G. MALÉCOT
Professeur à la Faculté des Sciences de Lyon

LES MATHÉMATIQUES DE L'HÉRÉDITÉ

PRÉFACE DE L. BLARINGHEM
Membre de l'Institut

MASSON ET Cie, ÉDITEURS
LIBRAIRES DE L'ACADÉMIE DE MÉDECINE
120, BOULEVARD SAINT-GERMAIN, PARIS (VIe)
1948

AVANT-PROPOS

— *Le but de cet opuscule est d'appliquer le calcul des probabilités à des démonstrations rigoureuses et suffisamment générales d'un certain nombre de formules, classiques ou inédites, de la génétique et de la théorie mathématique de l'évolution. Mais, plutôt que de présenter un schéma unique qui eût paru trop abstrait aux biologistes, j'ai préféré exposer diverses méthodes adaptées chacune à un problème concret; une fois ainsi vulgarisées les notions fondamentales de la génétique mathématique, les bases seront posées pour une indispensable expérimentation, et la voie sera frayée vers une synthèse ultérieure. Je m'excuse des imperfections de cette première ébauche et j'accueillerai avec intérêt toutes les observations et toutes les critiques que l'on voudra bien m'adresser* (1). *J adresse l'expression de ma reconnaissance à M. le professeur G. Darmois et à l'Institut de Statistique de Paris, grâce auxquels ce travail a vu le jour. J'adresse également mes remerciements à M. le professeur Blaringhem, pour ses précieux encouragements, et à la Maison Masson, pour le soin avec lequel elle a édité cet ouvrage.*

N. B. — *Les numéros entre ( ) dans le texte renvoient à la bibliographie à la fin du volume.*

(1) En particulier en ce qui concerne la théorie de la migration, jusqu'à présent inédite, et qui devra être confrontée avec les données expérimentales (voir recherches en cours de M. Lamotte sur *Cepæa Nemoralis*, au laboratoire de l'Ecole Normale Supérieure).

PRÉFACE

*M. Gustave Malécot présente dans cet ouvrage, une condensation très poussée, peut-être même trop poussée des systèmes de la biométrique actuellement en honneur dans l'analyse des problèmes de l'hérédité. J'ai écouté avec intérêt les premières leçons qu'il donna à l'Institut Henri Poincaré aux étudiants et aux adultes inscrits pour l'obtention des Certificats d'aptitude aux applications de la Méthode statistique et au Diplôme de Statisticien délivré par l'*Institut de Statistique *de l'Université de Paris. M. Alfred Barriol, Secrétaire général de la Société de Statistique de Paris pour les opérations financières et les carrières qui exigent ces connaissances, M. Georges Darmois, Professeur de Calcul des Probabilités et Physique mathématique à la Faculté des Sciences, sont plus qualifiés que moi pour en montrer l'intérêt. Mais il s'agit sans doute d'engager les biologistes, les médecins à prendre dans leurs études et leurs observations quelques précautions indispensables pour rendre comparables les mesures et les qualités. C'est à ce seul titre que je donne un chaleureux compliment au normalien, qui depuis 15 ans étudie en mathématicien expert les applications possibles du calcul des probabilités à la prévision des résultats de mariages consanguins ou même de la ségrégation de races locales ou même de lignées animales ou végétales soumises à une sélection prolongée. Pour les médecins, la lecture de cet ouvrage leur donnera à la fois la prudence dans l'application des formules et une audace, actuellement à peine éveillée, dans la recherche des tares et des aptitudes plus ou moins héréditaires dans les populations humaines.*

Francis Galton, vers 1880, après Ad. Quetelet (1835), avant Alphonse Bertillon dont les Instructions signalétiques *(1893) sont en usage familier, a propagé une mode qui pourrait inspirer nos cinéastes; il réalisa les* « portraits composites ». *Tous les Indiens, dit-il, ont, à première vue, des traits communs qui les distinguent des autres races; les Anglais de même ne sont pas confondus avec les Espagnols; et surtout, les membres d'une même famille ont*

dans leurs traits, dans leurs gestes, dans leur marche une allure qui malgré l'âge et le sexe donne un « air de famille ». *Galton réunit les photographies des individus de chaque classe, les superpose après les avoir ramenées à la même dimension quant à l'écart des yeux, visage sur visage, œil sur œil, et place les épreuves devant l'objectif et une plaque à développement lent, d'abord le portrait le plus récent et sur la même plaque le suivant en date, le troisième... Un développement unique réunit tous les traits non pas en un portrait informe, mais en une image où chacun se retrouve avec les traits communs renforcés qui laisse imperceptibles les particularités individuelles; rides, cheveux, barbe ne forment qu'un lavis léger. S'il peut superposer le grand-père jusqu'aux petits-enfants il obtient le* portrait composite *de cette descendance qui laisse deviner ce que seront les suivants.*

Les génétistes modernes s'attachent au contraire à suivre un caractère, la couleur des cheveux, celle des yeux..., la forme des doigts, l'implantation des ongles et avec les entraînements de l'époque, ce serait plutôt le portrait du talent de coiffeur, comme la silhouette trahirait l'art du tailleur d'habits. En bref, il faut avant toute analyse préciser l'objet et c'est la raison qui fait suivre de préférence le caractère saillant anormal, celui qui trahit une mutation, une substitution de gènes, ou même son absence. C'est dans une substitution, ou plutôt une réduction extrême de la diastase amylogène qui distingue le Pois ridé du Pois rond qui mit Grégor Mendel (1865) sur la voie des règles de la transmission des caractères dits indépendants; la règle mendélienne domine toutes les études de génétique modernes et se vérifie chez les Pois, comme chez les Souris, lorsqu'il s'agit de couples de caractères parfaitement distincts et avec des lignées contrôlées durant cinq à dix générations.

Lorsqu'il s'agit du Pois, l'étape amylacée (rond) de l'embryon A domine l'état sucré, dit récessif a *à la première génération; la disjonction se produit à la seconde génération suivant les règles mathématiques du calcul des probabilités dans les proportions* nA + 2nAa *qui sont ronds et* naa *qui sont sucrés, ridés quoique bien mûrs. Mlle Cécile Bourdouil guidée par Marc Bridel (1931) a pu démontrer qu'il s'agit uniquement de l'absence, ou plutôt de l'inactivité d'une diastase qui ne nuit en rien à la fécondité; la diastase est un corps qu'on sait isoler et conserver actif hors de l'embryon; c'est la définition même de l'*indépendance *notée avec soin par Grégor Mendel. C'est aussi le trait qui définit le Maïs sucré de Hugo de Vries par rapport au Maïs amylacé, dans la mise en valeur des*

règles de Mendel en 1899, point de départ définitif des études de Génétique, et dont la portée générale fut établie par les découvertes à la même date de la double fécondation par Nawaschine et Léon Guignard (1).

L'ouvrage de M. G. Malécot exprime très clairement cette répartition mathématique des descendances de la Belle de Nuit dont le pigment rouge domine l'absence dans la variété à fleurs blanches; la forme des crêtes du Coq exige pour être analysée la présence de trois facteurs (diastases) indépendants et la teinte très variée des mulâtres un nombre plus grand encore et certainement variable avec les populations examinées. Cette étude exige des précautions de langage et avec les découvertes cytologiques modernes une précision qui est la distinction des facteurs, *des* gènes, *des* loci. *J'insiste tout particulièrement sur les précautions prises dans les définitions.*

Le second chapitre fait apparaître les dangers des unions consanguines et c'est un résultat précis d'établir que le danger est doublé pour un enfant de doubles cousins germains, « qu'il est illogique de tolérer le mariage des doubles cousins germains et le mariage oncle-nièce et d'interdire le mariage entre demi-frères, qui présente exactement le même danger ». Le paragraphe intitulé Les variables aléatoires mendéliennes dans une population homogène isogamique stationnaire *est aussi de la plus haute portée pratique; la variabilité héréditaire est maintenue et cette démonstration justifie les enseignements de W. Johannsen (1909) dont j'ai eu connaissance auprès du Maître à Copenhague en février 1903, ce qui m'a aidé dans le choix des lignées pures d'Orges de Brasserie.*

Je ne puis qu'approuver l'effort très sérieux de l'auteur dans ses analyses des problèmes soulevés sur l'origine des races à population limitée ou en régression et du rapprochement qu'il fait de ses résultats avec ceux qui sont obtenus par la sélection dans le domaine des races animales et des plantes de grande culture ne supportant pas l'autofécondation prolongée. J'en ai des exemples très nets dans des lignées hybrides de Seigles vivaces et de Seigles annuels et la préparation de semences de Betteraves à sucre aboutit à des conclusions analogues. Et le paragraphe intitulé Influence de la Migration *suggère des aperçus sur les réussites modernes, obtenues par exemple à Lyon avec la Rose de Chine qui fournit les* Pernettia *si répandues et appréciées depuis un demi-siècle, sur la*

(1) **Voir une mise au point *Hérédité, Mutation et Evolution*, l'œuvre de Hugo de Vries au Palais de la Découverte, 1937, chez Masson et Cie.**

variabilité presque indéfinie des Dahlias et des Chrysanthèmes, en un mot sur tout ce qui touche à l'agriculture et à l'horticulture ; sans doute aussi aux races chevalines et aux groupements humains.

Un tel effort mérite l'éloge et provoque le souhait d'obtenir dans la prochaine décade le développement des applications possibles des formules mathématiques dans tous les domaines de la biologie.

Louis Blaringhem.

CHAPITRE PREMIER

LA LOTERIE MENDÉLIENNE

HÉRÉDITÉ ET LOIS DE MENDEL

Rappelons les lois de Mendel, en prenant l'exemple de la belle de nuit (*mirabilis jalapa*). Si nous en croisons un pied à fleurs blanches et un pied à fleurs rouges, nous n'obtenons que des pieds à fleurs roses. Mais si nous croisons entre eux ces pieds à fleurs roses, ils donnent naissance en moyenne à 1/4 de fleurs blanches, 1/2 de fleurs roses, 1/4 de fleurs rouges. Il y a réapparition des caractères des grands-parents. C'est le phénomène de la *disjonction mendélienne* ou *ségrégation*. Ce phénomène s'explique en admettant que la couleur des fleurs de la belle de nuit est conditionnée par une paire d'unités héréditaires ou *facteurs* qui peuvent présenter chacune l'un ou l'autre des 2 états ou *gènes* que nous désignerons par A ou *a*, de sorte qu'un pied pourra être porteur de la paire :

AA	auquel cas il sera	à fleurs	rouges.
A*a*	»	»	roses.
aa	»	»	blanches.

Les 3 états que peut ainsi présenter la paire sont appelés les *génotypes* ou *zygotes*. AA et *aa* sont les *homozygotes*, A*a* l'*hétérozygote*.

Le croisement se traduit par le mécanisme suivant : la paire de facteurs de chaque plante issue du croisement, de chaque « enfant » (dirons-nous) s'obtient en tirant au sort, à pile ou face si l'on veut, un des 2 facteurs du père et un des 2 facteurs de la mère.

Le croisement d'un AA avec un *aa* ne donne donc que des A*a* (1re génération).

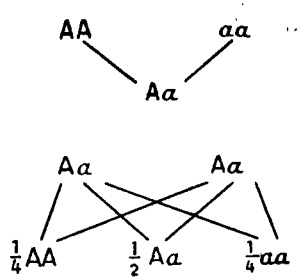

Fig. 1.

Mais le croisement des A*a* entre eux donne :

AA avec probabilité 1/4 (puisqu'il faut tirer A chez chaque parent).
A*a* » 1/2
aa » 1/4.

Cette interprétation concorde bien avec les fréquences observées à la 2e génération sur un grand nombre d'individus.

Les lois de Mendel se sont montrées remarquablement explicatives pour tous les phénomènes de l'hérédité, et il est permis de considérer qu'à de rares exceptions près toute l'hérédité se ramène au mécanisme mendélien. Mais il faut pour cela admettre que la présence des gènes peut se manifester d'une façon différente de celle que nous venons d'exposer.

A) Reprenons l'exemple donné par Mendel du croisement des petits pois à grains lisses et ridés. En 1re génération, on n'a que des petits pois à grains lisses ; en les croisant entre eux, on obtient 3/4 de grains lisses et 1/4 de grains ridés. Ce cas rentre parfaitement dans le schéma précédent, à condition d'admettre qu'à la fois les AA et les A*a* sont à grains lisses, seuls les *aa* étant à grains ridés.

Dans ce cas l'hétérozygote A*a* a même apparence extérieure qu'un des homozygotes, et ne s'en différencie que par les caractères de sa descendance, ce qui nous conduit à opposer au *génotype* ou constitution héréditaire, le *phénotype* ou apparence extérieure. Dans le cas considéré les 3 génotypes ne donnent que 2 phénotypes. On dit qu'il y a *dominance* du gène A sur le gène *a*, ou que *a* est *récessif*. L'hétérozygote présente le même phénotype que l'homozygote dominant.

La dominance peut être incomplète, l'hétérozygote étant plus voisin d'un des homozygotes, mais s'en distinguant néanmoins.

B) Il existe des caractères conditionnés par plusieurs paires de facteurs, ou *plurifactoriels*. Par exemple, la forme de la crête des coqs dépend de 3 paires de facteurs : la 1re paire avec les gènes C (présence de crête) dominant sur *c* (crête rudimentaire) ; la 2e paire R (en rose) dominant sur *r* (dentelée) ; la 3e D (double) dominant sur *d* (simple). Par suite de la dominance, les génotypes :

C C R R *d d*
C *c* R R *d d*
C *c* R *r d d*
etc.

donneront le même phénotype : crête en rose, mais :

C C *r r d d*
C *c r r d d*

donnent la crête dentelée,

c c r r D D
c c r r D *d*

donnent le type de Bréda (crête rudimentaire double).

L'étude des croisements montre que la ségrégation se fait indépendamment pour les différentes paires. Par exemple le croisement d'une race à crête en rose, doublement hétérozygote, Cc*Rrdd*, avec une Bréda *cc rr* DD donc des individus qui présentent tous le couple D*d*, mais qui présentent :

C*c* ou *cc* avec probabilité 1/2
R*r* ou *rr* avec probabilité 1/2

donc, puisque ces 2 paires subissent des ségrégations indépendantes il y aura en moyenne :

1/4 de C *c* R *r* D *d*
1/4 de *c c* R *r* D *d*
1/4 de C *c r r* D *d*
1/4 de *c c r r* D *d*

C) Dans tous les cas qui précèdent, les caractères observés ne présentaient qu'une échelle discontinue d'états. De cette hérédité « alternative », Karl Pearson, qui fut avec Galton le fondateur de la biométrie, distinguent l'hérédité « continue » ou « mélangée », comme par exemple celle de la taille ou de la couleur de la peau dans l'espèce humaine. Si on observe un nombre suffisant d'enfants issus d'un couple déterminé, on constate que les tailles des enfants se groupent autour d'une valeur moyenne dépendant des tailles des 2 parents, suivant une courbe en cloche, les grands écarts étant rares mais néanmoins possibles. Il semble y avoir un mélange des caractères des parents, compliqué par des fluctuations. De même le croisement de mulâtres fait apparaître chez leurs descendants une grande variété de teintes, groupés surtout autour d'une teinte intermédiaire, mais avec de temps en temps un type franchement noir ou franchement blanc. Tous ces résultats s'expliquent parfaitement par les lois de Mendel, en admettant que la taille ou la couleur de la peau, résultent de l'addition des effets d'un grand nombre n de facteurs mendéliens se disjoignant indépendamment. Pour fixer les idées, si dans chaque paire de facteurs les 2 gènes possibles ont pour effets respectifs d'ajouter ou de retrancher 1 mm. à la « taille normale » et qu'on croise 2 individus dont toutes les paires sont hétérozygotes, $A_1a_1A_2a_2 \ldots A_na_n$, il y aura chez chaque enfant les probabilités :

1/4 1/2 1/4

pour que chaque paire présente les états :

$$A_iA_i \qquad A_ia_i \qquad a_ia_i$$

et apporte donc à la taille les contributions :

2 mm. 0 mm. — 2 mm.

comme si cela résultait de 2 parties indépendantes de pile ou face.

Si les n différentes paires sont stochastiquement indépendantes, la taille d'un enfant ne sera autre que la somme des gains ou des pertes dans $2n$ parties indépendantes de pile ou face, somme dont la loi de probabilité est, comme nous savons, pour n infiniment grand, une loi de Gauss, qui s'ajuste bien à la courbe en cloche expérimentale. Nous verrons qu'il en est encore de même si on suppose plus généralement une dominance complète ou incomplète dans chaque paire, et des contributions différentes pour les différentes paires. Ce schéma général de l'*hérédité mendélienne multifactorielle* sera développé en détail dans le chapitre II, comme expliquant tous les résultats de la *biométrie* fondée par Galton et Pearson (5) (15) (1).

LES CHROMOSOMES

La base physiologique de l'hérédité mendélienne a été découverte dans les bâtonnets ou *chromosomes* qui constituent le noyau des cellules reproductrices ou *gamètes*. Ces chromosomes sont en nombre fixe N (24 chez un homme, 4 chez la drosophile), et présentent parfois entre eux des différences qui permettent de reconnaître dans 2 gamètes différentes les chromosomes *homologues* (voir figure ci-contre pour la drosophile). Quand une gamète du père et une gamète de la mère s'unissent l'œuf ainsi obtenu présente N paires de chromosomes homologues, et cette constitution subsiste dans toutes les cellules que l'œuf produit par division ; donc finalement dans toutes les cellules de l'individu adulte, à l'exception de ses cellules reproductrices ou *gamètes* : celles-ci sont engendrées par une division ou *disjonction* qui ne laisse subsister de chaque paire, qu'un chromo-

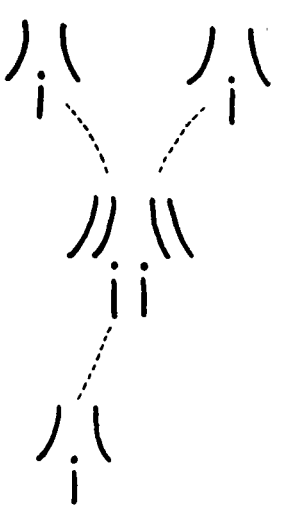

Fig. 2.

