Ann. Génét. Sél. anim., 1971, **3** (4), 449-462.

CALCUL AUTOMATIQUE DES COEFFICIENTS D'IDENTITÉ

C. CHEVALET

Laboratoire de Génétique cellulaire
Centre de Recherches de Toulouse, I. N. R. A.,
31 - Castanet-Tolosan

RÉSUMÉ

Une méthode de calcul automatique par simulation des coefficients d'identité est présentée. La population est fractionnée en générations successives : chaque génération est décrite par un certain nombre d' « états d'identité » équiprobables, indépendants et compatibles avec la généalogie ; la description d'une génération est calculée automatiquement à partir de la description de la génération précédente. Les résultats sont approchés, la précision est améliorée au prix d'un accroissement des temps de calcul ou d'une diminution de la capacité de traitement (800 individus par génération) due au calcul exact des coefficients de parenté et de consanguinité.

Ce moyen de calcul permet d'exploiter les méthodes d'estimation des composantes de la variance génotypique dans les populations consanguines et, conjointement avec ces méthodes, d'élaborer de nouveaux outils statistiques pour la génétique quantitative.

INTRODUCTION

L'expression des covariances génotypiques entre deux individus apparentés fait intervenir, dans le cas général, non seulement les coefficients de parenté et de consanguinité attachés à ces individus, mais aussi les coefficients d'identité, définis par GILLOIS (1964, 1965, 1966 *a*, *b*). Ce sont les probabilités des quinze « situations d'identité » qui sont susceptibles de se présenter en quatre gènes homologues pris chez les deux individus considérés, et elles ne dépendent que des liens d'ascendance entre eux (tabl. 1).

Calcul des coefficients de parenté

Des méthodes de calcul, adaptées à un traitement automatique, ont été proposées pour les coefficients de parenté et de consanguinité. On peut en distinguer deux sortes :

La première résulte de l'application de la formule, démontrée par MALÉCOT (1948) :

$$f = \sum_{C_i} (1 + f_i)\, 2^{-n_i - 1}$$

il faut dans le graphe que constitue le pedigree d'un individu isoler les chaînes de parenté C_i qui relient les deux gènes de ce zygote, compter pour chacune le nombre n_i de chaînons, et rechercher le coefficient de consanguinité f_i de son sommet. La programmation de cette recherche peut être faite directement (EMIK et TERRIL,

TABLEAU I

Les quinze situations d'identité entre quatre gènes

	Situation d'identité	Notations numériques $g_1\ g_2\ g_3\ g_4$	Probabilité
S_1	$(g_1 \equiv g_2 \equiv g_3 \equiv g_4)$	1 1 1 1	δ_1
S_2	$(g_1 \equiv g_2 \equiv g_3)\ 0\ (g_4)$	1 1 1 2	δ_2
S_3	$(g_1 \equiv g_2 \equiv g_4)\ 0\ (g_3)$	1 1 2 1	δ_3
S_4	$(g_1 \equiv g_3 \equiv g_4)\ 0\ (g_2)$	1 2 1 1	δ_4
S_5	$(g_2 \equiv g_3 \equiv g_4)\ 0\ (g_1)$	1 2 2 2	δ_5
S_6	$(g_1 \equiv g_2)\ 0\ (g_3 \equiv g_4)$	1 1 2 2	δ_6
S_7	$(g_1 \equiv g_2)\ 0\ (g_3)\ 0\ (g_4)$	1 1 2 3	δ_7
S_8	$(g_1)\ 0\ (g_2)\ 0\ (g_3 \equiv g_4)$	1 2 3 3	δ_8
S_9	$(g_1 \equiv g_3)\ 0\ (g_2 \equiv g_4)$	1 2 1 2	δ_9
S_{10}	$(g_1 \equiv g_3)\ 0\ (g_2)\ 0\ (g_4)$	1 2 1 3	δ_{10}
S_{11}	$(g_2 \equiv g_4)\ 0\ (g_1)\ 0\ (g_3)$	1 2 3 2	δ_{11}
S_{12}	$(g_1 \equiv g_4)\ 0\ (g_2 \equiv g_3)$	1 2 2 1	δ_{12}
S_{13}	$(g_1 \equiv g_4)\ 0\ (g_2)\ 0\ (g_3)$	1 2 3 1	δ_{13}
S_{14}	$(g_2 \equiv g_3)\ 0\ (g_1)\ 0\ (g_4)$	1 2 2 3	δ_{14}
S_{15}	$(g_1)\ 0\ (g_2)\ 0\ (g_3)\ 0\ (g_4)$	1 2 3 4	δ_{15}

