weitzman, M., 1992.** Quart. J. Econ. , 107, 363-405
- Weitzman, M., 1993.** Quart. J. Econ, 108, 157-183