(1) Les numéros renvoient à la bibliographie, à la fin du volume.

some, prélevé au hasard parmi les deux ; leur union au hasard avec les gamètes du conjoint donne les individus de la génération suivante.

Les lois de Mendel s'expliquent alors en admettant que les 2 facteurs d'une paire sont portés par 2 chromosomes homologues (voir figure). 2 parents hétérozygotes A*a* formeront chacun moitié de gamètes A et moitié de gamètes *a* : c'est la *disjonction*. L'union au hasard de ces gamètes donnera bien 1/4 d'individus AA, 1/2 d'individus A*a*, 1/4 d'individus *aa*.

Mais une difficulté se présente : l'indépendance des différentes paires de facteurs se conçoit si elles sont portées chacune par des paires différentes de chromosomes, car il est naturel de supposer que ces dernières subissent des disjonctions indépendantes.

Mais des facteurs portés par la même paire de chromosomes devraient être complètement liés, comme ceux ci-contre. En réalité en croisant un individu du type ci-contre avec l'homozygote $\begin{matrix} a \\ b \end{matrix}\Big| \Big| \begin{matrix} a \\ b \end{matrix}$, on trouve les génotypes :

A | a
B | b

AB*ab* *ab**ab* A*bab* *a*B*ab*

dans les proportions :

$$\frac{1-r}{2} \qquad \frac{1-r}{2} \qquad \frac{r}{2} \qquad \frac{r}{2}$$

r étant en général un nombre positif faible, mais non nul comme s'il y avait liaison absolue ; ceci ne peut s'expliquer que par une rupture des chromosomes avec échange des parties homologues avant disjonction, échange qui se produirait avec la probabilité *r*, ce qui entraîne bien pour les 4 types de gamètes formés les proportions ci-dessus ; c'est le phénomène de « *l'enjambement* » ou « crossing over ». L'analyse de ce phénomène a conduit à supposer que chaque facteur est localisé sur le chromosome en un point déterminé ; son *locus* (pluriel *loci*). Pour 2 facteurs portés par le même chromosome, *r*, probabilité d'enjambement, sera d'autant plus grand qu'ils seront plus éloignés, ce qui a permis, pour les 4 paires de chromosomes de la drosophile, de dresser la carte des loci des différents facteurs. Nous verrons au chapitre suivant que, du fait de l'enjambement la liaison (linkage) des facteurs sur un même chromosome ne les empêche pas à la longue d'être répartis comme s'ils étaient indépendants.

Le sexe est déterminé par un couple de facteurs mendéliens : XX pour la femelle et XY pour le mâle, sauf chez les lépidoptères et les oiseaux. Les 2 chromosomes qui les portent sont appelés les hétérosomes, les autres les autosomes. Les autres facteurs portés par les

hétérosomes sont dits « sex linked » ; on connaît principalement ceux qui sont portés par le chromosome X qui ne sont jamais masqués chez l'homme mais peuvent l'être chez la femme (ex. : daltonisme, hémophilie).

Quel que soit le processus physiologique par lequel les gènes présents dans les chromosomes agissent sur le développement de l'individu, il est naturel qu'une seule paire de gènes puisse agir sur plusieurs caractères par exemple : le gène récessif « albinos » a provoqué chez l'homozygote récessif *aa*, à la fois le pelage blanc et les yeux rouges (par privation de pigment). Inversement nous avons vu qu'un seul caractère comme la taille peut être influencé par un grand nombre de paires de gènes. En toute rigueur, chaque caractère dépend de l'ensemble des gènes présents, de l'ensemble de la « constitution génétique », mais pratiquement il arrive qu'un caractère ne soit influencé de façon appréciable que par un seul couple de gènes, et c'est ce qui a permis l'observation des lois de Mendel.

Nous avons admis jusqu'ici que les gènes occupant un locus déterminé ne pouvaient présenter que 2 états différents, que nous avons représentés par A et *a*, dits gènes *allèles*. En réalité, ils peuvent aussi présenter des états multiples A', A'', A''', ..., $A^{(n)}$, on dit alors qu'il y a *multiallélisme*. Il y a n homozygotes et $C_n^2 = \frac{n(n-1)}{2}$ hétérozygotes (1). L'origine des gènes allèles est à chercher dans le phénomène de la *mutation*, qui paraît être (du moins à notre échelle d'observation, car on ne peut préjuger de ce qui s'est passé à l'échelle des temps paléontologiques) le seul mode de modification des individus qui soit héréditaire, donc le seul qui ait une influence sur l'évolution de l'espèce. La mutation est un changement brusque affectant chez un individu, l'un de 2 loci homologues, donc transmis à la moitié de ses gamètes. Il apparaît ainsi dans une population d'individus tous homozygotes AA, donc indiscernables quant à la paire considérée, un hétérozygote A*a* qui, même s'il y a dominance de A, peut se révéler par sa descendance. La mutation a fait apparaître le gène nouveau *a* allèle du gène ancien A. Des mutations répétées affectant un même locus peuvent créer toujours le même gène *a* (mutation récurrente), ou ramener le gène *a* au gène A (mutation de retour) ou créer d'autres allèles (multiallélisme).

(1) Par exemple les 4 groupes sanguins de l'homme sont conditionnés par 3 allèles, A et B dominant sur O, donnant les 4 phénotypes A (AA ou AO), B(BB ou BO), AB (receveur universel), O(OO) (donneur universel).

LA RESSEMBLANCE ENTRE INDIVIDUS APPARENTÉS

On dit que 2 individus d'une population sont *apparentés* (1) s'ils ont un ou plusieurs ancêtres communs. Leur différence génétique doit être moins grande en moyenne que celle de 2 individus quelconques puisque certains de leurs gènes descendent d'un même gène d'un même ancêtre commun et ne peuvent donc, si on néglige les mutations, être différents, alors qu'ils pourraient toujours l'être chez des individus non apparentés.

Pour préciser le langage, nous distinguerons *facteurs*, *gènes* et *loci*. Nous appellerons gènes les différents états que peut présenter chaque facteur, sans considérer chez quel individu ils sont observés : 2 gènes correspondant au même facteur et observés soit chez le même individu, soit chez 2 individus différents seront dits identiques ou différents suivant qu'ils présentent le même état, A par exemple, ou 2 états allèles, par exemple A et *a*. Au contraire 2 loci seront dits identiques seulement s'ils dérivent par descendance mendélienne d'un même locus, d'un même ancêtre commun, sinon ils seront dits différents ; 2 loci identiques sont forcément occupés par des gènes identiques, s'il n'y a pas de mutations, mais 2 loci différents peuvent être occupés par 2 gènes identiques ou différents.

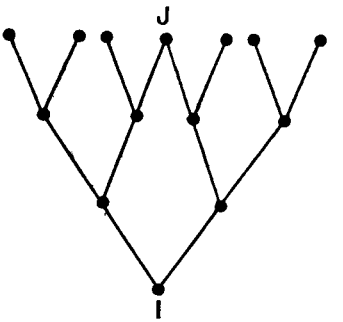

Fig. 3.

Tout individu I a 2 parents, 4 grands-parents, ..., 2^n ancêtres d'ordre n, dont certains peuvent être confondus. Un locus de I a les probabilités 1/2 de provenir du père, 1/2 de la mère, 1/4 de chacun des grands-parents, $\frac{1}{2^n}$ de provenir d'un certain ancêtre d'ordre n le long d'une *chaîne d'ascendance* déterminée (Un ancêtre de I peut lui être relié par plusieurs chaînes d'ascendance ; exemple : J sur la figure, on dit alors que c'est un ancêtre multiple ; il peut même être ancêtre d'ordre différent sur les différentes chaînes).

Nous appellerons *coefficient de parenté* f_{IL} de 2 individus I et L la

(1) Nous réserverons le nom de *parents* au père et à la mère.

probabilité pour que 2 loci homologues pris l'un sur I, l'autre sur L soient identiques, c'est-à-dire descendent d'un même locus. La probabilité complémentaire $1 - f_{IL}$ représente la probabilité pour que ces 2 loci proviennent d'ancêtres sans aucune parenté, c'est-à-dire soient stochastiquement indépendants (car alors la connaissance du gène qui occupe l'un ne donne aucun renseignement sur le gène qui occupe l'autre ; ces 2 gènes peuvent être identiques ou différents mais leurs probabilités sont indépendantes).

Nous appellerons *coefficient de consanguinité* f_M d'un individu M la probabilité pour que ses 2 loci homologues soient identiques. Comme l'un provient de son père et l'autre de sa mère, f_M n'est autre que le coefficient de parenté de ses 2 parents.

Evaluons le coefficient de parenté f_{IL} de 2 individus I et L. Il n'est $\neq 0$ que si I et L ont un ou plusieurs ancêtres communs A_1, A_2, etc., ce que nous supposerons. Supposons d'abord qu'il y en ait un seul, A, ancêtre d'ordre n de I et d'ordre p de L par 2 chaînes d'ascendance uniques dont l'ensemble constitue une *chaîne de parenté* reliant I et L.

La probabilité pour qu'un locus de I et un locus homologue de L proviennent tous deux de A est $(1/2)^{n+p}$; mais dans cette éventualité ils ont la probabilité 1/2 de provenir du même locus de A, et la probabilité 1/2 de provenir de loci différents auquel cas ils ne sont identiques qu'avec la probabilité f_A. D'où $f_{IL} = (1/2)^{n+p} \frac{1 + f_A}{2}$.

En particulier le coefficient de parenté d'un individu avec un ancêtre simple d'ordre n correspond à $p = 0$; le coefficient de parenté d'un individu avec lui-même à $n = p = 0$.

Traitons maintenant le cas général où I et L sont reliés par un nombre quelconque de chaînes de parenté, chaque chaîne étant la réunion de 2 chaînes d'ascendance remontant de I et L à un ancêtre commun A_i et n'ayant pas d'autre point commun que A_i ; 2 chaînes de parenté sont regardées comme distinctes même si elles ont une partie commune, pourvu qu'elles diffèrent par au moins un chaînon. Comme la transmission de loci identiques le long d'une chaîne de parenté déterminée exclut leur transmission le long de toute autre, le principe des probabilités totales donne :

$$f_{IL} = \Sigma(1/2)^{n_i+p_i}(1 + f_{A_i})/2.$$

La somme Σ *étant étendue à toutes les chaînes de parenté distinctes reliant* I *et* L, *la* i*ème comprenant* $n_i + p_i$ *chaînons et remontant à l'ancêtre commun* A_i *de coefficient de consanguinité* f_{A_i}.

Exemple : dans le cas de la figure ci-contre, en supposant que toutes les chaînes sont indiquées et que les têtes de généalogie sont sans consanguinité, les différentes chaînes et leurs contributions au coefficient f_{GF} sont :

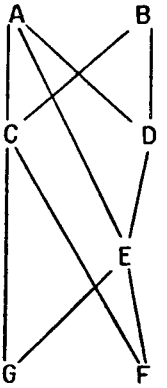

Fig. 4

GCF : 1/8 ; GEF : 5/32 ; GCAEF : 1/32 ;
FCAEG : 1/32 ; GCADEF : 1/64 ; FCADEG : 1/64 ;
GCBDEF : 1/64 ; FCBDEG : 1/64.

Donc :

$$f_{GF} = 13/32.$$

Supposons maintenant que les loci considérés puissent être affectés par des mutations de fréquence moyenne u par génération. La probabilité pour qu'un locus d'un individu reproduise sans modification le locus parental dont il dérive est alors $1 - u$; donc la probabilité pour qu'il se soit transmis sans modification le long d'une chaîne d'ascendance déterminée comportant n chaînons est $\left(\frac{1-u}{2}\right)^n$; le coefficient de parenté de 2 individus reliés comme indiqué ci-dessus n'est donc plus que $\Sigma[(1-u)/2]^{n_i+p_i}(1+f_{A_i})/2$; la correction ainsi introduite est d'ailleurs insignifiante pour les parentés rapprochées, car u est extrêmement faible ; elle ne prend d'influence, comme nous le verrons, que quand on fait intervenir des parentés remontant très haut.

CHAPITRE II

LES CORRÉLATIONS ENTRE APPARENTÉS

(DANS UNE POPULATION ISOGAMIQUE STATIONNAIRE)

Les probabilités des gènes et des génotypes.

Classons les individus d'une population F suivant l'état d'une paire déterminée de facteurs. Supposons d'abord qu'il n'y ait que 2 gènes allèles A et a, on observe alors les 3 génotypes AA, Aa, aa avec les fréquences respectives P, 2Q, R($P + 2Q + R = 1$). Nous appellerons « fréquences des gènes A et a » les quantités $p = P + Q$ et $q = Q + R$($p + q = 1$).

Ce sont les probabilités pour qu'un gène pris au hasard chez un individu quelconque de F présente les états A ou a. Dans chaque individu I de la population F, chacun des 2 loci homologues sera occupé par un gène A ou a avec la probabilité p ou q, mais il y aura en général une liaison entre les probabilités pour ces 2 loci, c'est-à-dire une corrélation entre les états de ces 2 loci ; cette corrélation est due à ce que la connaissance du gène qui occupe un des loci modifie en général les probabilités relatives à l'autre locus ; en effet les 2 conjoints de la génération précédente dont ils descendent respectivement peuvent s'être choisis suivant leur parenté (consanguinité), suivant leur ressemblance (homogamie) ou avoir laissé des descendants avec une loi de probabilité dépendant de leurs génotypes (fécondité différentielle) ; dans tous ces cas tout renseignement sur le génotype d'un des conjoints modifie les probabilités pour l'autre. Nous nous en tiendrons dans ce chapitre aux 2 cas suivants :

a) les conjoints se choisissent au hasard, la probabilité de trouver un conjoint est la même pour tous les individus, et la fécondité est la même pour tous les couples ; « random mating », « croisement au

hasard », *panmixie*. Alors la connaissance du gène qui occupe un des 2 loci de I ne nous apprend rien sur l'autre, les états de ces 2 loci sont stochastiquement indépendants. I présente donc les 3 génotypes AA, A*a*, *aa*, avec les probabilités p^2, $2pq$, q^2. Si la population est nombreuse, les fréquences observées P, 2Q, R, doivent être voisines de ces quantités. Pour le vérifier, il suffira de montrer que $Q^2 - PR$ est voisin de 0 (loi de Hardy), car on peut toujours poser $P = p^2 + \lambda$, $2Q = 2pq - 2\mu$, $R = q^2 + \gamma$; et puisqu'on a posé $P + Q = p$ et $Q + R = q (p + q = 1)$ on a $\lambda = \mu = \nu$, or :

$$Q^2 - PR = (pq - \lambda)^2 - (p^2 + \lambda)(q^2 + \lambda) = -\lambda$$

n'est nul que si $\lambda = 0$. On constate qu'il existe effectivement des populations naturelles où la loi de Hardy est vérifiée ; telle la population de coléoptères *Dermestes vulpinus* observée par Haldane et Philip (7) (le couple de facteurs étudié étant celui qui conditionne la couleur des ailes). Nous verrons qu'il en est de même pour l'espèce humaine en ce qui concerne les groupes sanguins.

b) les conjoints ne se choisissent que suivant leur consanguinité, sans tenir compte autrement de leur génotype ou de leur ressemblance ; la probabilité de trouver un conjoint est la même pour tous les individus et tous les couples ont la même fécondité : *consanguinité pure* ou *isogamie*. Alors un locus de n'importe quel individu, issu ou non d'un croisement consanguin, a toujours les mêmes probabilités p et q de porter les gènes A ou *a* ; de plus pour tout individu I dont on connaît le coefficient de consanguinité $f_I = f$, les 2 loci homologues ont comme nous avons vu la probabilité f d'être identiques et la probabilité $1 - f$ d'être stochastiquement indépendants ; donc les 3 génotypes chez cet individu ont les probabilités :

$$fp + (1 - f)p^2 = p^2 + fpq \qquad 2(1 - f)pq \qquad fq + (1 - f)q^2 = q(q + fp)$$

(Pour qu'on ait par exemple le 1er génotype, il faut que les 2 loci soient identiques et que l'un soit A, ou qu'ils soient indépendants et que tous deux soient A).

La consanguinité se traduit donc par une augmentation appréciable de la probabilité des homozygotes et une diminution de la probabilité des hétérozygotes. On a là l'explication du danger des mariages consanguins : les tares latentes dans l'espèce humaine sont en général conditionnées par des gènes récessifs rares, elles n'apparaissent donc que chez l'homozygote récessif *aa*. Si q est la fréquence, faible, du « gène taré » *a*, la probabilité pour qu'un individu I soit taré, c'est-à-dire du type *aa*, sera égale à q^2, c'est-à-dire extrêmement faible, si les parents de I ne sont pas apparentés ; mais elle monte à $q(q + fp) \# fq$

si I présente un coefficient de consanguinité f appréciable. Par exemple une tare produite par un gène de fréquence $q = 10^{-4}$ apparaîtra avec la probabilité 10^{-8} chez un enfant sans consanguinité, mais avec la probabilité $10^{-4}/16$ chez un enfant issu de cousins germains ($f = 1/16$) (1). Pour un enfant de doubles cousins germains ($f = 1/8$) le danger est doublé. Il est illogique de tolérer le mariage des doubles cousins germains et le mariage oncle-nièce et d'interdire le mariage entre demi-frères, qui présente exactement le même danger ($f = 1/8$) [Haldane (6)].