1949), ou par l'intermédiaire d'un calcul matriciel fondé sur la théorie des graphes (MARUYAMA et YASUDA, 1970). L'autre approche, développée d'abord par CRUDEN (1949), consiste à calculer systématiquement, de génération en génération, les coefficients de parenté relatifs à tous les couples possibles. Le passage d'une génération à la suivante résulte des formules suivantes :

si M (respectivement N) a pour parents I et J (respectivement K et L), si $\varphi(X, Y)$ désigne le coefficient de parenté de X et de Y, et si l'on pose par convention :

$$\varphi(X, X) = \frac{1}{2} + \frac{1}{2} f(X)$$

où $f(X)$ est le coefficient de consanguinité de X, on aura :

$$\varphi(M, M) = \frac{1}{2} + \frac{1}{2}\varphi(I, J)$$

$$\varphi(N, N) = \frac{1}{2} + \frac{1}{2}\varphi(K, L)$$

$$\varphi(M, N) = \frac{1}{2}\varphi(I, N) + \frac{1}{2}\varphi(J, N)$$

$$= \frac{1}{2}\varphi(M, K) + \frac{1}{2}\varphi(M, L)$$

$$= \frac{1}{4}\varphi(I, K) + \frac{1}{4}\varphi(I, L) + \frac{1}{4}\varphi(J, L) + \frac{1}{4}\varphi(J, K)$$

Calcul des coefficients d'identité

Le calcul des coefficients d'identité proposé par GILLOIS (1964, 1966 *a*) est une généralisation de la première méthode de calcul des coefficients de parenté, et fait usage des inf. 1/2 faisceaux, ensembles ordonnés dont les chaînes de parenté sont un cas particulier. Une tentative de systématisation a été présentée par JACQUARD (1966), mais elle comporte une simplification qui, dans le cas général, conduit à des résultats erronés.

Nous présentons ici une méthode pour calculer de génération en génération les coefficients d'identité pour tout couple d'individus d'une même génération. La population proposée est d'abord fractionnée en générations, puis décrite de façon à rendre les calculs aussi simples que possible, et à permettre le calcul de tout coefficient d'identité (CHEVALET, 1969). Déterminer les valeurs exactes nécessite un nombre d'opérations qu'un calculateur électronique ne saurait effectuer dans un temps limité : les résultats obtenus sont approchés. Par ailleurs, les programmes correspondant à cette méthode comportent le calcul exact des coefficients de parenté et de consanguinité par la méthode de Cruden. Seule la donnée pour chaque individu de la population : de son nom, du nom de son père, et du nom de sa mère est nécessaire.

I. — DÉCOMPOSITION D'UNE POPULATION EN GÉNÉRATIONS

A. — *Générations et générations immédiatement successives*

A tout sous-ensemble G de la population P, attachons les deux parties A(G) et D(G) :

A(G) est l'ensemble des individus de P qui sont ascendants d'individus de G, et qui sont hors de G.

D(G) est l'ensemble des individus de P qui sont descendants d'individus de G et qui sont hors de G.

Nous dirons qu'une partie G de P est une génération si :

1° la réunion de G, de A(G), et de D(G) recouvre toute la population P, et si :

2° l'intersection de A(G) et de D(G) est vide.

Deux générations G et G' sont « successives » si leur réunion est elle-même une génération. Elles sont aussi telles que tout individu de G', ou bien se trouve déjà dans G, ou bien a ses parents dans la réunion de G et G'. Deux générations H et H' sont « immédiatement successives » si H' se déduit de H :

1° par l'adjonction d'un descendant direct X (fils ou fille) d'individus de H,

2° éventuellement par l'élimination du ou des parents de X qui n'auraient dans D(H) pas d'autre descendant que X.

Ces définitions ne font appel qu'à la structure de l'ensemble P, partiellement ordonné par la relation d'ascendance.

B. — *Décomposition en une suite de générations immédiatement successives*

En vue des calculs, une population sera analysée de la façon suivante :

1° Déterminer une liste de générations successives G_1, G_2,..., G_n d'après la seule connaissance des pedigrees ou d'après l'histoire de la population.

2° Chercher un ordre d'apparition

$$(Z_{k+1}, Z_{k+2},..., Z_{k+n})$$

des individus entre la génération

$$G_i = (Z_1, Z_2... Z_k)$$

et la génération suivante G_{i+1}, de façon que l'individu $Z_l (k < l \leqslant k + n)$ ait ses parents dans l'ensemble

$$G_i U (Z_{k+1}, Z_{k+2},..., Z_{l-1}).$$

Alors les ensembles suivants :

$$\begin{aligned}
H_0 &= G_i && = (Z_1, Z_2,..., Z_k)\\
H_1 &= G_i U (Z_{k+1}) && = (Z_1,..., Z_k, Z_{k+1})\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
H_{n-1} &= G_i U (Z_{k+1}, Z_{k+2},..., Z_{k+n-1}) && = (Z_1,..., Z_k,..., Z_{k+n-1})\\
H_n &= G_{i+1}
\end{aligned}$$

forment une suite de générations immédiatement successives.