Traitons maintenant le cas plus général du *multiallélisme*. Supposons que les gènes allèles A^i aient les fréquences $p_i (\Sigma p_i = 1)$.

a) Dans le cas de la panmixie, les probabilités des divers génotypes sont :

pour un homozygote tel que : $A^iA^i : p^2$;
pour un hétérozygote tel que : $A^iA^j : 2p_ip_j$

(ce sont les coefficients du développement de $(\Sigma p_i l_i)^2$) (Ces formules donnent de bons ajustements pour la fréquence des groupes sanguins dans une population homogène ($p^2 + 2pr$, $q^2 + 2qr$, $2pq$, r^2)).

b) Dans le cas plus général d'isogamie, pour un individu de coefficient de consanguinité f, ces probabilités sont respectivement :

$$fp_i + (1-f)p_i^2 = p_i^2 + fp_i(1-p_i) \qquad \text{et} \qquad 2(1-f)p_ip_j$$

(ce sont les coefficients du développement de : $f\Sigma p_i l_i^2 + (1-f)(\Sigma p_i l_i)^2$).

La répartition des facteurs dans une population isogamique.

Nous appellerons *population isogamique* une population F dont les parents ne se sont choisis qu'au hasard ou en raison de leur parenté, mais à l'exclusion de toute homogamie, et où tous les couples ont eu la même fécondité. Nous supposons que la proportion de couples ayant un coefficient de parenté f_i est w_i (une proportion w_0 correspondant à l'accouplement au hasard $f_0 = 0$). w_i est donc la fréquence dans la population des individus de coefficient de consanguinité f_i. Nous avons vu que les probabilités des gènes allèles A et a (en en supposant deux seulement pour simplifier) sont les mêmes parmi ces individus que dans la population totale, soient p et q.

(1) Effectivement on constate que la moitié des cas d'ataxie de Friedreich proviennent de mariages consanguins, ainsi que le 1/3 des cas d'albinisme.

Les probabilités des trois génotypes dans la population totale, et par suite sensiblement leurs fréquences P, 2Q, R si la population est nombreuse, sont donc :

$$\Sigma\omega_i p(p+f_i q) \qquad \Sigma 2\omega_i pq(1-f_i) \qquad \Sigma\omega_i q(q+f_i p)$$

ce qui s'écrit aussi :

$$p(p+\alpha q) \qquad 2pq(1-\alpha) \qquad q(q+\alpha p)$$

en posant $\alpha = \Sigma\omega_i f_i$. α est le « coefficient de consanguinité moyen de la population », moyenne pondérée des coefficients de ses individus. Il avait été introduit *a priori* par Bernstein (1) pour mesurer l'écart par rapport à la panmixie. Son évaluation approchée a été tentée pour quelques populations humaines, à l'aide du relevé de l'état-civil des mariages consanguins. Il est en général faible ; pour une population paysanne autrichienne Reutlinger a trouvé 0,6 0/0, Orel a trouvé un peu plus de 1 0/0 pour une population israélite... Mais cette estimation est probablement très inférieure à la réalité, car les parentés éloignées, qui sont négligées, jouent un rôle aussi grand que les parentés rapprochées.

Après la répartition d'une paire de facteurs étudions maintenant la répartition simultanée dans la population F de *2 paires de facteurs* occupés par des gènes présentant respectivement les états A_i avec les probabilités p_i et les états B_j avec les probabilités x_j.

Un individu I pris au hasard dans F résulte de l'union de 2 gamètes issues de la génération précédente F'. Appelons P'_{ij} la probabilité pour qu'une gamète quelconque Γ' issue de F' présente dans ses chromosomes les gènes A_i et B_j, et P_{ij} la probabilité pour qu'il en soit de même pour une gamète quelconque Γ issue de F, c'est-à-dire pour une gamète produite par I. Cherchons la relation entre P_{ij} et P'_{ij}. Quand la gamète Γ produite par I présente les gènes A_i et B_j, ils proviennent, ou bien tous deux de la même gamète Γ', ou bien chacun d'une des gamètes Γ' ayant constitué I ; ces 2 éventualités présentent respectivement les probabilités 1/2 et 1/2 si les 2 gènes sont localisés dans 2 chromosomes différents, en vertu de la ségrégation indépendante ; mais elles présentent les probabilités $1-r$ et r si les 2 gènes sont localisés dans le même chromosome, en vertu du « crossing over » ; nous engloberons le 1er cas dans le 2e en convenant qu'il correspond à $r = 1/2$. On a alors : $P_{ij} = (1-r)P'_{ij} + r\pi_{ij}$, en appelant π_{ij} la probabilité d'union dans F d'une gamète portant A_i avec une gamète portant B_j.

Nous voyons donc que les différentes paires ne sont pas en général

stochastiquement indépendantes, puisque leur répartition dépend des répartitions aux générations précédentes, donc d'une répartition initiale qui peut être arbitraire. Mais nous allons montrer qu'il existe une « indépendance asymptotique » sous les hypothèses suivantes :

a) la population considérée est très nombreuse, ce qui fait qu'à chaque génération fréquences et probabilités sont sensiblement égales ;

b) elle est isogamique, ce qui revient comme nous avons vu à dire qu'aucun gène n'est avantagé, donc qu'à chaque génération les probabilités des gènes restent égales à leurs fréquences dans la génération précédente. Dès lors les fréquences p_i resteront constantes au cours des générations. Ce seront des constantes caractéristiques de la population et du système de gènes allèles considéré. On en déduit les probabilités des 3 génotypes pour les individus de coefficient de consanguinité f, c'est-à-dire leurs fréquences s'ils sont suffisamment nombreux.

c) le système de croisement adopté, quoiqu'il implique une parenté entre les 2 gamètes qui s'unissent, laisse indépendantes leurs probabilités de porter les différents gènes. Ce résultat, évident dans le cas de panmixie, n'est pas toujours valable pour les croisements consanguins ; par exemple quand la population se trouve divisée en groupes entre lesquels les croisements sont impossibles. On peut montrer par exemple qu'il s'applique au système de croisement frère-sœur, à condition que tous les individus de chaque génération soient frères et sœurs, sinon la population serait partagée en plusieurs groupes et des différences entre les gènes figurant dans ces groupes subsisteraient indéfiniment. Nous admettrons donc que le système de croisement choisi est tel qu'il laisse indépendantes les probabilités pour les 2 gamètes qui s'unissent de porter l'un le gène A_i, l'autre le gène B_j. Alors la probabilité π_{ij} d'union d'une gamète portant A_i avec une gamète portant B_j sera constante et égale à $p_i x_j$. La récurrence ci-dessus s'écrit alors :

$$P_{ij} - p_i x_j = (1 - r)(P'_{ij} - p_i x_j).$$

Si donc les P_{ij} d'une génération sont égaux à $p_i x_j$, ils le resteront toujours dans les générations suivantes : nous dirons alors que la population est *stationnaire*, et nous voyons que les gènes des différentes paires y sont stochastiquement indépendants. Dans une population non stationnaire, $P'_{ij} - p_i x_j$ tend vers 0 comme $(1 - r)^n$ quand le nombre n de générations tend vers l'infini, la population tend à devenir stationnaire ; nous supposerons dans le reste du chapitre, que *cet état d'équilibre est atteint ;* et, en particulier, qu'il y a *indépendance stochastique* des différents facteurs.

Les variables aléatoires mendéliennes dans une population isogamique stationnaire.

Portons notre attention sur un caractère déterminé, par exemple la taille, des individus constituant la population ; ce caractère pouvant être, soit quantitatif et mesurable, soit qualitatif et repérable conventionnellement sur une certaine échelle numérique. Appelons y la valeur numérique ainsi attribuée au caractère chez chaque individu. Pour un individu I pris au hasard dans la population, y est une *variable aléatoire*. Nous regarderons y comme la somme d'une variable aléatoire x représentant l'influence sur le caractère considéré de la constitution génétique de I, et d'une aléatoire e représentant l'influence du hasard et du milieu environnant sur le développement de ce caractère, e étant stochastiquement indépendante de x. Nous regarderons x comme la somme des contributions apportées au caractère par un certain nombre de paires de facteurs. Par exemple la contribution $\mathcal{H}$ de l'une de ces paires sera égale à i, j, ou k suivant que cette paire présente les états AA, Aa ou aa, états dont les probabilités sont $p^2 + fpq$, $2(1-f)pq$, $q(q+fp)$, en appelant p et q le sfréquences de A et a et f le coefficient de consanguinité de I. $\mathcal{H}$ sera appelée l'*aléatoire génotypique* relative au caractère et à la paire de facteurs considérés (1). Dans le cas où il y a dominance complète, $j = i$ ou k. Dans le cas où il n'y a pas de dominance, c'est-à-dire où l'hétérozygote est exactement intermédiaire entre les 2 homozygotes, on a $j = (i+k)/2$, ou encore $i = 2t$, $j = s + t$, $k = 2s$, et on vérifie immédiatement que l'aléatoire du 3e ordre $\mathcal{H}$ est alors la somme de 2 aléatoires du 2e ordre H et H' prenant chacune les valeurs t ou s avec les probabilités p ou q, et ayant la probabilité f d'être identiques et la probabilité $1-f$ d'être indépendantes. H et H', qui traduisent les états respectifs des 2 loci de la paire seront appelées les *aléatoires géniques*. Dans le cas où il y a dominance (complète ou non), nous pouvons conserver les aléatoires H et H' en prenant pour s et t des valeurs convenables, et poser $\mathcal{H} = \mathrm{H} + \mathrm{H}' + d$, le *résidu de dominance* d étant égal à $i - 2t$, $j - s - t$, $k - 2s$, suivant que H + H' est égal à $2t$, $s + t$, $2s$ (nous verrons plus tard comment il est le plus commode de choisir s et t). On a alors :

$$y = x + e = \mathrm{S}\mathcal{H} + e = \mathrm{S}(\mathrm{H} + \mathrm{H}' + d) + e$$

(1) Il sera toujours sous-entendu sauf avis contraire que l'épreuve dont dépend cette aléatoire est l'observation d'un individu pris au hasard parmi ceux de coefficient de consanguinité f.

le symbole S désignera une somme étendue à tous les couples de facteurs influant sur le caractère considéré (pour un caractère monofactoriel, elle ne comprend qu'un terme). La population étudiée étant supposée stationnaire, les différents termes de S sont, comme nous avons vu, des aléatoires indépendantes. e en est également supposée indépendante.

Pour simplifier les écritures, nous supposerons dorénavant que chacune de ces aléatoires est rapportée à une origine qui est sa valeur moyenne dans la population (ou dans un sous-groupe déterminé de cette population) ; cette hypothèse ne restreint pas la généralité, il suffit de convenir que le caractère étudié est mesuré en prenant *pour valeur o cette valeur moyenne*, qui est sensiblement la moyenne générale dans la population (ou le sous-groupe) s'ils sont nombreux. En désignant comme d'usage par le symbole $\mathfrak{M}$ la valeur moyenne d'une aléatoire, la convention que nous venons de faire se traduira alors par :

$$\mathfrak{M}(\mathcal{H}) = 0, \quad \mathfrak{M}(x) = 0, \quad \mathfrak{M}(e) = 0, \quad \mathfrak{M}(y) = 0$$

(et en choisissant convenablement s et t : $\mathfrak{M}(\mathrm{H}) = 0$, $\mathfrak{M}(d) = 0$).

La fluctuation du caractère y (c'est-à-dire le carré de son écart-type) dans la population (ou le sous-groupe) sera alors, en raison de l'indépendance :

$$\mathfrak{M}(y^2) = \mathfrak{M}(x^2) + \mathfrak{M}(e^2) = \mathrm{S}\mathfrak{M}(\mathcal{H}^2) + \mathfrak{M}(e^2)$$

ce que nous écrirons :

$$\sigma_y^2 = \sigma_x^2 + \sigma_e^2$$

(σ désignant l'écart-type).

Toutes ces formules subsistent dans le cas de multiallélisme.

Nous voyons que le fait pour une population d'être stationnaire entraîne la constance au cours des générations de la fluctuation, c'est-à-dire de la variabilité. Or la conservation de la variabilité est un fait d'expérience qui peut être regardé comme confirmant l'hypothèse de l'hérédité mendélienne des caractères. Les théories d'« hérédité mélangée » vers lesquelles ont penché certains biométriciens impliqueraient que la partie héréditaire x d'un caractère, si elle était égale à x_1 et x_2 chez les 2 parents, serait égale à $(x_1 + x_2)/2$ chez les enfants (le reste de la variabilité étant attribué au hasard et au milieu). Mais alors, dans le cas de la panmixie, et en supposant $\mathfrak{M}(x) = 0$, la fluctuation de x dans l'ensemble des enfants d'une population serait ($\mathfrak{M}(x_1 x_2)$ étant nulle) :

$$\mathfrak{M}\left(\frac{x_1 + x_2}{2}\right)^2 = \frac{1}{2}\,\mathfrak{M}(x^2),$$

moitié de la fluctuation des parents ; donc la fluctuation génétique de x tendrait rapidement vers 0 au cours des générations, et finalement, la seule variabilité serait ou bien celle produite par le milieu ou le hasard, mais les expériences de Johannsen sur les lignées pures ont montré que pour la plupart des caractères elle est faible ; ou bien celle produite par des mutations qui devraient alors être très fréquentes, ce qui est encore en contradiction avec l'expérience. « L'hérédité mélangée » est donc inadmissible, et le schéma mendélien, avec sa disjonction indéfinie des caractères parentaux, est un des plus simples parmi ceux qui entraînent la conservation de la variabilité héréditaire (Fisher (4)).

Nous allons maintenant montrer comment le schéma mendélien permet de retrouver les résultats de la biométrie : nous supposerons dorénavant que le caractère étudié est *multifactoriel* et dépend d'un *grand nombre de gènes* apportant des contributions toutes du même ordre de grandeur. x est alors la somme d'un grand nombre d'aléatoires indépendantes dont chacune est petite par rapport à l'écart-type σ_x de x, et d'après le théorème de Liapounoff la loi de probabilité de x est voisine de la loi de Gauss $\frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x^2}{2\sigma_x^2}} dx$. Si l'effet du hasard et du milieu sur le développement résulte d'actions multiples et indépendantes, e sera sensiblement gaussienne et y aussi, ce qui est conforme aux observations de Galton et Pearson sur la taille.

Mesurons le caractère y sur 2 individus apparentés I_1 et I_2, et soient y_1 et y_2 ses valeurs. On peut montrer de même que la loi de probabilité de l'ensemble des 2 aléatoires y_1 et y_2 est voisine d'une loi de Gauss-Bravais, elle peut donc être caractérisée par son coefficient de corrélation. La détermination expérimentale de ce coefficient, en relevant dans une population nombreuse un grand nombre de paires d'individus de même parenté, a été faite, pour différentes populations, par Galton, Pearson, Snow (5) (15). Nous allons en chercher la valeur théorique.

I. — LES CORRÉLATIONS ENTRE APPARENTÉS, DANS LE CAS DE NON-DOMINANCE

On a alors :

$$y = S\mathscr{E} + e = S(H + H') + e$$

et :

$$\mathfrak{M}(H + H') = 2\mathfrak{M}(H) = 2(pt + qs)$$

donc la convention $\mathfrak{M}(\mathscr{E}) = 0$ équivaut à $\mathfrak{M}(H) = 0$ c'est-à-dire :

$$\underline{pt + qs = 0.}$$

Soient y_1 et y_2 les aléatoires représentant les caractères de 2 individus I_1 et I_2 de coefficient de parenté f.

S'il n'y a pas dominance :

$$y_1 = S(H_1 + H'_1) + e_1 = H_1 + H'_1 + K_1 + K'_1 + \ldots + e_1$$
$$y_2 = S(H_2 + H'_2) + e_2 = \ldots$$

Pour calculer leur coefficient de corrélation r, il faut calculer la valeur moyenne de leur produit, qui se réduit à :

$$\mathfrak{M}(y_1 y_2) = S\mathfrak{M}(H_1H_2 + H'_1H_2 + H_1H'_2 + H'_1H'_2)$$

car en vertu de l'indépendance :

$$\mathfrak{M}(e_1H_2) = \mathfrak{M}(e_1)\mathfrak{M}(H_2) = 0, \text{ etc.},$$

de même $\mathfrak{M}(e_1e_2) = 0$; et, si K et K′ représentent les aléatoires géniques relatives à une autre paire quelconque :

$$\mathfrak{M}(K_1H_2) = \mathfrak{M}(K_1)\mathfrak{M}(H_2) = 0, \text{ etc.}$$

En outre chaque terme tel que $\mathfrak{M}(H_1H_2)$ se calcule en remarquant que les aléatoires H_1 et H_2 traduisent l'état de 2 loci homologues pris au hasard sur I_1 et I_2, c'est-à-dire ont une probabilité f d'être identiques et $1 - f$ d'être indépendantes. D'où :

$$\mathfrak{M}(H_1H_2) = f\mathfrak{M}(H_1^2)$$

(f est donc le coefficient de corrélation de H_1 et H_2).

Donc :

$$\mathfrak{M}(y_1y_2) = 4fS\mathfrak{M}(H_1^2).$$

D'autre part, si f_1 et f_2 sont les coefficients de consanguinité de I_1 et I_2, on a :

$$\mathfrak{M}(y_1^2) = S\mathfrak{M}(H_1 + H'_1)^2 + \mathfrak{M}(e^2) = 2(1 + f_1)S\mathfrak{M}(H_1^2) + \mathfrak{M}(e^2)$$

puisque :

$$\mathfrak{M}(H_1H'_1) = f_1\mathfrak{M}(H_1^2).$$

D'où le coefficient de corrélation cherché :

$$r = \mathfrak{M}(y_1y_2)/\sqrt{\mathfrak{M}(y_1^2)\mathfrak{M}(y_2^2)} = 2f/\sqrt{(1 + f_1 + \xi^2)(1 + f_2 + \xi^2)}$$

en appelant ξ^2 le rapport $\mathfrak{M}(e^2)/2S\mathfrak{M}(H_1^2)$.

Pour un caractère ne dépendant que de l'hérédité et pour des individus sans consanguinité r se réduit à $r_0 = 2f$: que nous appellerons

la *corrélation fondamentale*, ce qui donne les coefficients bien connus : 1/2 pour parent-enfant et pour 2 frères ; 1/4 pour demi-frères, ou grand-père et petit-fils, ou pour oncle-neveu, ou pour doubles cousins germains ; 1/8 pour cousins germains, etc.

Mais la consanguinité chez chacun des individus que l'on compare, l'action sur eux du hasard ou du milieu, se traduit par des f_1, f_2, $\xi^3 \neq 0$, elle a donc pour effet de réduire la corrélation fondamentale.

II. — LES CORRÉLATIONS DANS LE CAS DE DOMINANCE, POUR DES INDIVIDUS AYANT UN COEFFICIENT DE CONS NGUINITÉ NUL

Les probabilités des gènes A et a étant toujours p et q, celles des 3 génotypes chez ces individus seront p^2, $2pq$, q^2. Supposons toujours les aléatoires $\mathcal{A}$ rapportées à des origines telles que :

$$\mathfrak{M}(\mathcal{A}^2) = p^2 i + 2pqj + q^2 k = 0.$$

d prend avec les probabilités p^2, $2pq$, q^2, les valeurs $i - 2t$, $j - s - t$, $k - 2s$; t et s désignant les valeurs que peut prendre chacune des aléatoires H ou H′ (valeurs jusqu'ici arbitraires). Avec R. A. Fisher (3) nous choisirons les valeurs qui rendent minimum $\mathfrak{M}(d^2)$, d'où :

$$(D) \qquad \begin{cases} p(i - 2t) + q(j - s - t) = 0 \\ p(j - s - t) + q(k - 2s) = 0 \end{cases}$$

(en annulant les dérivées partielles par rapport à t et s).

On obtient ainsi pour t et s des valeurs bien déterminées :

$$\begin{cases} t = pi + qj \\ s = pj + qk \end{cases}$$

qui satisfont à $\mathfrak{M}(H) = 0$ (car $pt + qs = 0$). On a donc :

$$\mathfrak{M}(d) = \mathfrak{M}(\mathcal{A}) - \mathfrak{M}(H) - \mathfrak{M}(H') = 0.$$

Mais de plus nous remarquons que les équations (D) signifient que *la valeur moyenne de* d *est nulle quand on fixe la valeur de* H (ou celle de H′) (si on fixe $H = t$, H′ qui est indépendant de H puisque l'individu est sans consanguinité, présentera les valeurs de t ou s avec les probabilités p ou q, et d aura les valeurs correspondantes $i - 2t$ et $j - s - t$, dont la valeur moyenne est bien nulle d'après (D)).

Il en résulte que $\mathfrak{M}(d\mathrm{H}) = \mathfrak{M}(d\mathrm{H}') = 0$. Donc :

$$\mathfrak{M}(\mathfrak{X}^2) = \mathfrak{M}(\mathrm{H} + \mathrm{H}')^2 + \mathfrak{M}(d^2) = 2\mathfrak{M}(\mathrm{H}^2) + \mathfrak{M}(d^2)$$
$$\mathfrak{M}(y^2) = \mathrm{S}2\mathfrak{M}(\mathrm{H}^2) + \mathrm{S}\mathfrak{M}(d^2) + \mathfrak{M}(e^2)$$

Soient pour 2 individus apparentés I_1 et I_2 les valeurs :

$$y_1 = \mathrm{S}(\mathrm{H}_1 + \mathrm{H}'_1 + d_1) + e_1$$
$$y_2 = \mathrm{S}(\mathrm{H}_2 + \mathrm{H}'_2 + d_2) + e_2.$$

Par hypothèse H_1 et H'_1 sont indépendantes, et de même H_2 et H'_2. H_1 ne peut donc être en corrélation (positive) à la fois avec H_2 et H'_2. Supposons par exemple que H_1 ne soit en corrélation qu'avec H_2 ; H'_1 ne pourra alors être en corrélation qu'avec H'_2 ; soient φ et φ' les coefficients de corrélation respectifs. On a alors :

$$\mathfrak{M}(y_1y_2) = \mathrm{S}[\mathfrak{M}(\mathrm{H}_1\mathrm{H}_2) + \mathfrak{M}(\mathrm{H}'_1\mathrm{H}'_2) + \mathfrak{M}(d_1d_2) + \mathfrak{M}(d_1\mathrm{H}_2) + \mathfrak{M}(d_1\mathrm{H}'_2) + \mathfrak{M}(d_2\mathrm{H}_1) + \mathfrak{M}(d_2\mathrm{H}'_1)]$$

car les termes tels que $\mathfrak{M}(e_1\mathrm{H}_2)$, $\mathfrak{M}(e_1d_2)$, $\mathfrak{M}(e_1e_2)$, sont nuls puisque les aléatoires qui y figurent sont indépendantes et ont des valeurs moyennes nulles.

Nous allons montrer de plus que les 4 derniers termes écrits sont nuls. Si par exemple l'on fixe la valeur de H_2, d_2 ne dépend plus que de H'_2 donc est indépendante de H_1, et la valeur moyenne du produit H_1d_2 est le produit des valeurs moyennes de H_1 et de d_2, alors que nous savons que cette dernière est nulle. Donc la valeur moyenne de H_1d_2, étant nulle quand H_2 est fixé, est nulle aussi quel que soit H_2. De même pour les 3 autres termes. On a donc :

$$\mathfrak{M}(y_1y_2) = \mathrm{S}[\mathfrak{M}(\mathrm{H}_1\mathrm{H}_2) + \mathfrak{M}(\mathrm{H}'_1\mathrm{H}'_2) + \mathrm{M}(d_1d_2)] = (\varphi + \varphi')\mathrm{S}\mathfrak{M}(\mathrm{H}^2) + \mathrm{S}\mathfrak{M}(d_1d_2)$$

et tout revient à calculer $\mathfrak{M}(d_1d_2)$.

a) ***Les 2 individus sont apparentés par un seul de leurs loci.***

Il n'y a alors que 2 des 4 aléatoires géniques qui ne soient pas indépendantes, par exemple H_1 et H_2 ; soit φ leur coefficient de corrélation, qui résulte du mode de parenté. Si on fixe H_1, d_1 ne dépend plus que de H'_1 donc devient indépendante de H_2, H'_2 et de d_2, et sa valeur moyenne est alors nulle. Donc la valeur moyenne de d_1d_2 est nulle quand on fixe H_1 et par suite quel que soit H_1 : $\mathfrak{M}(d_1d_2) = 0$. On a alors $\mathfrak{M}(y_1y_2) = \mathrm{S}\mathfrak{M}(\mathrm{H}_1\mathrm{H}_2) = \varphi\mathrm{S}\mathfrak{M}(\mathrm{H}^2)$. D'où le coefficient de corrélation de y_1 et y_2 :

$$r = \frac{\mathfrak{M}(y_1y_2)}{\mathfrak{M}(y_1^2)} = \frac{\varphi}{2(1 + \eta^2 + \xi^2)}$$

en posant :

$$\eta^2 = \frac{S\mathfrak{M}(d^2)}{2S\mathfrak{M}(H^2)}, \qquad \xi^2 = \frac{\mathfrak{M}(e^2)}{2\mathfrak{M}(H^2)}.$$

Cela peut s'écrire encore $r = (\varphi/2)\tau^2/\sigma^2$ en désignant par τ^2 la « fluctuation génétique » $2S\mathfrak{M}(H^2)$, et par σ^2 la « fluctuation totale » $2S\mathfrak{M}(H^2) + S\mathfrak{M}(d^2) + \mathfrak{M}(e^2)$. Pour éviter d'avoir à évaluer φ remarquons que, puisque φ ne dépend pas de la dominance, cette formule peut s'écrire : $r = r_0\tau^2/\sigma^2$, r_0 désignant la « corrélation fondamentale » définie précédemment. La dominance joue donc, dans le cas d'individus apparentés par un seul locus, exactement le même rôle que le hasard et le milieu, en réduisant toutes les « corrélations fondamentales » dans un même rapport fixe < 1. Cette formule donne en particulier pour les corrélations simples en ligne directe ou en ligne collatérale : $(1/2)^n\tau^2/\sigma^2$, avec $n = 1$ pour la corrélation père-fils, $n = 2$ pour grand-père-petit-fils, demi-frères, ou oncle-neveu, $n = 3$ pour cousins germains, etc... Mais elle ne s'applique pas aux frères ou aux doubles cousins germains, qui sont apparentés à la fois par les 2 loci. Nous montrerons au *b*) que dans ces cas la réduction subie par la corrélation fondamentale est moins importante, du fait qu'il existe une corrélation positive ($\mathfrak{M}(d_1d_2) > 0$) entre les résidus de dominance d_1 et d_2 des 2 individus.

Mais remarquons pour finir que, si on cherche la corrélation partielle entre caractères y chez un individu I_1 et chez un de ses ancêtres I_2 en supposant fixée la valeur du caractère chez un ancêtre intermédiaire I_3, qui en est séparé par n et p chaînons respectivement, on peut appliquer, si les régressions sont linéaires (ce qui est réalisé quand les aléatoires y sont gaussiennes et à liaison gaussienne), la formule classique :

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{(1/2)^{n+p}\tau^2/\sigma^2 - (1/2)^n(\tau^2/\sigma^2)(1/2)^p(\tau^2/\sigma^2)}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

et ce coefficient, en général positif, n'est nul que si $\tau^2/\sigma^2 = 1$, c'est-à-dire, s'il n'y a ni dominance ni influence du milieu ; ce n'est que dans ce cas que, connaissant le caractère chez un ancêtre de I_1, la connaissance chez les ancêtres antérieurs dans la même ligne ne nous apprendrait rien de plus sur I_1 (pas d'« hérédité ancestrale »). Mais il y a pratiquement toujours dominance ou influence du milieu, ce qui fait que la connaissance du caractère chez un ancêtre laisse subsister une corrélation positive entre ancêtres antérieurs et descendants. Cette « loi d'hérédité ancestrale » mise en évidence expérimentalement par Galton et Pearson n'est donc nullement en contradiction (comme l'avaient cru Bateson et Weldon) avec les lois de Mendel. Certes, il

résulte des lois de Mendel que la connaissance de la constitution génétique d'un ancêtre rend indifférente, pour la prévision de la descendance, toute connaissance des ancêtres antérieurs. Mais notre étude montre simplement que la connaissance du caractère y chez cet ancêtre n'est, quand il y a dominance ou influence du milieu, qu'un renseignement insuffisant sur sa constitution génétique, et ce renseignement peut alors être précisé par la connaissance des ancêtres antérieurs.

b) *Les individus sont apparentés par leurs 2 loci.*

Posons $\mathcal{H}_1 = H_1 + H'_1 + d_1$ et $\mathcal{H}_2 = H_2 + H'_2 + D_2$. Il s'agit de calculer $\mathfrak{M}(\mathcal{H}_1\mathcal{H}_2)$ sachant que H_1 et H_2 ont un coefficient de corrélation φ, H'_1 et H'_2 un coefficient φ', et que ces 2 couples d'aléatoires sont indépendants entre eux.

La fonction génératrice $V(x, y, z, u)$ de l'ensemble de ces 4 aléatoires liées est donc le produit des fonctions génératrices $V_1(x, y)$ et $V_2(z, u)$ des 2 couples H_1 et H_2, H'_1 et H'_2.

Rappelons que la fonction génératrice d'aléatoires prenant des valeurs respectives α, β, etc., est par définition la valeur moyenne de $x^\alpha y^\beta \ldots$ (alors que la fonction caractéristique est la valeur moyenne de $e^{\alpha x} e^{\beta y} \ldots$). Donc :

$$\begin{aligned} V_1(x, y) &= p(p + \varphi q)x^t y^t + pq(1 - \varphi)(x^t y^s + x^s y^t) + q(q + \varphi p)x^s y^s \\ &= (px^t + qx^s)(py^t + qy^s) + pq(x^t - x^s)(y^t - y^s) \end{aligned}$$

et $V_2(z, u)$ s'en déduit en y remplaçant x par z, y par u, et φ par φ'. On sait que la fonction génératrice de l'ensemble des 2 variables $H_1 + H'_1$ et $H_2 + H'_2$ s'obtient en faisant $x = z$ et $y = u$ dans le produit V_1V_2, c'est donc $W(x, y) = V_1(x, y)V_2(x, y) = \Sigma P_{\alpha\beta} x^\alpha y^\beta$, le coefficient $P_{\alpha\beta}$ de $x^\alpha y^\beta$ dans $W(x, y)$ représentant par définition la probabilité pour qu'on ait à la fois $H_1 + H'_1 = \alpha$ et $H_2 + H'_2 = \beta$, donc pour que $\mathcal{H}_1$ et $\mathcal{H}_2$ aient des valeurs déterminées $u(\alpha)$ et $u(\beta)$. La connaissance de W permet donc de calculer $\mathfrak{M}(\mathcal{H}_1\mathcal{H}_2) = \Sigma P_{\alpha\beta} u(\alpha)u(\beta)$ en remplaçant dans W, x^α par $u(\alpha)$ et y^β par $u(\beta)$ c'est-à-dire x^{2i} et y^{2i} par i, etc. ; calculons donc :

$$\begin{aligned} W(x, y) = {} & (px^t + qx^s)^2(py^t + qy^s)^2 \\ & + pq(\varphi + \varphi')(px^t + qx^s)(x^t - x^s)(py^t + qy^s)(y^t - y^s) \\ & + p^2q^2\varphi\varphi'(x^t - x^s)^2(y^t - y^s)^2 ; \end{aligned}$$

en remplaçant x^{2i} et y^{2i} par i, x^{i+s} et y^{i+s} par j, x^{2s} et y^{2s} par k, on obtient alors :

$$\mathfrak{M}(\mathfrak{H}_1\mathfrak{H}_2) = (p^2i + 2pqj + q^2k)^2 + pq(\varphi + \varphi')(pi + (q-p)j - qk)^2 + p^2q^2\varphi\varphi'(k - 2j + i)^2.$$

C'est une forme bilinéaire symétrique de φ et de φ' dont les coefficients sont bien déterminés dans une population donnée et ne dépendent pas de φ et φ'. Il en est de même de :

$$r = \frac{\mathfrak{M}(y_1y_2)}{\sigma y_1 \sigma y_2} = \frac{\mathfrak{M}(x_1x_2)}{\mathfrak{M}(y^2)} = \frac{S\mathfrak{M}(\mathfrak{H}_1\mathfrak{H}_2)}{\mathfrak{M}(y^2)} = \frac{Spq(\varphi + \varphi')[pi + (q-p)j - qk]^2 + p^2q^2\varphi\varphi'(k-2j+i)^2}{\sigma^2}.$$

Nous calculerons les coefficients en donnant à φ et φ' des valeurs particulières. Nous avons vu que pour $\varphi' = 0$, r doit se réduire à $\frac{\varphi}{2}\tau^2/\sigma^2$. On peut donc écrire :

$$\underline{r = (\varphi + \varphi')\tau^2/2\sigma^2 + \varphi\varphi'\varepsilon^2/\sigma^2}$$

en posant toujours :

$$\sigma^2 = \mathfrak{M}(y^2) = S[2\mathfrak{M}(H^2) + \mathfrak{M}(d^2)] + \mathfrak{M}(e^2) \qquad \text{(fluctuation totale),}$$

et :

$$\tau^2 = S\mathfrak{M}(H + H')^2 = 2S\mathfrak{M}(H^2) \qquad \text{(fluctuation génétique)}$$

et en introduisant une quantité ε^2 qui est égale à $S\mathfrak{M}(d^2)$ (fluctuation de dominance), car si on fait $\varphi = \varphi' = 1$, $\mathfrak{H}_1$ et $\mathfrak{H}_2$ deviennent alors identiques, on doit donc avoir :

$$\frac{\tau^2 + \varepsilon^2}{\sigma^2} = \frac{S\mathfrak{M}(\mathfrak{H}^2)}{\mathfrak{M}(y^2)} = \frac{S[2\mathfrak{M}(H^2) + \mathfrak{M}(d^2)]}{\sigma^2}$$

[ce calcul fournit incidemment :

$$\tau^2 = S2pq[pi + (q-p)j - qk]^2 \quad \text{et} \quad \varepsilon^2 = Sp^2q^2(k - 2j + i)^2].$$

En comparant à la formule $\mathfrak{M}(y_1y_2) = (\varphi + \varphi')S\mathfrak{M}(H^2) + S\mathfrak{M}(d_1d_2)$, on voit que $\mathfrak{M}(d_1d_2) = \varphi\varphi'\varepsilon^2 = \varphi\varphi'\mathfrak{M}(d^2)$; le coefficient de corrélation entre les résidus de dominance d_1 et d_2 de I_1 et I_2 est donc $\varphi\varphi'$, produit des coefficients de corrélation entre les aléatoires géniques. Il est nul si I_1 et I_2 sont apparentés par un seul locus, mais positif si I_1 et I_2 sont apparentés par leurs 2 loci, ce qui a alors pour effet d'augmenter la corrélation. Par exemple pour des frères :

$$\varphi = 1/2 \qquad \varphi' = 1/2 \qquad r = (1/2)\frac{\tau^2 + \varepsilon^2/2}{\sigma^2}$$

(cette corrélation est donc supérieure à la corrélation père-fils quand il y a dominance ; une autre cause de supériorité apparaîtrait si l'on tenait compte de ce que les actions du milieu sur 2 frères ne peuvent plus être regardées comme indépendantes s'ils sont élevés ensemble). Pour des doubles cousins :

$$\varphi = \varphi' = 1/4 \qquad r = (1/4)\frac{\tau^2 + \varepsilon^2/4}{\sigma^2}$$

(cette corrélation est supérieure à la corrélation oncle-neveu).

Le phénomène de dominance se traduit donc statistiquement par une supériorité des coefficients de corrélation pour les parentés doubles sur ceux des parentés simples correspondantes. Mais cette supériorité s'atténue rapidement à mesure que la parenté devient plus distante, car le produit $\varphi\varphi'$ devient alors rapidement négligeable.