C. — *Calculs associés à la décomposition*

Tout calcul de coefficients génétiques d'une génération à la suivante résultera de l'itération du processus élémentaire qui conduit d'une génération H_j à une génération H_{j+1} immédiatement successive.

La méthode de CRUDEN est compatible avec cette décomposition :

H_j étant décrite par la matrice $\Phi_j = ((\varphi_{m,n}))$ des coefficients de parenté des couples $(Z_m Z_n)$; $m, n \leqslant k + j$, la matrice Φ_{j+1} s'obtient en adjoignant à Φ_j les coefficients de parenté entre Z_{k+j+1} et $Z_m (m = 1, 2..., k + j + 1)$. Or Z_{k+j+1} a ses parents dans l'ensemble :

$$H_j = (Z_1, Z_2,..., Z_{k+j}).$$

Soient Z_p et Z_q ces parents. Il vient :

$$\varphi_{k+j+1} = \frac{1}{2} \varphi_{p,m} + \frac{1}{2} \varphi_{q,m} \qquad \text{pour} \qquad m = 1,2..., k + j$$

et :

$$\varphi_{k+j+1,k+j+1} = \frac{1}{2} + \frac{1}{2} \varphi_{p,q}$$

où les quantités $\varphi_{p,m}$, $\varphi_{q,m}$, $\varphi_{p,m}$ sont des éléments de Φj.

Le résultat du calcul est présenté pour le pedigree de la figure 1 (tabl. 2).

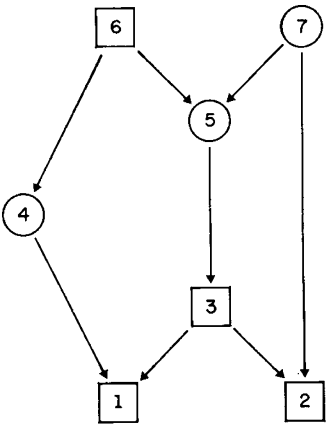

FIG. 1. — *Exemple de pedigree:*

P = ((7, 0, 0) ; (6, 0, 0) ; (5, 6, 7) ; (4, 6, 0) ; (3, 0, 5) ; (2, 3, 7) ; (1, 3, 4)).

Il admet les générations suivantes :

$$\begin{aligned} G_1 &= H_0 = (7\,;\,6) \\ & \quad H_1 = (7\,;\,6\,;\,5) \\ G_2 = H_2 &= K_0 = (7\,;\,5\,;\,4) \\ & \quad K_1 = (7\,;\,5\,;\,4\,;\,3) \\ & \quad K_2 = (7\,;\,5\,;\,3\,;\,1) \\ G_3 &= K_3 = (1\,;\,2). \end{aligned}$$

Dans les calculs des tableaux 2 et 3, les générations K_2 et K_1 sont remplacées par les ensembles K_1' et K_2', où l'on a retiré l'individu (5) et qui suffisent pour la suite du calcul.

TABLEAU 2

Matrice des coefficients de parenté calculés par la méthode de CRUDEN *pour la décomposition* (H_0, H_1, $H_2 = K_0$, K_1', K_2', K_3) *du pedigree de la figure 1*

7	6	5	4	3	1	2	
1/2	0	1/4	0	1/8	1/16	.	7
0	1/2	1/4	.	.	.	.	6
1/4	1/4	1/2	1/8	.	.	.	5
0	.	1/8	1/2	1/16	.	.	4
1/8	.	.	1/16	1/2	9/32	.	3
1/16	.	.	.	9/32	17/32	11/64	1
	.	.	.	.	11/64	9/16	2

II. — DESCRIPTION D'UNE POPULATION PAR LA RELATION D'IDENTITÉ

A. — *Situations et états d'identité*

La généralisation immédiate de la méthode de CRUDEN consisterait à décrire chaque génération H, non plus par la suite des probabilités d'identité de couples de gènes (les coefficients de parenté), mais par la suite des probabilités de toutes les situations d'identité relatives à tous les quadruplets de gènes présents dans H.

Si H comporte N zygotes diploïdes, donc 2N gènes homologues, il y a $\binom{2N}{4}$ quadruplets de gènes distincts, et le nombre de paramètres nécessaires, compte tenu des relations entre les coefficients d'identité relatifs aux mêmes gènes, est de :

$$4\binom{2N}{4} + \binom{2N}{3} + \binom{2N}{2},$$

de l'ordre de $\frac{8}{3}$ N^4 quand N est grand. Même avec un calculateur de grande puissance on ne peut envisager de manipuler un tel nombre de paramètres. De plus les calculs sont alourdis par le traitement des permutations entre gènes.