Extensions diverses. — 1° Les résultats subsistent s'il y a multiallélisme, car $V_1(x, y)$ est encore une forme linéaire de φ, donc $W(x, y)$, $\mathfrak{M}(\mathcal{H}_1\mathcal{H}_2)$, et ρ, sont des formes bilinéaires symétriques de φ et φ' dont les coefficients se déterminent encore en faisant $\varphi' = 0$, puis $\varphi = \varphi' = 1$.

2° Les résultats peuvent être étendus (3) (11) au cas où les effets sur le caractère considéré des différents couples de gènes ne sont pas additifs (généralisation de la dominance).

3° Le calcul peut être modifié (3) (11) (17) pour tenir compte de la ressemblance entre parents (homogamie) qui a pour effet de majorer toutes les corrélations.

4° Si dans la mesure statistique de la corrélation on sépare les sexes, on trouve en général des résultats différents suivant le sexe, en raison de la contribution au caractère considéré de gènes « liés au sexe ». Les calculs peuvent s'étendre à ce cas (Hogben (8)).

III. — LES CORRÉLATIONS DANS LE CAS DE DOMINANCE, POUR DES INDIVIDUS QUELCONQUES

Pour 2 individus I_1 et I_2 de coefficients de consanguinité non nuls, le calcul de la corrélation (dans le cas de dominance), est beaucoup moins simple, car les 4 aléatoires H_1, H_2, H'_1, H'_2, sont alors liées entre elles. Il est alors indispensable de déterminer la fonction génératrice de l'ensemble de ces 4 aléatoires, ce qui ne peut se faire, pour un schéma de parenté quelconque, que de proche en proche, par la méthode suivante : étant donné un groupe d'individus I_1, I_2, etc., I_n, désignons par $P_{\alpha\alpha'\beta\beta'\ldots}$ la probabilité conjointe pour que leurs $2n$ loci homologues soient dans les états représentés par α, α', β, β', ... (cha-

cune de ces quantités ayant l'une des valeurs t ou s). La fonction génératrice relative aux $2n$ loci sera donc :

$$\varphi(a_1, a_2, b_1, b_2, \ldots) = \sum P_{\alpha\alpha'\beta\beta'\ldots} a_1^{\alpha} a_2^{\alpha'} b_1^{\beta} b_2^{\beta'} \ldots$$

elle jouit des propriétés suivantes :

Si on réunit 2 groupes d'individus sans corrélation, les fonctions φ se multiplient.

Si on fait abstraction d'un des individus, par exemple I_1, la fonction génératrice relative aux individus restants se déduit de φ en y faisant :

$$a_1 = a_2 = 1.$$

Si on adjoint au groupe un enfant E d'un couple pris dans le groupe, par exemple un enfant de I_1 et I_2, la fonction génératrice du groupe ainsi étendu, comprendra 2 variables de plus, relatives à E, soient l_1 et l_2 : elle ne sera autre, d'après la loi de Mendel, que :

$$\sum P_{\alpha\alpha'\beta\beta'\ldots} \frac{l_1^{\alpha} + l_1^{\alpha'}}{2} \frac{l_2^{\beta} + l_2^{\beta'}}{2} a_1^{\alpha} a_2^{\alpha'} \ldots$$

ce qui est égal à :

$$\frac{1}{4}[\varphi(a_1 l_1, a_2, b_1 l_2, b_2) + \varphi(a_1 l_1, a_2, b_1, b_2 l_2) + \varphi(a_1, a_2 l_1, b_1 l_2, b_2) + \varphi(a_1, a_2 l_1, b_1, b_2 l_2)]$$

On peut ainsi passer de proche en proche des probabilités pour un groupe initial donné aux probabilités pour tout groupe en dérivant par des croisements donnés. Mais les calculs sont rarement simples.

CHAPITRE III

L'ÉVOLUTION D'UNE POPULATION MENDÉLIENNE

Nous avons étudié jusqu'ici une population mendélienne stationnaire, c'est-à-dire une population où la répartition des gènes ne change pas d'une génération à l'autre ; il ne pouvait en être ainsi que si la population était très nombreuse, et si les différents gènes allèles ne conféraient à leurs porteurs ni avantage ni désavantage (gènes *neutres*). Nous allons nous affranchir de ces hypothèses, en considérant successivement une population *limitée*, puis une population où il y a *sélection* des gènes. Nous verrons qu'alors la répartition des gènes ne reste plus constante, mais qu'elle évolue au cours du temps, et nous aurons à répondre à deux questions : vers quoi tend cette évolution ? et à quelle vitesse se poursuit-elle ?

I. — INFLUENCE DE L'EFFECTIF DE LA POPULATION SUR DES GÈNES NEUTRES

a) *Effectif constant*

Considérons une population *à effectif constant* de K individus, se reproduisant par panmixie, et envisageons d'abord des gènes dont le taux de mutation puisse être négligé. Partons d'une génération initiale F_0 et désignons les générations successives, que nous supposerons séparées, par F_1, F_2, ... (le cas où les générations ne seraient pas séparées et où des croisements entre générations différentes seraient possibles introduit des complications de calcul, mais ne modifie pas essentiellement les résultats). Malgré la panmixie les individus de la $n^{ième}$ génération F_n présenteront certainement de la consanguinité quand n sera suffisamment grand, car chacun aura au plus K ancêtres

distincts d'ordre n, alors qu'il en a en tout 2^n. Nous ne pourrions calculer le coefficient de consanguinité de l'un d'eux que si nous connaissions toutes les chaînes de parenté reliant ses parents, c'est-à-dire si nous connaissions le schéma complet des croisements depuis le début. Mais nous allons voir qu'il est facile de caractériser *a priori* la consanguinité moyenne à la $n^{ième}$ génération par un nombre f_n. f_n sera par définition la probabilité *évaluée a priori*, pour que les 2 loci d'un individu pris au hasard dans F_n soient identiques, c'est-à-dire proviennent d'un même locus d'un ancêtre commun. Dans chaque expérience réalisée, le coefficient de consanguinité *a posteriori* variera suivant l'individu considéré, mais f_n représentera sa valeur moyenne pour un grand nombre d'individus, c'est-à-dire dans un grand nombre d'expériences.

Puisque les gènes considérés sont supposés neutres, les probabilités *a priori* des différents allèles seront les mêmes pour toutes les générations, soient p et q si on suppose 2 allèles seulement. Les formules :

$$p(p+f_nq) \qquad 2pq(1-f_n) \qquad q(q+f_np)$$

représenteront donc les probabilités *a priori* des 3 génotypes pour la $n^{ième}$ génération et aussi *leurs fréquences dans un grand nombre d'expériences* où l'on partirait toujours des mêmes fréquences initiales p et q pour les gènes A et a. Nous allons calculer f_n par récurrence dans différents cas en faisant abstraction des mutations.

1° *Sexes séparés*

Considérons d'abord une population animale *à sexes séparés*, comprenant des nombres constants N_1 et N_2 de mâles et de femelles, formant la sous-population des mâles ${}_1F$ et la sous-population des femelles ${}_2F$. Puisqu'il y a panmixie, les 2 loci d'un individu A_n de F_n sont tirés au hasard l'un dans ${}_1F_{n-1}$, l'autre dans ${}_2F_{n-1}$; la probabilité pour qu'ils proviennent d'un même individu de ${}_1F_{n-2}$ ou de ${}_2F_{n-2}$ est :

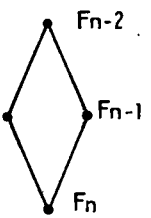

Fig. 5

$$\frac{1}{2}\frac{1}{N_1}\frac{1}{2}+\frac{1}{2}\frac{1}{N_2}\frac{1}{2}=\frac{1}{N}$$

(en appelant N la moyenne harmonique de $2N_1$ et $2N_2$:

$$1/N=1/4N_1+1/4N_2)\ ;$$

la probabilité complémentaire $1-\frac{1}{N}$ est la probabilité pour qu'ils proviennent d'individus différents de F_{n-2} ; dans ces 2 cas respectifs

les probabilités d'identité des 2 loci sont $(1+f_{n-2})/2$ et f_{n-1} (car f_{n-1} représente la probabilité pour que 2 loci pris chez 2 individus différents de F_{n-2} soient identiques). Donc f_n, probabilité d'identité des 2 loci de F_n, est donné par :

$$f_n = \frac{1}{N}\,\frac{1+f_{n-2}}{2} + \left(1-\frac{1}{N}\right) f_{n-1}.$$

De cette récurrence linéaire à coefficients constants, on déduit facilement f_n. On se ramène d'abord à une récurrence homogène en remarquant que l'équation est vérifiée par $f_n = \text{constante} = 1$ et posant $\alpha_n = 1 - f_n$ d'où :

$$\alpha_n = [1-(1/N)]\alpha_{n-1} + (1/2N)\,\alpha_{n-2}.$$

α_n sera une combinaison linéaire de 2 solutions de la forme $\alpha_n = k^n$, k étant donné par l'équation caractéristique :

$$k^2 - [1-(1/N)]k - 1/2N = 0.$$

$$\alpha_n = \lambda\left[\left(1-1/N+\sqrt{1+1/N^2}\right)/2\right]^n + \mu\left[\left(1-1/N-\sqrt{1+1/N^2}\right)/2\right]^n.$$

λ et μ étant déterminés par la donnée des 2 valeurs initiales α_0 et α_1 :

$$\alpha_0 = \lambda + \mu \qquad \alpha_1 = \alpha_0(1-1/N)/2 + (\lambda-\mu)\sqrt{1+1/N^2}/2.$$

Remarquons en passant que le croisement indéfini frère-sœur, étudié en détail par Haldane et Fisher, rentre comme cas particulier dans cette formule, en faisant $N = 2$:

$$\alpha_n = \lambda\left[\left(1+\sqrt{5}\right)/4\right]^n + \mu\left[\left(1-\sqrt{5}\right)/4\right]^n.$$

Si N est grand on a :

$$\sqrt{1+1/N^2}\,\lambda = \left(\sqrt{1+1/N^2}-1+1/N\right)\alpha_0/2 + \alpha_1$$
$$\lambda = \alpha_1 + \alpha_0/2N + 0(1/N^2).$$

λ a donc pour partie principale α_1 si $\alpha_1 \neq 0$, c'est-à-dire si la population initiale n'est pas formée d'homozygotes identiques. Le terme en μ est très rapidement négligeable par rapport au terme en λ, car leur rapport est équivalent $(-1/2N)^n\mu/\lambda$, on a donc dès que n est un peu grand :

$$\alpha_n = 1 - f_n \# \alpha_1(1-1/2N)^n \# \alpha_1 e^{-n/2N}.$$

Nous en déduisons les importantes conclusions suivantes :

f_n tend vers un quand n tend vers l'infini, c'est-à-dire que l'on tend asymptotiquement vers une population dans laquelle les 2 loci de

chaque individu auraient la probabilité 1 d'être identiques, donc une population dont tous les loci seraient identiques, une population d'homozygotes identiques : pour des gènes neutres et sans mutations, *une panmixie indéfinie dans une population limitée aboutit toujours à l'homogénéité complète.* Ce résultat, en apparence étonnant, provient de ce qu'un gène peut toujours s'éteindre, si le choix au hasard des 2N loci de la génération suivante se trouve porter toujours sur son allèle, alors qu'inversement un gène éteint ne reparaît pas. Les probabilités *a priori* des gènes A et *a* dans la génération F_n sont certes toujours constantes et égales à p et q, mais cela signifie maintenant que la population finale a la probabilité p de ne contenir que des AA et la probabilité q de ne contenir que des *aa*. La différence paraît essentielle avec le cas d'une population illimitée où les 3 génotypes coexistent indéfiniment dans les proportions p^2, $2pq$, q^2. Mais il faut remarquer que l'homogénéité asymptotique est atteinte extrêmement lentement si N est grand : pour que $\alpha_n = 1 - f_n$ soit réduit au 1/10 de sa valeur, il faut un nombre n de générations tel que $e^{-n/2N} = 0{,}1$, donc $n = 2NL10$; une réduction appréciable de la différence $1 - f$ qui mesure l'écart à l'homogénéité nécessite *un nombre de générations de l'ordre de grandeur du nombre d'individus dans la population.* Ces résultats ont d'importantes conséquences biologiques : plusieurs biologistes ont insisté sur le rôle du hasard dans l'extinction des gènes neutres. On constate en effet dans de nombreuses espèces animales ou végétales la divergence de « races géographiques » qui, après avoir été séparées par une barrière telle que bras de mer ou chaîne de montagnes, évoluent vers des états homozygotes différents, l'une finissant par exemple par ne présenter que des génotypes AA alors qu'une autre ne présenterait que des *aa*. Cela pourrait certes s'expliquer par une sélection différente suivant la situation géographique, le gène A étant avantageux à un endroit et le gène *a* en un autre. Mais comme il en est fréquemment ainsi pour des gènes neutres, sans valeur sélective, il faut admettre alors que cette évolution résulte de l'homogénéisation d'une population limitée : homogénéisation qui provient de l'extinction au hasard, et qui par suite élimine tantôt l'un des gènes tantôt l'autre. Il a été parfois fait un usage abusif de cette explication, car il faut bien souligner que *cette extinction au hasard ne peut avoir lieu en un temps assez court que si la population est très peu nombreuse.* Nous prendrons pour exemple la formule sanguine des Amérindiens (indiens d'Amérique). Tous ces indiens paraissent dériver d'une même souche, malgré leur variabilité morphologique et malgré des vestiges de métissages locaux avec des immigrants d'Océanie ou de Mélanésie (race préhistorique de Lagoa-Santa, races actuelles de l'Amazonie). Or il est très curieux de constater qu'ils sont

la seule race au monde à présenter exclusivement un seul groupe sanguin, le groupe O (OO).

Pourtant les groupes A et B résultent de mutations extrêmement anciennes, puisque probablement antérieures à la séparation des lignées du chimpanzé et de l'homme, et ils doivent avoir existé de tout temps dans l'Asie extrême-orientale, qui est le foyer probable des grandes migrations humaines. Les différents groupes sanguins paraissent être sans valeur sélective, car ils coexistent, en Asie et en Europe, sous tous les climats et sous toutes les latitudes. Il semble donc que les Indiens dérivent d'un groupe d'immigrants asiatiques dans lesquels les gènes A et B auraient disparu par extinction au hasard. Mais le groupe a dû se développer rapidement après son arrivée dans le Nouveau Continent, il n'a pu rester très peu nombreux que pendant quelques générations, après quoi les proportions des gènes n'auraient pu varier que très lentement. Or, pour que l'état homozygote ait pu être atteint en un très petit nombre de générations, il faudrait que le groupe n'ait compris qu'un très petit nombre d'individus. L'hypothèse de l'extinction au hasard des gènes A et B en Amérique conduit donc à considérer que la plus grande partie des Amérindiens descendraient génétiquement d'un petit nombre d'Asiatiques (Mongoloïdes) venus en Amérique (peut-être par le détroit de Behring). Cela confirme la thèse des ethnographes américains, mais non celle d'après laquelle la race amérindienne résulterait de métissages entre Mongoloïdes, Australiens et Mélanésiens venus à plusieurs reprises par mer. Cette immigration depuis la Mélanésie n'a eu d'influence manifeste que sur des régions très isolées, comme le domaine des Sirionos (forêt vierge de l'Amazonie). Notons de plus qu'une autre série de groupes sanguins, la série MM-MN-NN montre dans toutes les races une proportion appréciable de MN et de NN, sauf chez les Indiens où il n'y en a presque point, ce qui corrobore notre hypothèse. Certes il y a beaucoup d'autres gènes pour lesquels la race indienne présente des hétérozygotes, mais ces gènes peuvent résulter pour une bonne part de mutations postérieures à l'occupation de l'Amérique. La lenteur de l'extinction au hasard des gènes dans une population qui compterait au moins quelques centaines d'individus est confirmée par l'exemple des Tziganes, nomades venus de l'Inde en Europe il y a plus de mille ans, et qui ont remarquablement conservé le type Hindou, car ils se marient presque exclusivement entre eux. La population isolée de quelques milliers d'individus qu'ils forment en Allemagne, de même qu'en France, a conservé la même répartition en groupes sanguins que les Hindous, malgré la séparation millénaire : 40 0/0 de B, la plus forte densité au monde.

2° *Sexes réunis*

Considérons maintenant une population végétale de N individus *monoïques*, c'est-à-dire où les 2 sexes sont réunis sur une même plante ; l'autofécondation est maintenant possible, mais supposons qu'elle ne soit ni plus ni moins probable que la fécondation croisée. Les 2 loci d'un individu A_n de F_n ont alors la probabilité 1/N de provenir du même individu de F_{n-1} (auquel cas ils sont identiques avec la probabilité $(1+f_{n-1})/2$) et la probabilité $1-1/N$ de provenir d'individus différents, auquel cas ils sont identiques avec une probabilité que nous désignons par φ_n ; d'où :

$$f_n = (1+f_{n-1})/2N + (1-1/N)\varphi_n.$$

φ_n est la probabilité pour que 2 loci pris sur 2 individus *différents* de F_{n-1} soient identiques ; ces 2 loci ont encore la probabilité 1/N de provenir du même individu de F_{n-2}, d'où :

$$\varphi_n = (1+f_{n-2})/2N + (1-1/N)\varphi_{n-1}$$

ce qui s'écrit :

$$f_n - (1+f_{n-1})/2N = (1-1/N)(1+f_{n-2})/2N + (1-1/N)[f_{n-1} - (1+f_{n-2})/2N]$$

$$f_n = 1/2N + (1-1/2N)f_{n-1},$$

d'où :

$$\alpha_n = 1 - f_n = (1-1/2N)\alpha_{n-1} \qquad \alpha_n = \alpha_0(1-1/2N)^n.$$

α_n tend encore vers 0, f_n tend vers 1. Le cas classique de l'autofécondation indéfinie est obtenu pour $N=1$, $1-f_n$ décroît alors de moitié à chaque génération, l'homogénéité presque complète est atteinte assez rapidement. L'autofécondation répétée d'une espèce végétale est un procédé assez rapide pour obtenir une lignée homozygote pour presque tous les facteurs. Mais, si N est grand, l'homogénéité ne s'établit que très lentement. α_n est alors de l'ordre de $e^{-n/2N}$ comme dans le cas des sexes séparés. La réunion des 2 sexes sur la même plante ne modifie que de façon insignifiante l'évolution de la population (sous l'hypothèse naturellement que l'autofécondation n'est pas plus favorisée que la fécondation croisée, car nous venons de voir qu'une autofécondation exclusive conduit au contraire rapidement à l'homogénéité).

B) *Effectif non constant*

Supposons maintenant que l'*effectif* de la population *ne soit plus constant*, mais varie au cours du temps. Considérons le cas où les sexes sont séparés. Les nombres N_1 et N_2 sont fonctions du rang i de la génération F_i. Posons $1/4N_1 + 1/4N_2 = 1/N(i)$.

La formule :

$$f_n = (1 + f_{n-2})/2N + (1 - 1/N) f_{n-1}$$

subsiste à condition de prendre pour N la valeur $N(n-2)$. Nous en déduisons, en augmentant de 2 les indices pour simplifier les écritures :

$$\alpha_{n+2} = [1 - 1/N(n)]\alpha_{n+1} + \alpha_n/2N(n).$$

Nous avons cette fois une récurrence linéaire homogène à coefficients variables. Nous la résoudrons en posant $\alpha_n = k_0 k_1 \dots k_n$, les k_i étant des constantes à déterminer, liées par la relation :

$$k_{n+2} k_{n+1} = [1 - 1/N(n)] k_{n+1} + 1/2N(n)$$

c'est-à-dire :

$$k_{n+2} = 1 - 1/N(n) + 1/[2N(n) k_{n+1}]$$

ce qui permet de calculer de proche en proche les k_i en partant de k_1 supposé > 0. Les k_{n+2} sont alors tous > 0 et même $> 1 - 1/N(n)$ (à partir de k_2). Ils sont tous < 1 à partir de k_3 car $k_{n+2} < 1$ équivaut à $k_{n+1} > 1/2$. Il en résulte que les α_n sont positifs et décroissants, donc tendent quand n tend vers l'infini vers une limite $\alpha \geqslant 0$. f *tend donc toujours par valeurs croissantes vers une limite* $1 - \alpha \leqslant 1$. Log α est la somme de la série $\log k_0 + \log k_1 + \dots \log k_n + \dots$

Pour que la limite de f soit < 1, c'est à-dire que les hétérozygotes ne soient jamais éliminés complètement, il faut et il suffit que cette série soit convergente. Il faut pour cela que $k_n \to 1$, ce qui nécessite d'après la formule de récurrence que $N(n)$ soit infini avec n ; cette dernière condition est suffisante pour que $k_n \to 1$, car, en posant :

$$k_{n+2} = 1 - u_{n+2} \quad [0 < u_{n+2} < 1/N(n)],$$

la récurrence s'écrit :

$$u_{n+2} = [1/2N(n)][1 - 2u_{n+1})/(1 - u_{n+1})] < 1/2N(n)$$

donc :

$$u_{n+2} \to 0 \qquad \text{si} \qquad N(n) \to \infty .$$

Donc : 1° *si* $N(n)$ *reste fini quand* $n \to \infty$, la limite de k_n est < 1, celle de $\log k_n$ est < 0 donc $\log \alpha = -\infty$, $\alpha = 0$, f *tend vers 1* ;

2° *si* $N(n)$ *tend vers l'infini avec n* :

$$k_n = 1 - u_n \to 1, \qquad \log k_n = \log (1 - u_n) \to 0.$$

Pour étudier la série de terme général $\log k_n$, remarquons que :

$$u_{n+2} = (1 - \varepsilon_n)/2N(n)$$

avec :

$$\varepsilon_n = u_{n+1}/(1 + u_{n+1}) \to 0$$

donc :

$$u_{n+2} \sim 1/2N(n)$$

et la série $\log k_{n+2} = \log (1 - u_{n+2})$ est convergente à condition que la série u_{n+2} soit convergente.

a) *Si* $N(n)$ *croît au plus comme une fonction linéaire de n*, la série est divergente, $f \to 1$.

b) *Si* $N(n)$ *croît au moins comme* n^{1+k} $(k > 0)$, la série est convergente, $f \to$ *une limite* < 1 et il n'y a pas disparition complète des hétérozygotes.

On trouverait le même résultat dans le cas où les 2 sexes sont réunis sur la même plante.

C) *Rôle de la mutation*

Mais il est manifeste que la plupart du temps l'hétérogénéité génétique d'une population, la présence de nombreux génotypes hétérozygotes provient non de ce que la population est extrêmement nombreuse, mais de ce qu'il y apparaît de temps en temps de nouveaux gènes, soit par *mutation*, soit par *immigration* d'individus provenant d'une population différente. Soit u_1 la fréquence moyenne de mutation par génération pour un locus déterminé, u_2 la proportion moyenne d'immigrants à chaque génération, ces immigrants étant supposés provenir de populations suffisamment nombreuses pour qu'on puisse admettre qu'il n'y a pas de parenté entre eux. Posons $u = u_1 + u_2$. La probabilité pour qu'un locus de A_n provienne d'un locus non muté d'un « autochtone » (c'est-à-dire non immigré) de la génération précédente est : $1 - u_1 - u_2 = 1 - u$. Par suite, dans le cas des sexes séparés :

$$f_n = (1 - u)^4[(1 + f_{n-2})/2N + (1 - 1/N)\varphi_{n-1}]$$

φ_{n-1} représentant la probabilité pour que 2 loci pris chez 2 autochtones différents de F_{n-2} soient identiques. Mais le coefficient de consangui-

nité f_{n-1} d'un individu de F_{n-1} est évidemment $(1-u)^2\varphi_{n-1}$. On en déduit :

$$f_n = (1-u)^2(1+f_{n-2})/2N + (1-u)^2(1-1/N)f_{n-1}.$$

En supposant u^2 négligeable, la valeur d'équilibre de f_n est donnée par :

$$f = (1-4u)(1+f)/2N + (1-2u)(1-1/N)f$$

$$f = \frac{1-4u}{4Nu+1} \# \frac{1}{1+4Nu}.$$

Pour voir comment f tend vers cette limite, posons $\alpha_n = f - f_n$; on a :

$$\alpha_n = (1-4u)\alpha_{n-2}/2N + (1-2u)(1-1/N)\alpha_{n-1}$$

cette équation admet 2 solutions $\alpha_n = k_n$, k étant racine de :

$$k^2 - (1-2u)(1-1/N)k - (1-4u)/2N = 0$$

donc :

$$2k = (1-2u)(1-1/N) \pm \sqrt{(1-2u)^2(1-1/N)^2 + 2(1-4u)/N}$$

la plus grande des 2 racines est donnée par :

$$2k = 1 - 2u - 1/N + (1 - 4u - 2/N + 8u/N + 2/N - 8u/N)^{1/2} + O(u^2) + O(1/N^2)$$

$$k = 1 - 2u - 1/2N + O(u^2) + O(1/N^2).$$

Donc :

$$\alpha_n = f - f_n \sim (1-2u-1/2N)^n \# e^{-2nu-n/2N} = e^{-(4Nu+1)n/2N}$$

la valeur d'équilibre est atteinte plus rapidement dans le cas où il y a mutation ou immigration, et beaucoup plus rapidement si $4Nu$ est grand. Quand les sexes sont réunis sur la même plante, on trouve encore le même résultat.

En résumé, nous voyons que le coefficient de consanguinité f_n tend toujours vers une valeur limite f. Si elle est égale à 1 ou très voisine de 1, cela signifie qu'il y a presque certitude que la population deviendra génétiquement homogène au bout d'un temps suffisant. Si elle est différente de 1, cela signifie qu'il subsistera en général des hétérozygotes dans la population finale [la probabilité *a priori* pour qu'un individu pris au hasard dans cette population soit hétérozygote étant $2pq(1-f)$]. Or, si l'effectif de la population ne peut être regardé comme augmentant indéfiniment, $f = 1/(1+4Nu)$ est notablement inférieur à 1 à condition que $4Nu$ ne soit pas petit, c'est-à-dire que la fréquence u de mutation et d'immigration par génération soit de

l'ordre de grandeur de 1/N au moins ; ou encore que le nombre total 2Nu de gènes nouveaux introduits à chaque génération par mutation ou migration soit au moins de l'ordre de l'unité. Quel que soit l'effectif de la population, il suffit que la mutation ou l'immigration affecte à chaque génération quelques individus pour qu'il y ait une notable persistance d'hétérozygotes.

II. — INFLUENCE DE LA SÉLECTION

Nous allons étudier la répartition, au cours du temps, d'un couple de facteurs présentant 2 allèles seulement, A et a :

Nous désignerons par p et q ($p = 1 - q$), les probabilités de A et a chez les adultes reproducteurs de la génération F_n. Les probabilités $p + \delta p$ et $q + \delta q$ ($\delta q = -\delta p$), dans la génération suivante F_{n+1} seront en général différentes de p et q. Nous admettrons que la variation δq résulte d'un certain nombre de causes produisant chacune une variation petite (de sorte que nous puissions négliger son carré) :

1° du fait des mutations récurrentes et réversibles, il y a à chaque génération, dans les cellules reproductrices de F_n, transformation en A d'une proportion moyenne u_1 de gènes a et transformation en a d'une proportion v_1 des gènes A, ce qui donne pour la fréquence moyenne de a dans les cellules reproductrices : $q - u_1 q + v_1(1 - q)$; la variation de q produite par la mutation est fonction linéaire de q.

2° il faut tenter de tenir compte de l'immigration car, dès que l'on considère, non pas la totalité des individus de l'espèce considérée, mais seulement une population locale, cette population n'est presque jamais complètement isolée et échange toujours des individus avec les populations voisines. Il en résulte que, si ce que nous avons appelé F_n désigne l'ensemble des individus « autochtones », c'est-à-dire nés sur les lieux mêmes, la population reproductrice différera de F_n ; elle ne sera formée que pour une fraction $1 - k$ en moyenne d'individus de F_n, la fraction restante k étant formée d'individus immigrés ; si nous supposons que ces individus proviennent d'un ensemble de populations dont la composition peut *être regardée comme constante au cours du temps* et caractérisée par une fréquence q_m pour le gène a, la fréquence moyenne de a dans la population reproductrice sera :

$$(1 - k)q + kq_m = q + k(q_m - q).$$

La variation de q provoquée par la migration est donc une fonction linéaire de q, comme dans le cas des mutations. On peut l'écrire sous la même forme : $-u_2 q + v_2(1 - q)$ en posant $kq_m = v_2$ et $k = u_2 + v_2$,

c'est-à-dire $u_2 = k(1 - q_m)$. La variation produite à la fois par les mutations et l'immigration est alors :

$$\delta_1 q = -u_1 q + v_1(1-q) - u_2 q + v_2(1-q) = -uq + v(1-q)$$

en posant :

$$u = u_1 + u_2 = u_1 + k(1 - q_m) \qquad \text{et} \qquad v = v_1 + v_2 = v_1 + kq_m.$$

Cette variation est donc une fonction linéaire de q.

Si les effets des mutations et de l'immigration se réunissent ainsi en une même formule, cela tient à l'hypothèse simplificatrice assez grossière que nous avons faite en supposant que les immigrants proviennent de populations extérieures *dont la composition reste constante au cours du temps* ; en réalité ces populations évoluent, et de plus, la plupart du temps la population que nous étudions réagit sur elle par émigration, de sorte que ce qu'il faudrait étudier, c'est l'évolution d'un ensemble de population réagissant les uns sur les autres par migration. Cette étude sera faite ultérieurement.

Pour le moment nous admettrons que les cellules reproductrices de la génération F_n modifiée par immigration présentent le gène a avec la fréquence $q_1 = q + \delta_1 q = q - uq + v(1-q)$.

3° Nous admettrons que les gamètes produites par ces cellules et concourant à la reproduction ne présentent plus le gène a avec la probabilité q_1 mais avec une probabilité différente q_2, parce que ce gène présente un avantage ou un désavantage pour les gamètes qui en sont porteuses (sélection gamétique) ; nous supposerons exactement que les probabilités des 2 gènes, au lieu d'être q_1 et $1 - q_1$, sont $q_2 = \alpha q_1$ et $1 - q_2 = \beta(1 - q_1)$, α et β ayant un rapport constant, que nous supposerons voisin de 1, que nous désignerons par $1 - s$, et qui caractérise le degré de viabilité des gamètes, l'intensité de la « sélection gamétique » (on peut s'il y a lieu supposer s positif, en nommant a le gène défavorable).

Comme on doit avoir :

$$\alpha q_1 + \beta(1 - q_1) = 1 \qquad \text{avec} \qquad \alpha/\beta = 1 - s$$

on a :

$$\beta(1-s)q_1 + \beta(1 - q_1) = 1 \qquad 1/\beta = 1 - sq_1 \qquad \alpha \neq 1 - s + sq_1 + 0(s^2).$$

4° Nous ferons encore l'hypothèse de consanguinité pure, qui entraîne que chaque gamète concourant à la reproduction a, quelle que soit sa constitution, la même probabilité de rencontrer une autre gamète ; la consanguinité éventuelle ayant toutefois pour effet d'augmenter la probabilité pour que cette autre gamète présente le même

gène que la première. Si nous appelons λ le coefficient de consanguinité moyen de la génération F_{n+1}, les gamètes qui s'unissent pour former les individus naissants, ou « zygotes » de la génération F_{n+1} présentent chacune A et a avec les probabilités p_2 et q_2, mais ont entre elles en moyenne une corrélation de coefficient λ ; les 3 zygotes AA, Aa, aa ont donc les probabilités :

$$P = p_2(p_2 + \lambda q_2) \qquad 2Q = 2p_2q_2(1 - \lambda) \qquad R = q_2(q_2 + \lambda p_2).$$

5° Mais nous admettrons que les 3 zygotes n'ont pas la même probabilité de se développer et de parvenir à l'état adulte reproducteur. Nous admettrons que les probabilités des adultes sont :

$$\lambda P \qquad 2\mu Q \qquad \gamma R,$$

les 3 quantités λ, μ, ν, ayant des rapports constants et voisins de 1 : $\nu/\lambda = 1 - \sigma$, $\mu/\lambda = 1 - h\sigma$, les 2 constantes σ et h caractérisant le degré de viabilité des zygotes, ou l'intensité de la « sélection zygotique » (l'hétérozygote aura une viabilité intermédiaire entre celles des 2 homozygotes si $0 < h < 1$; il sera supérieur aux deux si $h < 0$, inférieur aux deux si $h > 1$, dans le cas où $\sigma > 0$; et inversement si $\sigma < 0$).

On doit avoir :

$$\lambda P + 2\mu Q + \nu R = 1, \qquad \lambda[(1 - \sigma)R + 2(1 - h\sigma)Q + P] = 1$$

d'où :

$$\lambda = 1/(1 - \sigma R - 2h\sigma Q) = 1 + \sigma R + 2h\sigma Q + 0(\sigma^2).$$

La probabilité $q_2 + \delta_3 q$ de a chez les adultes reproducteurs de F_{n+1} est donc :

$$q_2 + \delta_3 q = \mu Q + \nu R = \lambda[(1 - h\sigma)Q + (1 - \sigma)R]$$
$$= (q_2 - h\sigma Q - \sigma R)/(1 - \sigma R - 2h\sigma Q)$$

et, en groupant les termes en Q et R et réduisant à 1 le dénominateur :

$$\delta_3 q = h\sigma p_2 q_2(1 - \lambda)(2q_2 - 1) - \sigma p_2 q_2(q_2 + \lambda p_2) + 0(\sigma^2)$$
$$= \sigma p_2 q_2[(2h - 1)(1 - \lambda)q_2 + h\lambda - \lambda - h] + 0(\sigma^2)$$

on peut y remplacer q_2 par q_1 et p_2 par p_1, à $0(s\sigma)$ près [puisque $q_2 = q_1 + 0(s)$]. On voit alors que la somme des variations $\delta_2 q$ et $\delta_3 q$ dues à la sélection gamétique et à la sélection zygotique est de la forme :

$$\delta_2 q + \delta_3 q = q_1(1 - q_1)(l + wq_1)$$

(en négligeant les termes du 2e ordre en s et σ), avec :

$$t = -s - \lambda\sigma - h\sigma(1-\lambda) \qquad \text{et} \qquad w = \sigma(2h-1)(1-\lambda).$$

Si s, σ et $h\sigma$ ont le même signe, t sera de même signe; t sera appelé *le coefficient de sélection totale*. L'ensemble de ces 2 sélections produit donc une variation qui est une fonction du 3e degré de q s'annulant pour $q=0$ et $q=1$. La sélection cesse en effet de jouer quand a est éteint ($q=0$) ou fixé ($q=1$) (Il y a là une différence essentielle avec la variation produite par les mutations ou l'immigration, qui était une fonction du 1er degré ne s'annulant pas aux bornes, car cette variation agissait même sur un gène éteint ou fixé).

La fonction s'abaisse au 2e degré si $w=0$, c'est-à-dire si $\sigma=0$ (sélection uniquement gamétique), ou si $h=1/2$ (hétérozygote exactement intermédiaire au point de vue viabilité), ou si $h=1$ (population uniquement formée d'homozygotes).

On peut remplacer dans la formule q_1 par q, si l'on néglige les termes du 2e ordre, en u, v, s, et σ. On obtient alors, pour variation totale $\delta q = \delta_1 q + \delta_2 q + \delta_3 q$ au cours d'une génération :

polynôme du 3e degré que nous appellerons $\delta(q)$; ses coefficients ont été regardés comme assez petits pour que leurs produits et carrés soient négligeables (1).

$$\delta q = \underbrace{-uq + v(1-q)}_{\text{mutation et immigration}} + \underbrace{q(1-q)(t+wq)}_{\text{sélection}}$$

C'est là la variation de la probabilité q de a, et sensiblement de la fréquence *dans une population très nombreuse.*

1° Cas d'une population très nombreuse.