Nous utiliserons au contraire une description qui conduit à des calculs très simples, mais à des résultats approchés pour les coefficients d'identité : un ensemble de N zygotes diploïdes est décrit par la suite des situations d'identité relatives à leurs 2N gènes homologues, avec leurs probabilités. Une situation d'identité entre m gènes sera notée par une suite de m nombres entiers : chaque gène g_i est représenté par un nombre I_i, de façon que deux gènes g_i et g_j soient représentés par le même nombre $I_i = I_j$ si, et seulement si, ils sont identiques.

Chaque valeur numérique qui apparaît dans la suite,

$$(I_1, I_2, \ldots, I_m)$$

peut être identifiée à une « classe d'identité ».

Ainsi la situation S_7 :

$$(g_1 \equiv g_2) \circ (g_3) \circ (g_4)$$

peut être représentée par l'une des suites :

$$\begin{matrix} 1\ 1\ 2\ 3 \\ 5\ 5\ 1\ 3, \end{matrix}$$

ou $x\ x\ y\ z$, d'une façon générale, si $x \neq y, y \neq z, z \neq x$.

Pour reconnaître si deux suites représentent la même situation d'identité, on les ramènera à une forme canonique où, l'ordre des gènes étant fixé, la $n^{ième}$ classe d'identité qui apparaît est représentée par l'entier n (notations du tableau 1).

Deux suites distinctes qui conduisent ou non à la même forme canonique seront dites « états d'identité ». Ces « états d'identité » peuvent représenter des génotypes :

Si (1) et (2) représentent deux gènes homologues hétéroactifs (allèles), les états (1 1) et (2 2), pour un zygote, sont indiscernables du point de vue de la relation d'identité, mais représentent deux génotypes homozygotes distincts. Ce système de notation a été introduit par BOUFFETTE (1966).

B. — *Calculs associés à l'introduction d'un individu*

Supposons alors qu'une génération H, comportant N zygotes diploïdes,

$$(Z_1, Z_2, \ldots, Z_i, \ldots, Z_j \ldots Z_N),$$

et donc 2N gènes homologues soit décrite par la suite des situations d'identité entre ces 2N gènes et par leurs probabilités, c'est-à-dire par une famille de σ suites de 2N + 1 nombres : 2N entiers et une probabilité :

$$(I_1^\alpha, I_2^\alpha, \ldots, I_{2i-1}^\alpha, I_{2i}^\alpha, \ldots I_{2j-1}^\alpha, I_{2j}^\alpha, \ldots I_{2N}^\alpha ; p^\alpha)$$

$$\alpha = 1, 2, \ldots \sigma.$$

L'étape élémentaire consiste à introduire un nouveau zygote, Z_{N+1}, dont les parents Z_i et Z_j sont dans H. Le gène d'origine paternelle g_{2N+1} est identique à l'un des deux gènes g_{2i-1} et g_{2i} du père Z_i, et chaque éventualité à la probabilité 1/2 en l'absence de toute sélection gamétique et pour un gène autosomal. Par conséquent si les gènes de H étaient dans la situation S^α (ce qui a la probabilité p^α), l'ensemble de gènes :

$$(g_1, g_2, \ldots, g_{2i-1}, g_{2i}, \ldots, g_{2N}, g_{2N+1})$$

sera dans l'état d'identité :

$$(I_1^\alpha, I_2^\alpha, \ldots, I_{2i-1}^\alpha, I_{2i}^\alpha, \ldots I_{2N}^\alpha, I_{2i-1}^\alpha)$$

avec la probabilité $(1/2)\ p^\alpha$ et dans l'état d'identité :

$$(I_1^\alpha, I_2^\alpha, \ldots, I_{2i-1}^\alpha, I_{2i}^\alpha, \ldots I_{2N}^\alpha, I_{2i}^\alpha) \text{ avec la probabilité } (1/2)\ p^\alpha.$$

En regroupant ensuite, parmi les 2σ états obtenus, ceux qui représentent la même situation, on obtient la description de l'ensemble

$$(g_1, g_2, \ldots, g_{2N}, g_{2N+1}).$$

On introduit de la même façon le gène d'origine maternelle de Z_{N+1} ; et si un parent est inconnu, le gène qu'il a transmis sera considéré comme non-identique à tous les autres et représenté par un nombre qui n'apparaît pas dans les situations d'identité initiales.

Ce processus est illustré pour le pedigree de la figure 1 (tabl. 3).

III. — CALCUL APPROCHÉ D'UN COEFFICIENT D'IDENTITÉ

A. — *Nécessité d'une approximation*

Le processus ainsi défini consiste à déterminer toutes les structures géniques d'identité et de non-identité possibles avec leurs probabilités, pour une généalogie donnée. Les calculs associés sont simples et la programmation est aisée, mais le nombre des situations d'identité construites défie la puissance de tout calculateur :

— ce nombre est multiplié par 4 après chaque introduction d'un zygote diploïde, par 4^N après N introductions. La réduction de ce nombre du fait de l'élimination des ancêtres devenus inutiles et du passage aux formes canoniques n'est pas estimé théoriquement mais des calculs complets pratiqués sur de très petites populations (une dizaine de zygotes par génération) ne font pas apparaître de gain sensible ;

— une limite supérieure de ce nombre est donnée par le nombre $S(n)$ des situations d'identité possibles entre n gènes rangés dans un certain ordre. On montre d'abord que si $S(n, p)$ désigne le nombre de situations d'identité entre n gènes pris dans p classes d'identité ($p \leqslant n$), on a la relation :

$$S(n, p) = p\ S(n - 1, p) + S(n - 1, p - 1)$$

avec :

$$S(1,1) = 1$$

TABLEAU 3

Calcul complet des états et des situations d'identité des générations H_0, H_1, K_0, K_1, K_2, K_3

Les situations sont en italique et leurs probabilités notées entre parenthèses. On remarque que la situation (1122) a une probabilité nulle, alors que l'algorithme de Jacquard conduit à la valeur (1/128) car il ne tient pas compte du parcours commun (5)-(3).

H_0	H_1	$H_2 = K_0$		K_1'		K_2'		$K_3 = G_3$	
|6|7|	|7|6|5|	|7|5|4|	|7|5|4|	|7|4|3|	|7|4|3|	|7|3|1|	|7|3|1|	|1|2|	|1|2|
								3431	
								3432	
							123134	3411	*1213*
						125153		3412	(13/32)
	123431	123135	*123134*	123461	*123451*	125154	(1/8)	1431	
		123145		123463		125113		1432	
	(1/4)		(1/4)		(1/4)	125114	*123114*	1411	*1233*
							(1/8)	1412	(1/16)
								3431	
							123434	3432	*1234*
						125353		3441	
		124135		124561		125354	(1/16)	3442	(5/16)
	123441	124145	*123145*	124563	*123453*	125333			
	(1/4)		(1/4)		(1/4)	125334		3531	
								3532	
							123435	3541	*1231*
							(3/16)	3542	(1/16)
1234								4432	
(1)							*123444*	4431	*1211*
								4441	
						125653	(1/16)	4442	(1/16)
		123235		123462		125654			
	123432	123245	*123234*	123463	*123456*	125663			
	(1/4)		(1/4)		(1/4)	125664		4531	
							123445	4532	*1223*
								4541	
							(3/16)	4542	(1/32)
								3431	
							123234	3432	*1123*
								3421	
						125253		3422	
		124235		124562		125254	(1/8)		(1/32)
	123442	124245	*123245*	124563	*123452*	125223		2431	
	(1/4)		(1/4)		(1/4)	125224	*123224*	2432	*1112*
								2421	
							(1/8)	2422	(1/32)

Les quantités $S(n, p)$ sont les nombres de Stirling de deuxième espèce, et l'on a :

$$S(n,p) = \sum_{k=1}^{k=p} \frac{(-1)^{p-k}k^n}{k!(p-k)!}$$

$$S(n) = \sum_{p=1}^{p=n} S(n,p) = \sum_{k=1}^{k=p} \sum_{q=0}^{q=n-k} \frac{(-1)^q k^n}{q!\,k!} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$$

$S(n)$, moment d'ordre n de la loi de Poisson de paramètre 1 est un infiniment grand avec n, d'ordre supérieur à $(n!)$.

B. — *Principe d'un calcul approché*

Pour estimer un coefficient d'identité, on reprend sa définition : une situation d'identité S étant définie relativement à p gènes donnés dans une généalogie, son coefficient d'identité δ est la probabilité que S se réalise dans le processus aléatoire de la transmission des gènes. Par conséquent, si l'on est capable de simuler s fois ce processus, de façons indépendantes, la probabilité que la situation S soit réalisée r fois est :

$$\binom{s}{r}\delta^r(1-\delta)^{s-r}$$

et la quantité $\frac{r}{s}$, rapport du nombre de réalisations de S au nombre d'expériences, est une estimation de δ.