L'écart entre la probabilité et la fréquence est négligeable. La fréquence q varie, d'une génération à l'autre, d'une quantité $\delta(q)$ (supposée petite). q est une fonction du temps (mesuré en générations) dont

(1) On pourrait naturellement formuler $\delta(q)$ sans faire ces approximations, mais l'expression obtenue alors serait peu maniable, sauf dans certains cas particuliers, comme celui étudié par G. Teissier (16) des gènes léthaux : *aa* non viables : $\sigma=1$, on ne peut pas négliger son carré, mais la formule se simplifie néanmoins du fait qu'il n'y a que 2 génotypes en présence.

la différence première est la fonction $\delta(q)$. L'intégration de cette différence première se ramène sensiblement à la quadrature :

$$\frac{dq}{dt} = \delta(q) \qquad t = \int_{q_0}^{q} \frac{dq}{\delta(q)}$$

q_0 étant la valeur initiale dans la génération $t = 0$.

Mais nous allons plutôt faire une discussion graphique en vue d'obtenir la limite de q quand $t \to +\infty$ (répartition asymptotique des

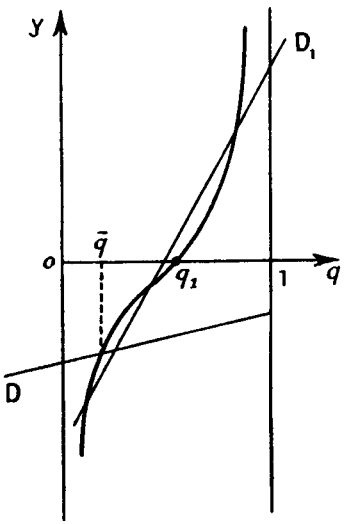

Fig. 6.

gènes). Nous supposerons que les taux de mutation ou d'immigration u et v ne sont pas nuls. Remarquons alors que le polynôme du 3e degré $\delta(q)$ est égal à $v > 0$ pour $q = 0$ et à $-u < 0$ pour $q = 1$. Il ne s'annule donc pas pour $q = 0$ ou $q = 1$ et il admet entre 0 et 1 soit 1 soit 3 zéros. Précisons en écrivant :

$$\frac{\delta(q)}{q(1-q)} = t + wq - \left(\frac{u}{1-q} - \frac{v}{q}\right) = y_1 - y_2$$

et représentons dans un plan (q, y) la droite D : $y_1 = t + wq$ et la courbe C :

$$y_2 = \frac{u}{1-q} - \frac{v}{q} \qquad \text{avec} \qquad 0 < q < 1$$

cette courbe monte du point $\begin{cases} q=0 \\ y=-\infty \end{cases}$ au point $\begin{cases} q=1 \\ y=+\infty \end{cases}$ car :

$$y'_2 = \frac{u}{(1-q)^2} + \frac{v}{q^2}.$$

Elle traverse Oq en un point $q_1 = \frac{v}{u+v}$.

A. — Supposons qu'elle soit ***rencontrée par la droite D en un seul point Q*** d'abscisse $\overline{q}$. $\delta(\overline{q}) = 0$, donc une fréquence initiale q_0 égale à $\overline{q}$ resterait constante au cours des générations (répartition stationnaire). Dans le cas général $y_1 - y_2$, donc aussi $\delta(q)$, est > 0 si $q < \overline{q}$ et < 0 si $q > \overline{q}$; $\delta(q)$ est donc toujours de signe contraire de $q - \overline{q}$; la différence $q - \overline{q} = r$ va en diminuant constamment en valeur absolue à partir de sa valeur initiale $r_0 = q_0 - \overline{q}$; pour voir si elle tend vers 0 et à quelle vitesse, étudions le quotient $\frac{-\delta(q)}{q-\overline{q}}$; c'est un polynôme du 2e degré au plus, positif et ne s'annulant jamais. Appelons $m > 0$ son minimum dans la région où varie q, c'est-à-dire entre q_0 et $\overline{q}$ (si q_0, et par suite q, est suffisamment voisin de $\overline{q}$, on pourra écrire sensiblement $\delta(q) = \delta'(\overline{q})(q - \overline{q})$, et par suite prendre approximativement $m = \delta'(\overline{q})$). Donc, $\Delta r = r' - r$ désignant la variation de r d'une génération à la suivante, on a :

$$\frac{-\Delta r}{r} > m \qquad \frac{-\Delta \mid r \mid}{\mid r \mid} > m \qquad \Delta \mid r \mid < -m \mid r \mid$$

$\mid r' \mid = \mid r \mid + \Delta \mid r \mid < (1-m) \mid r \mid$, donc quand n générations sont écoulées depuis le début $\mid r \mid < (1-m)^n \mid r_0 \mid$; donc $r = q - \overline{q}$ tend vers 0 au moins comme l'exponentielle $(1-m)^n$. La répartition stationnaire $q = \overline{q}$ considérée au début est stable, *et tout autre répartition tend asymptotiquement vers elle, l'écart* $\mathrm{r} = \mathrm{q} - \overline{\mathrm{q}}$ *étant réduit au bout de* n *générations dans un rapport certainement inférieur à* $(1 - \mathrm{m})^n$.

Signalons quelques cas particuliers importants :

a) Pas de sélection ; les mutations et l'immigration agissent seules : $w = t = o$. D coïncide avec oq, la valeur asymptotique $\overline{q}$ est égale à

$$q_1 = \frac{v}{u+v} \; (= 1/2 \text{ si } u = v).$$

$$\delta(q) = -uq + v(1-q) = -(u+v)(q-\overline{q})$$

donc $m = u + v$; $q - \bar{q}$ est réduit en n générations dans un rapport $< (1 - u - v)^n$; cette réduction n'est appréciable que quand n est de l'ordre de grandeur de $\frac{1}{u+v}$; si u et v se réduisent aux taux de mutation, extrêmement faibles (de l'ordre de 1/100.000), il n'y a d'approche sensible à la valeur asymptotique que si le nombre n de générations écoulées est de l'ordre de 100.000 ; il sera pratiquement impossible d'observer une population devenue stationnaire sous l'action des mutations seules. D'ailleurs l'irrégularité des taux de mutation aussi bien que des taux de migration, rend la validité de la formule assez illusoire ; mais pratiquement ce sera souvent la sélection qui jouera le rôle principal.

b) Sélection seulement gamétique, ou hétérozygote exactement intermédiaire en viabilité : $w = 0$, D est horizontale, d'ordonnée t (coefficient de sélection totale) ; $t < 0$ si pour fixer les idées le gène a est défavorisé ; q tend vers la valeur asymptotique $\bar{q}$ qui est $< q_1 = \frac{v}{u+v}$. Calculons $\bar{q}$; on a :

$$\delta(q) = -tq^2 + (t - u - v)q + v$$

les racines sont :

$$\frac{-t + u + v \pm \sqrt{(t-u-v)^2 + 4vt}}{-2t}$$

$\delta(1)$ est < 0, donc $\bar{q}$, qui est compris entre 0 et 1, est la plus petite des racines, l'autre racine $\bar{\bar{q}}$ s'obtient en prenant le signe + devant le radical, et on a :

$$\delta(q) = -t(q - \bar{q})(q - \bar{\bar{q}})$$

on prendra donc pour m, minimum de $\frac{-\delta(q)}{q - \bar{q}}$, le minimum de $-t(\bar{\bar{q}} - q)$, c'est-à-dire la plus petite des 2 quantités $-t(\bar{\bar{q}} - \bar{q})$ et $-t(\bar{\bar{q}} - q_0)$.

Dans le cas particulier usuel où u et v (supposés réduits aux taux de mutation) sont petits par rapport au coefficient de sélection totale t, les racines sont données par :

$$\frac{-t + u + v}{-2t}\left[1 \pm \sqrt{1 + \frac{4vt}{(-t + u + v)^2}}\right]$$

qui est équivalent à :

$$(1/2)\left[1 \pm \left(1 + \frac{2v}{t}\right)\right] \qquad \text{donc} \qquad \bar{q} \# -v/t \qquad \bar{\bar{q}} \# 1 + v/t$$

la valeur asymptotique $\bar{q} = -v/t$ est petite ; la sélection élimine presque complètement le gène défavorisé a ; seul le taux de mutation v empêche sa disparition complète.

A moins que q_0 ne soit voisin de $\bar{\bar{q}}$ c'est-à-dire de 1, m est de l'ordre de $-t$, alors qu'il ne serait égal qu'à $u+v$ s'il n'y avait pas sélection ; la valeur asymptotique est donc atteinte beaucoup plus rapidement.

B. — La courbe C peut être ***rencontrée par une droite telle que D_1 en 3 points*** (fig. 6) ; ce cas se présentera quand C aura 2 tangentes réelles de même coefficient angulaire w que D_1, et que D_1 sera comprise entre ces tangentes ; or les tangentes parallèles à D_1 ont alors leurs points de contact donnés par $\frac{u}{(1-q)^2} + \frac{v}{q^2} = w$, équation qui a 2 solutions q réelles entre 0 et 1 si w est supérieur au minimum $(u^{1/3}+v^{1/3})^3$ du 1er membre entre 0 et 1 ; si de plus t est compris dans l'intervalle $t_1 \ldots t_2$ des ordonnées à l'origine de ces tangentes, l'équation $\delta(q) = 0$ aura 3 solutions entre 0 et 1 ; soient par ordre de grandeur $\bar{q}_1$, $\bar{q}_2$, $\bar{q}_3$; chacune de ces valeurs donne une répartition stationnaire, se conservant au cours du temps ; mais si on part d'une valeur initiale q_0 différente la figure 6 montre que :

a) Si $q_0 < \bar{q}_2$, $\delta(q) = y_1 - y_2$ est de signe contraire de $q - \bar{q}_1$; la différence $r = q - \bar{q}_1$ va en diminuant en valeur absolue à partir de sa valeur initiale $r_0 = q_0 - \bar{q}_1$; si on appelle $m > 0$ le minimum de $\frac{-\delta(q)}{q-\bar{q}}$ dans l'intervalle $q_0 \ldots \bar{q}_1$, l'écart r est encore réduit au bout de n générations dans un rapport $< (1-m)^n$, q *tend vers la valeur asymptotique* $\bar{q}_1$.

b) Si $q_0 > \bar{q}_2$, le même raisonnement montre que q tend vers la valeur asymptotique $\bar{q}_3$. La racine intermédiaire $\bar{q}_2$ de $\delta(q)$ correspond donc à un état stationnaire instable ; suivant que q_0 est $<$ ou $>$ que $\bar{q}_2$, ou tend vers les valeurs stationnaires stables $\bar{q}_1$ ou $\bar{q}_3$.

C. — Etudions directement le cas d'une ***sélection quelconque***, mais ***avec mutations et migrations négligeables*** : $u = v = 0$; ce cas ne rentre pas directement dans l'étude précédente, la courbe C étant alors dégénérée ; on a alors :

$$\delta(q) = q(1-q)(t+wq) = wq(1-q)(q-\alpha) \qquad \text{avec} \qquad \alpha = -t/w$$

(α peut être intérieur ou extérieur à l'intervalle 0 ... 1).

Les valeurs stationnaires sont :

$$q = 0, \qquad q = 1, \qquad \text{et} \qquad q = \alpha \qquad \text{si } 0 < \alpha < 1.$$

a) Si $\alpha > 1$ ou $\alpha < 0$, $\delta(q)$ a un signe constant; si pour fixer les idées il est négatif, q décroît toujours; — $\delta(q)/q$ a un minimum positif m; et on en déduit que q tend vers 0 plus rapidement que $(1-m)^n$: *le gène* a *est éliminé quelle que soit sa fréquence initiale* (s'il y avait de très faibles taux de mutation, *a* subsisterait avec une fréquence faible, comme dans le cas de la sélection gamétique). Si $\delta(q)$ est positif, $q \to 1$: *fixation du gène* a *quelle que soit sa fréquence initiale.*

b) Si $0 < \alpha < 1$, 2 cas sont à distinguer :

1° *si* $w > 0$, $\delta(q)$ a toujours le signe de $q - \alpha$; la variation de q, donc de $q - \alpha$, est du signe de $q - \alpha$. $q - \alpha$ augmente en valeur absolue à partir de sa valeur initiale $q_0 - \alpha$. On voit comme précédemment que q tend vers 0 si $q_0 < \alpha$ et q tend vers 1 si $q_0 > \alpha$. *Il y a encore élimination d'un des gènes, mais cette fois la nature du gène éliminé dépend de la fréquence initiale;*

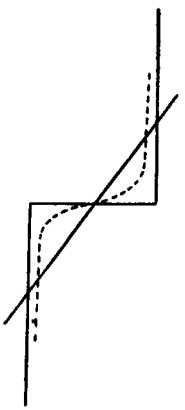

Fig. 7.

2° *si* $w < 0$, $\delta(q)$ a toujours le signe contraire de $q - \alpha$; on voit encore que l'écart $r = q - \alpha$ décroît en valeur absolue et tend vers 0. *Dans la répartition asymptotique les 2 gènes* a *et A coexistent, avec les fréquences stationnaires* α *et* $1 - \alpha$. Il est aisé de voir qu'il en est ainsi dans le cas particulier important : $s = 0$, $\sigma < 0$, $h > 1$: sélection uniquement zygotique, hétérozygote supérieur en viabilité aux 2 homozygotes, à condition que la consanguinité ne soit pas trop forte. En effet on a $w < 0$, et

$$\alpha = -t/w = \left(h + \frac{\lambda}{1-\lambda}\right)/(2h-1) \text{ est positif}$$

mais n'est < 1 que si $\frac{\lambda}{1-\lambda} < (h-1)$, ce qui nécessite $1/(1-\lambda) < h$, c'est-à-dire $\lambda < 1 - 1/h$.

Remarque : on vérifie aisément que la discussion de C) : $u = v = 0$, rentre comme cas particulier dans la discussion graphique de A) ou de B) à condition de regarder alors la courbe C comme dégénérée en la ligne brisée :

$$\begin{cases} q = 0 \\ y < 0 \end{cases} \quad \begin{cases} y = 0 \\ 0 < q < 1 \end{cases} \quad \begin{cases} q = 1 \\ y > 0 \end{cases}$$ (fig. 7, ligne en trait plein) (1).

(1) C sera en effet rencontrée en un seul point par la droite D si $\alpha > 1$ ou $\alpha < 0$ (cas *a*) ou si $0 < \alpha < 1$ et $w < 0$ (cas *b*, 2°); et en 3 points si $0 < \alpha < 1$ et $w > 0$ (cas *b*, 1°).

Il en résulte que, si u et v sont petits par rapport à t et w mais non nuls (ligne en pointillé), la discussion sera sensiblement la même que dans le cas C), à cela près que l'élimination et la fixation seront remplacées par un équilibre asymptotique correspondant à une fréquence $\overline{q}$ voisine de 0 ou de 1.

2° *Cas d'une population limitée.*

Soit N le nombre d'individus à chaque génération. Si q est la fréquence de a dans F_n, nous avons vu que la probabilité de a dans F_{n+1} sera $q + \delta(q)$, $\delta(q)$ pouvant être représenté en 1re approximation par un polynôme du 3e degré. Mais la fréquence q_1 de a dans F_{n+1} différera de la probabilité $q + \delta(q)$, car cette fréquence est une variable aléatoire dont $q + \delta(q)$ ne représente que la valeur moyenne ; quand la loi de probabilité de cette aléatoire q_1 est connue (en fonction naturellement de la valeur de q), soit $\theta(q, q_1)dq_1$, les fréquences de a dans les générations successives se présentent comme des aléatoires en chaîne simple de Markoff, la chaîne étant définie par la loi de passage $\theta(q, q_1)dq_1$ (que nous supposerons indépendante du rang n de la génération, ce qui est admissible si N est constant). Si nous admettons que le passage en une génération, ou en tout cas en un certain nombre de générations, de n'importe quelle fréquence q à n'importe quelle fréquence q_1 est possible (hypothèse qui implique que les taux de mutation u et v ne soient pas nuls, sans quoi on ne pourrait passer des valeurs $q = 0$ ou $q = 1$ à des valeurs différentes), $\theta(q, q_1)$ est alors toujours > 0 et le théorème de Markoff nous apprend alors que la loi de probabilité *a priori* $\varphi_n(q)dq$ de la fréquence q dans la génération F_n tend quand n tend vers ∞ vers une loi limite $\varphi(q)dq$ indépendante de la valeur initiale de q.

Il est possible d'obtenir la forme de ces lois en faisant quelques hypothèses sur la loi de passage $\theta(q, q_1)dq_1$ (loi de probabilité de q_1 quand q est fixé) ; nous supposerons que c'est une loi de Gauss de valeur moyenne $q + \delta(q)$ ($\delta(q)$ étant petit et tel que $\delta(0) \geqslant 0$ et $\delta(1) \leqslant 0$), et de fluctuation petite $\sigma^2 = w(q) \geqslant 0$ (ne s'annulant que pour $q = 0$ et $q = 1$).

Si nous supposons par exemple que les 2N gamètes qui donnent naissance à F_{n+1} sont tirées au hasard parmi un nombre infiniment grand de gamètes produites par F_n et présentant sensiblement les fréquences q et $1 - q$ pour a et A, on sait que la loi de probabilité de la fréquence de a dans F_{n+1} sera sensiblement gaussienne, et que la fluctuation de cette fréquence par rapport à sa valeur moyenne sera

$\sigma^2 = w(q) = \frac{q(1-q)}{2N}$, qui ne s'annule bien que pour $q=0$ et $q=1$.