Chaque génération est décrite par s états d'identité possibles et équiprobables. Le p^{ieme} état $E_p(j+1)$ de la génération H_{j+1} est déduit du p^{ieme} état $E_p(j)$ de H_j en retenant un seul état parmi tous ceux qui peuvent être issus de $E_p(j)$: tout gène g introduit est identique à l'un des deux gènes g_1 et g_2 du parent dont il est issu et chaque éventualité $(g = g_1)$ ou $(g = g_2)$ a la probabilité 1/2. On tire un nombre au hasard λ, uniformément distribué sur l'intervalle [0, 1], et l'on fait :

$$g = g_1 \quad \text{si } 0 < \lambda \leqslant 1/2$$
$$g = g_2 \quad \text{si } 1/2 < \lambda \leqslant 1$$

(programme RANDU dans le langage FORTRAN IV).

Si l'on considère que les valeurs $\lambda_1, \lambda_2, \ldots$, générées par RANDU sont une suite de réalisations indépendantes de la même loi uniforme sur [0, 1], on obtient finalement s structures possibles indépendantes et équiprobables de la population. Les plus lointains ancêtres connus étant supposés indépendants, on partira pour les s descriptions, de la structure où chaque gène est non-identique à tous les autres et identifié par le nom du zygote qui le contient et le signe + ou — selon qu'il est d'origine paternelle ou maternelle.

Dans ces calculs, il est inutile de se ramener à la forme canonique des situations d'identité, il suffit de le faire quand, une génération étant décrite par s états, on s'attache à un ensemble donné de gènes (deux ou quatre).

C. — *Propriétés de l'approximation*

A la suite des s simulations du processus, on dispose, pour un couple donné (I, J) dans la généalogie, de la suite de s états d'identité possibles, répartis parmi les

15 situations d'identité possibles. Soient s_1, s_2,..., s_{15} les nombres de réalisations de chacune.

La probabilité de cet événement était de :

$$p = \frac{s!}{s_1!\, s_2!\ldots s_{15}!} (\delta_1)^{s_1}(\delta_2)^{s_2}\ldots (\delta_{15})^{s_{15}}$$

et les estimateurs :

$$\hat{\delta}_i = \frac{s_i}{s}$$

ont les propriétés suivantes, de la loi multinomiale :

$$\mathrm{E}(\hat{\delta}_i) = \delta_i$$

$$\mathrm{Var}\ (\hat{\delta}_i) = \frac{\delta_i(1 - \delta_i)}{s}$$

$$\mathrm{Cov}\ (\hat{\delta}_i\hat{\delta}_j) = -\frac{\delta_i\delta_j}{s}$$

Si l'on considère plusieurs couples, (I, J) et (K, L) par exemple, dans une population, les estimations précédentes des coefficients d'identité, relatifs au couple (IJ) et au couple (KL), sont corrélées du fait qu'elles sont obtenues à partir du même échantillon de structures possibles de la population :

Si $\Delta(\mathrm{IJ}, k\ ;\ \mathrm{KL}, l)$ désigne la probabilité que le couple (IJ) soit dans la situation S_k, et que simultanément le couple (KL) soit dans la situation S_l (il s'agit d'un coefficient d'identité concernant 8 gènes), la covariance entre les estimateurs de δ_k (IJ) et de δ_l(KL), tirées d'un même ensemble de s épreuves simulées, est :

$$\mathrm{Cov}\ (\hat{\delta}_k(\mathrm{IJ})\hat{\delta}_l(\mathrm{KL})) = \frac{1}{s}(\Delta(\mathrm{IJ}, k\ ;\ \mathrm{KL}, l) - \delta_k(\mathrm{IJ})\delta_l(\mathrm{KL}))$$

Cette formule est d'ailleurs tout à fait générale et recouvre les précédentes quand on y fait (IJ) = (KL) avec $k \neq l$ ou $k = l$.

Des expressions du même type concernent l'estimation des coefficients de consanguinité : étant donné un couple (I, J) et s épreuves simulées, r_1 et r_J les nombres d'épreuves où I et J, respectivement, sont en état d'identité, les estimations justes de f_1 et f_J :

$$\hat{f}_1 = \frac{r_1}{s} \qquad \text{et} \qquad \hat{f}_J = \frac{r_J}{s}$$

ont pour moments du deuxième ordre :

$$\mathrm{Var}\,\hat{f}_1 = \frac{1}{s} f_1(1 - f_1)$$

$$\mathrm{Var}\,\hat{f}_J = \frac{1}{s} f_J(1 - f_J)$$

$$\mathrm{Cov}\,\hat{f}_1\hat{f}_J = \frac{1}{s}(\delta_1(\mathrm{IJ}) + \delta_6(\mathrm{IJ}) - f_1 f_J)$$

$\delta_1(\mathrm{IJ}) + \delta_6(\mathrm{IJ})$ est la probabilité que I et J soient simultanément en état d'identité (cf. tabl. 1).

En théorie ces estimations peuvent être rendues aussi précises qu'on le veut, en accroissant le nombre des simulations. En pratique ce nombre, s, est limité par les temps de calcul nécessaires, proportionnels à s et au nombre d'individus dans la population.

Le coefficient de variation de l'estimateur d'un coefficient dont la valeur est δ, étant $\sqrt{\frac{1-\delta}{\delta s}}$, une erreur relative de ε sera atteinte si $\lambda\sqrt{\frac{1-\delta}{\delta s}} \leqslant \varepsilon$, avec $\lambda = 2$ au seuil 5 p. 100, ou $\lambda = 2{,}6$, au seuil de 1 p. 100 ; donc si δ exède une valeur limite :

$$\delta_L = 1/(1 + \frac{\varepsilon^2}{\lambda^2} s)$$

Le tableau 4 présente la valeur δ_L en fonction du seuil λ, de l'erreur relative ε, et du nombre de simulations s.

Au contraire, la longueur des intervalles de confiance, aux seuils λ définis dans les mêmes conditions, est majorée uniformément en δ :

$$l = 2\lambda\sqrt{\frac{\delta(1-\delta)}{s}}, \qquad \text{et} \qquad \delta(1-\delta) \leqslant 1/4 \text{ pour tout } \delta.$$

D'où :

$l \leqslant \frac{\lambda}{\sqrt{s}}$, λ prenant par exemple les valeurs 2, au seuil 5 p. 100, et 2,6 au seuil de 1 p. 100.

La troisième colonne du tableau 4 donne les valeurs de cette longueur.

TABLEAU 4

Valeur limite inférieure δ_L qu'un coefficient d'identité δ doit dépasser pour être estimé, en s simulations, avec une erreur relative inférieure à ε, au seuil de 5 p. 100 ($\lambda = 2$) ou de 1 p. 100 ($\lambda = 2{,}6$)

(l est la longueur de l'intervalle de confiance, indépendante de δ)

S	λ	l	10 %	20 %	33 %	ε
1 000	2	0,063	0,286	0,091	0,035	
	2,6	0,082	0,405	0,145	0,058	
1 500	2	0,052	0,210	0,063	0,023	
	2,6	0,067	0,310	0,101	0,039	
3 000	2	0,037	0,117	0,032	0,012	
	2,6	0,048	0,184	0,054	0,020	
10 000	2	0,020	0,039	0,010	0,004	
	2,6	0,026	0,063	0,017	0,006	

IV. — PROGRAMMATION

La détermination d'une suite de générations G_1, G_2,... dépend essentiellement de la structure de la population P : dans un groupe humain les générations devront être construites arbitrairement alors que dans un élevage de volailles elles sont définies par les plans annuels de croisement. Aussi les programmes réalisés supposent-ils que ces générations sont déjà connues.

La description de la génération G étant écrite sur une bande magnétique B (ou tout autre support), une suite de sous-programmes assure le calcul et l'écriture de la description relative à la génération suivante G' sur une bande B' : lecture de la nouvelle génération ; mise sous forme symbolique de G' en vue des calculs ; calculs des coefficients de parenté et de consanguinité par la méthode de CRUDEN ; calcul par simulation d'un nombre donné *s* d'états d'identité avec l'aide du sous-programme RANDU du langage FORTRAN IV.

Le calcul des coefficients de parenté et de consanguinité par la méthode de CRUDEN exige le recours à une mémoire extérieure (disque magnétique) dont le temps d'accès est long (25 à 70 ms), et la capacité limitée. La capacité des programmes actuels est de 400 ou 800 individus diploïdes par génération, et la vitesse de 2,5 secondes par individus introduit (sur IBM 360-50, équipé d'une mémoire de 128-K octets).

Au contraire le seul calcul par simulation peut admettre des populations plus importantes, avec un temps de calcul proportionnel au nombre d'appels au sous-programme RANDU, et donc au produit du nombre *s* d'états d'identité voulu, par le nombre d'individus introduits.

CONCLUSION

La méthode de calcul des coefficients d'identité que nous avons présentée a été conçue principalement en vue d'une utilisation en Génétique quantitative : elle traite d'une population entière, fractionnée en générations, et non d'un couple isolé de deux individus.

L'usage d'une description relative à toute la population évite de procéder à des calculs particuliers pour tous les couples d'une génération, calculs qui comprendraient une recherche des généalogies communes s'étendant jusqu'aux plus lontains ancêtres connus. Les propriétés de la description utilisée permettent de déterminer la structure d'une génération à partir de celle de la génération précédente sans tenir compte des liens de parenté plus éloignés. La simulation d'un nombre donné d' « états d'identité » équiprobables et indépendants conduit à des valeurs approchées des coefficients d'identité de tout couple d'individus. La précision des calculs par simulation, l'accroissement de la capacité de traitement des programmes sont limités par les temps de calcul nécessaires et les caractéristiques techniques des calculateurs électroniques.

Indépendamment des notions de situations et de coefficients d'identité, on peut utiliser les méthodes de simulation développées ici dans les populations dont les

généalogies sont connues : les noms des classes d'identité sont ceux des ancêtres qui les ont portées les premiers, et leur répartition probable chez les descendants peut mettre en évidence des classes d'individus apparentés, peut orienter une étude statistique des corrélations entre les performances de ces descendants et les probabilités qu'ils ont de porter un gène d'une origine donnée, et aboutir à la mise en évidence statistique d'un facteur héréditaire mendélien important. Inversement, les noms des classes d'identité peuvent être regroupés pour représenter des classes de gènes isoactifs, si certains sont observables (comme les groupes sanguins). L'écart entre la structure probable de la population, établie par le moyen de la simulation, et sa structure réelle, observée, est une manifestation des phénomènes de sélection gamétique, et peut en être une mesure.

Les applications des coefficients d'identité sont rendues possibles par cette méthode de calcul : ainsi la recherche, dans une famille, des individus, qui par leur proximité génétique d'un parent, sont susceptibles de donner un greffon (FEINGOLD *et al.*, 1970) peut être systématisée, bien qu'une méthode de calcul plus légère et adaptée aux généalogies simples qu'on rencontre généralement chez les hommes puisse être préférée. Le calcul dans une population entière des coefficients de parenté et de consanguinité, par la méthode de CRUDEN, ou par la méthode approchée dans les grandes populations, permettra de dégrossir dans un premier temps les effets de la consanguinité dans des populations fermées comme celles des élevages avicoles. Mais c'est surtout l'analyse des corrélations génétiques dans les populations consanguines qui devient possible. Dans ce cas l'expression des covariances génotypiques intra-population ou inter-populations fait intervenir les coefficients d'identité, et quatre ou cinq composantes (GILLOIS, 1964, 1966 *b* ; CHEVALET, 1971). L'estimation de ces composantes est possible, dans une seule population, si les généalogies présentent une certaine structure et la construction des estimateurs peut être faite automatiquement à partir des valeurs calculées des coefficients d'identité et de parenté.

Reçu pour publication en août 1971.

SUMMARY

AUTOMATIC CALCULATION OF IDENTITY COEFFICIENTS

A theory for computing the identity coefficients by means of a simulation procedure is presented here. The population is divided into successive generations, that can overlap ; any generation is described by a given number of equiprobable and independent « identity states » ; the description of a generation is computed from the description of the preceding one, and with help of a pseudo-random numbers generator. Results are approximate and may be improved by increasing the number of the simulated states. With a 128-K 8-bits bytes computer, generations of 800 diploid zygotes are treated at the speed of 2.5 seconds per zygote. This method allows applications in quantitative genetics of inbred populations.

RÉFÉRENCES BIBLIOGRAPHIQUES

BOUFFETTE J., 1966. *Expression de la covariance génotypique chez les tétraploïdes.* Thèse 3e cycle, Fac. Sciences Lyon.

CHEVALET C., 1969. Calcul automatique des coefficients d'identité, de parenté, et de consanguinité. *Ann. Génét. Sél. anim.*, **1**, 181-182.

CHEVALET C., 1971. Calcul *a priori* intra et inter-populations des variances et covariances génotypiques entre apparentés quelconques. *Ann. Génét. Sél. anim.*, **4**, 463-477.

CRUDEN D., 1949. The computation of inbreeding coefficients for closed populations. *J. Hered.*, **40**, 248-251.

EMIK L. O. et TERILL C. E., 1949. Systematic procedures for calculating inbreeding coefficients. *J. Hered.*, **40**, 51.

FEINGOLD N., FEINGOLD J., FREZAL J., DAUSSET J., 1970. Probability of graft compatibility. *Ann. Hum. Genet.*, **33**, 285-290.

GILLOIS M., 1964. *La relation d'identité en génétique.* Thèse, Fac. Sciences Paris, 294 p.

GILLOIS M., 1965. Relation d'identité en génétique. *Ann. Inst. Henri-Poincaré*, B, **2**, 1-94.

GILLOIS M., 1966 *a*. Le concept d'identité et son importance en génétique. *Ann. Gén.*, **9**, 58-65.

GILLOIS M., 1966 *b*. Note sur la variance et la covariance génotypiques entre apparentés. *Ann. Inst. Henri-Poincaré*, B, **2**, 349-352.

JACQUARD A., 1966. Logique du calcul des coefficients d'identité. *Population*, 751-776.

MALECOT G., 1948. *Les Mathématiques de l'hérédité.* Masson, Paris.

MARUYAMA T. et YASUDA N., 1970. Use of graph theory in computation of inbreeding and kinship coefficients. *Biometrics*, **26**, 209-220.