(Plus généralement, si en raison de la consanguinité dans F_n les N zygotes de F_{n+1} sont chacun tiré au hasard avec les probabilités :

$$P = p(p+\lambda q), \qquad 2Q = 2pq(1-\lambda), \qquad R = q(q+\lambda p)$$

pour les 3 états AA, Aa, aa (λ étant le coefficient de consanguinité moyen de F_{n+1}) la fluctuation de la fréquence q_1 de a dans F_{n+1} par rapport à son espérance mathématique q sera encore $\sigma^2 = \frac{q(1-q)}{2N}$).

a) ***Equation fondamentale.*** — Dans le passage de la génération F_n à la génération F_{n+1}, la loi de probabilité *a priori* de la fréquence passe de :

$$\varphi_n(q)dq \text{ à } \varphi_{n+1}(q_1)dq_1 = dq_1 \int_0^1 \varphi_n(q)\theta(q, q_1)dq$$

si on appelle M_i et M'_i les moments de la loi de probabilité dans F_n et dans F_{n+1} on a :

$$M_i = \int_0^1 q^i \varphi_n(q)dq$$

$$M'_i = \int_0^1 q_1^i \varphi_{n+1}(q_1)dq_1 = \int_0^1 \left[\int_0^1 q_1^i \theta(q, q_1)dq_1\right]\varphi_n(q)dq$$

$$= \int_0^1 \mu_i(q)\varphi_n(q)aq$$

(en intervertissant les intégrations, ce qui est légitime pour des fonctions bornées et intégrables dans des intervalles finis).

Les $\mu_i(q)$ étant les moments de la loi de Gauss $\theta(q, q_1)dq_1$, dont la moyenne est $q+\delta(q)$ et la fluctuation $w(q)$; δ et w étant petits, ces moments se calculent en développant suivant les puissances de t la fonction caractéristique :

$$\exp.\,[(q+\delta)\tau + w\tau^2/2] = 1 + (q+\delta)\tau + w\tau^2/2 + \ldots + [(q+\delta)\tau + w\tau^2/2]^i/i\,! + \ldots$$

On voit que, en négligeant les termes de l'ordre de w^2 et de δ^2 :

$$\mu_i/i\,! = (q+\delta)^i/i\,! + (i-1)(q+\delta)^{i-2}w/2(i-1)\,! + 0(w^2)$$

$$\mu_i = q^i + i\delta q^{i-1} + \frac{i(i-1)}{2} wq^{i-2} + 0(w^2) + 0(\delta^2) + 0(w\delta)$$

donc la variation des moments d'une génération à la suivante est :

$$(1)\quad \left\{\begin{aligned} M'_i - M_i &= \int_0^1 [\mu_i(q) - q\,]\varphi_n(q)dq \\ &= i\int_0^1 \delta(q)q^{i-1}\varphi_n(q)dq + \frac{i(i-1)}{2}\int_0^1 \omega(q)q^{i-2}\varphi_n(q)dq. \end{aligned}\right.$$

Si on suppose que $\delta(q)$ et $\omega(q)$ sont représentables par des polynômes, ce qui est une hypothèse d'une grande généralité pratique, d'ailleurs vérifiée exactement par les formes particulières que nous avons indiquées, on écrira :

$$\delta(q) = \sum_{l\geq 0} A_l q^l,\quad \omega(q) = \sum_{l\geq 0} B_l q^l.$$

En assimilant la variation petite $M'_i - M_i$ à une dérivée dM_i/dt (le temps t étant mesuré en générations), les équations (1) se transforment en un système différentiel pour les moments :

$$(2)\qquad \frac{dM_i}{dt} = i\sum A_l M_{i-1+l} + \frac{i(i-1)}{2}\sum B_l M_{i-2+l}$$

ce système n'est pas résoluble directement, car en général il intervient dans les deuxièmes membres des moments d'ordre plus élevé que dans les premiers ; mais il nous permet d'obtenir une équation aux dérivées partielles pour la fonction caractéristique (ou transformée de Laplace), de la loi de probabilité $\varphi(q, t)dq$ dont les moments sont les $M_i(t)$. Cette transformée est en effet :

$$F(s, t) = \int_0^1 e^{sq}\varphi(q, t)dq = \sum_{p\geq 0} M_p(t)s^p/p!$$

de dérivées

$$\frac{\partial^k F}{\partial s^k} = \sum M_p s^{p-k}/(p-k)!$$

(elle est toujours définie, puisqu'on n'intègre qu'entre 0 et 1). En multipliant les équations (2) par $s^{i-1}/i!$ et sommant depuis $i=0$ jusqu'à $+\infty$ on obtient :

$$(3)\qquad \frac{1}{s}\frac{\partial F}{\partial t} = \sum A_l \frac{\partial^l F}{\partial s^l} + \frac{s}{2}\sum B \frac{\partial^l F}{\partial s^l}.$$

D'après les propriétés bien connues de la transformation de Laplace,

en posant

$$F(s, t) = \mathcal{L}[\varphi(q, t)],$$

on a :

$$\frac{\partial F}{\partial t} = \mathcal{L}\left[\frac{\partial \varphi}{\partial t}\right] \qquad \frac{\partial^i F}{\partial s^i} = \int_0^1 e^{sq} q^i \varphi(q, t) dq = \mathcal{L}[q^i \varphi(q, t)]$$

$$\frac{1}{s}\frac{\partial F}{\partial t} = \frac{1}{s}\int_0^1 e^{sq}\frac{\partial \varphi}{\partial t} dq = \frac{1}{s}\int^1 e^{sq}\frac{\partial V}{\partial q} dq = -\int_0^1 e^{sq} V dq = -\mathcal{L}[V],$$

en posant :

$$V = \int_0^q \frac{\partial \varphi}{\partial t} dq = \frac{\partial}{\partial t}\left[\int_0^q \varphi dq\right]$$

et remarquant que $V \equiv 0$ pour $q = 0$ et pour $q = 1$.

Puisque 2 fonctions dont les transformées de Laplace sont les mêmes sont identiques presque partout, on tire de (3) :

$$-\frac{\partial}{\partial t}\left[\int_0^q \varphi(q, t) dq\right] = \sum A_i q^i \varphi(q, t) - 1/2 \frac{\partial}{\partial q}\left[\sum B_i q^i \varphi(q, t)\right]$$

c'est-à-dire :

$$(4) \qquad \boxed{\frac{\partial}{\partial t}\left[\int_0^q \varphi(q, t) dq\right] = (1/2)\frac{\partial}{\partial q}[w(q)\varphi(q, t)] - \delta(q)\varphi(q, t)}$$

Telle est l'équation fondamentale.

b) ***Loi de probabilité asymptotique.*** — Si nous appelons $\varphi(q)dq$ loi de probabilité asymptotique pour t infini, ce sera d'après la théorie de Markoff la loi de probabilité stationnaire vérifiant (4) c'est-à-dire telle que :

$$(5) \qquad (1/2)\frac{\partial}{\partial q}[w\varphi] - \delta\varphi = 0,$$

c'est donc la loi dont la densité de probabilité est :

$$(5') \qquad \varphi(q) = [K/w(q)] \exp.\left[2\int \frac{\delta(q)}{w(q)} dq\right].$$

En particulier, quand :

$$w = q(1-q)/2N \quad \text{et} \quad \delta(q)/q(1-q) = -u/(1-q) + v/q + t + wq,$$

on a :

$$(6) \qquad \underline{\varphi(q) = K_1 q^{4Nv-1}(1-q)^{4Nu-1} \exp.\,[2N(wq^2 + 2tq)]dq,}$$

K_1 étant déterminé de façon que l'intégrale entre 0 et 1 soit égale à 1.

Cette formule qui a été donnée par S. Wright (17) dans des cas particuliers mais sans démonstration générale représente la probabilité pour que, dans une population limitée de N individus, un gène a ayant des coefficients de mutation, migration et sélection donnés, ait, au bout d'un nombre de générations infiniment grand, une fréquence comprise entre q et $q + dq$. Elle représente donc aussi la loi de répartition asymptotique du gène a, au bout d'un temps infiniment grand, dans un nombre infiniment grand de populations de même effectif N et où tous les coefficients seraient les mêmes.

Indiquons quelques *cas particuliers* :

a) Si $u = 0$, ou $v = 0$, K_1 est nécessairement nul, puisque l'intégrale entre 0 et 1 de $1/q$ ou de $1/(1-q)$ est infinie. Ce résultat ne fait que traduire la certitude de l'élimination ou de la fixation finale des gènes sur lesquels la mutation ou la migration n'agissent pas.

b) Si $4Nu$ et $4Nv$ sont > 1, c'est-à-dire si l'effectif de la population est suffisamment élevé et les taux de mutation ou de migration pas trop faibles, $\varphi(q)$ est nul pour $q = 0$ et $q = 1$, et est représenté par une courbe en cloche (fig. 8) ayant une ou plusieurs dominantes q_1 données par l'équation $\frac{\partial \varphi}{\partial q} = 0$,

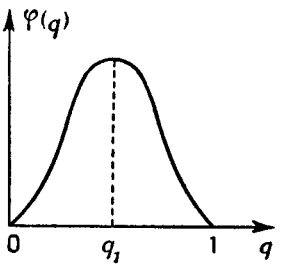

Fig. 8.

c'est-à-dire $4N\delta(q_1) + 2q_1 - 1 = 0$ (équation qui, pour N très grand, devient $\delta(q) = 0$, c'est-à-dire redonne la ou les valeurs asymptotiques $\bar{q}$ dans une population très nombreuse, valeurs étudiées au 1°).

Nous obtiendrons des résultats simples en supposant qu'il y a une seule dominante q_1, et que q en reste voisin avec une probabilité peu différente de 1 ; on peut alors en première approximation remplacer $\delta(q)$ par une fonction linéaire de q, soit :

$$\delta(q) = -k(q - \bar{q}), \tag{7}$$

$\bar{q}$ étant nécessairement la valeur asymptotique dans une population très nombreuse [1°, cas A) ; ou cas C) sauf *b*) 1°] et k étant d'après la formule de Taylor égal à $-\delta'(\bar{q})$, donc de l'ordre de grandeur du plus grand des nombres u, v, w, t [La formule (7) serait d'ailleurs rigoureuse si $\delta(q)$ était linéaire, c'est-à-dire s'il n'y avait pas de sélection : 1°, A), cas *a*)]. La dominante q_1 est alors donnée par :

$$-4Nk(q_1 - \bar{q}) + 2q_1 - 1 = 0 \qquad \text{d'où} \qquad q_1 = \frac{4Nk\bar{q} - 1}{4Nk - 2} \neq \bar{q}$$

si $4Nk$ est grand, la dominante coïncide alors sensiblement avec la valeur asymptotique dans une population infinie. La distribution asymptotique (5') s'écrit alors :

$$\varphi(q) = K_1 q^{4N\bar{k}q-1}(1-q)^{4Nk(1-\bar{q})-1} \qquad \text{avec} \qquad K_1 = B(4k\bar{p}, 4k\bar{q}), \ (^1)$$

et ses moments sont donnés par les formules (2) qui s'écrivent :

$$0 = i(-kM_i + k\bar{q}M_{i-1}) + \frac{i(i-1)}{4N}(M_{i-1} - M_i),$$

d'où, en partant de $M_0 = 1$:

$$M_1 = \bar{q} \qquad \text{et} \qquad 2k(M_2 - \bar{q}^2) = \frac{1}{2N}(\bar{q} - M_2)$$

d'où la fluctuation :

$$\sigma^2 = M_2 - \bar{q}^2 = \frac{\bar{q}(1-\bar{q})}{4Nk+1} \# \frac{\bar{q}(1-\bar{q})}{4Nk}$$

si $4Nk$ est grand, c'est-à-dire si l'ordre de grandeur de k est supérieur à celui de $1/N$; σ^2 est alors faible, la distribution est « concentrée », et

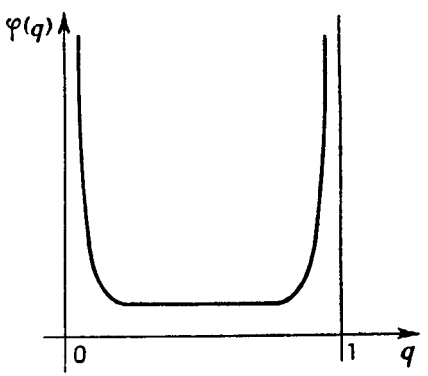

Fig. 9.

c'est dans ce cas qu'il est légitime d'admettre que q varie pratiquement dans un intervalle de faible étendue et que $\delta(q)$ est linéaire.

Nous voyons que sur un grand nombre de populations se trouvant dans les mêmes conditions et ayant toutes le même effectif N, suffisamment élevé pour que $4Nk$ soit grand, les fréquences asymptotiques observées

(1) B désignant l'intégrale eulérienne.

seront presque toutes groupées au voisinage de la valeur correspondant à une population infinie. La mesure expérimentale de la fluctuation de ces fréquences permettra, connaissant N, de déterminer k (alors que celle de $\overline{q}$ donne $\frac{v}{u+v}$ ou $-\frac{v}{t}$ ou $\frac{h+\lambda/1-\lambda)}{2h-1}$).

c) Si $4Nu$ et $4Nv$ sont <1, $\varphi(q)$ est infini pour $q=0$ et $q=1$ et est représenté par une courbe en U (fig. 9). K_1 est d'autant plus petit que u et v sont plus petits. Ce sont alors les fréquences proches de $q=0$ et $q=1$ qui ont de beaucoup la plus forte probabilité. La plupart des gènes sont près de se fixer ou de disparaître, et n'en sont empêchés que pour les mutations récurrentes ou la migration renouvelée. Il y a ainsi une différence essentielle entre le cas d'une population très limitée ou à taux de mutation et migration très faibles, avec tendance à l'extinction ou à la fixation des gènes, et le cas d'une population nombreuse avec presque stabilisation de chaque gène au voisinage d'une fréquence déterminée.

c) ***Evolution de la loi de probabilité au cours du temps.*** — Appelons $\varphi(q, t)$ et $\Phi(q, t)=\int^q \varphi(q, t)dq$ les lois de probabilité élémentaire et intégrale à la date t, vérifiant (4); appelons $\varphi(q)$ et $\Phi(q)=\int_0^q \varphi(q)dq$ les lois asymptotiques, déduites de (5); désignons par $R(q, t)=\Phi(q, t)-\Phi(q)$ l'écart entre la loi intégrale à l'instant t et la loi asymptotique. Cet écart est connu à l'instant initial : $R(q,0)\equiv R_0(q)$ donné; il satisfait aux conditions, aux limites $R(0,t)\equiv R(1, t)\equiv 0$; et il vérifie évidemment l'équation (obtenue en retranchant (5) de (4)) :

$$\frac{\partial R}{\partial t}=\frac{1}{2}\frac{\partial}{\partial q}\left[w(q)\frac{\partial R}{\partial q}\right]-\delta(q)\frac{\partial R}{\partial q}. \tag{7}$$

On le déterminera donc en cherchant les solutions de (7) qui s'annulent pour $q=0$ et $q=1$ et sont de la forme $R\equiv K(q)\times L(t)$; ces solutions doivent vérifier :

$$\frac{L'(t)}{L(t)}=\frac{w}{2}\frac{K''(q)}{K(q)}+\left(\frac{w'}{2}-\delta\right)\frac{K'(q)}{K(q)}$$

ce qui nécessite :

$$L(t)=e^{-\lambda t} \tag{8}$$

$$wK''/2+(w'/2-\delta)K'+\lambda K=0 \quad \text{avec} \quad K(0)=K(1)=0. \tag{9}$$

La dernière équation ne peut être satisfaite avec ces conditions aux limites que si la constante λ appartient à la suite des « valeurs propres » λ_i qui sont réelles, positives, et qu'on peut supposer rangées par ordre de grandeur croissante. En appelant $K_i(q)$ la « solution propre » correspondant à λ_i, on voit que toute série :

$$R(q, t) \equiv \sum_i A_i e^{-\lambda_i t} K_i(q)$$

vérifiera à la fois (7) et les conditions aux limites. Elle vérifiera de plus la condition initiale $R(q, 0) \equiv R_0(q)$ si les coefficients A_i sont choisis de façon que $\sum_i A_i K_i(q) \equiv R_0(q)$, c'est-à dire s'ils sont donnés par le développement de la fonction $R_0(q)$ en série de fonctions $K_i(q)$. On sait que ce développement est possible pour une fonction $R_0(q)$ continue et nulle aux bornes $q = 0$ et $q = 1$. Pour le formuler il suffit de mettre l'équation (9) sous la forme réduite :

$$\frac{d^2\overline{K}}{dr^2} = \frac{-2\lambda}{w(q)\varphi^2(q)}\overline{K},$$

en désignant par r la nouvelle variable $\int_0^q \varphi(q)dq$ [qui n'est autre que la fonction de probabilité totale $\Phi(q)$]. On sait alors que les solutions propres $\overline{K}_i(r)$ sont orthogonales (et peuvent être choisies normées) par rapport à la fonction $1/w\varphi^2(q)$, c'est-à-dire que $\int_0^1 \frac{\overline{K}_i(r)\overline{K}_j(r)}{w\varphi^2(q)}\,dr = 0$, ou, en revenant à la variable q :

$$\int_0^1 \frac{K_i(q)K_j(q)}{w\varphi(q)}\,dq = 0 \qquad \text{avec} \qquad \int_0^1 \frac{[K_i(q)]^2}{w\varphi(q)}\,dq = 1.$$

Les coefficients A_i du développement de $R_0(q)$ sont donc de la forme :

$$A_i = \int_0^1 \frac{R_0(q)K_i(q)}{w\varphi(q)}\,dq,$$

ce qui donne la solution du problème :

(10) $$R(q, t) = \sum_{i=1}^{\infty} A_i e^{-\lambda_i t} K_i(q) \qquad \text{(série uniformément convergente).}$$

On voit que l'ordre de grandeur de la décroissance de l'écart $R(q, t)$ entre la loi à l'instant t et la loi asymptotique est l'ordre de $e^{-\lambda_1 t}$, λ_1 étant la première valeur propre (à moins que l'écart initial $R_0(q)$ ne

soit orthogonal, relativement à $1/w\varphi(q)$, à la fonction $K_1(q)$). La vitesse du processus est ainsi caractérisée.

Il est aisé de résoudre complètement le problème dans le cas étudié précédemment, où $\delta(q)$ peut être remplacé par la fonction linéaire $\delta(q) = -k(q-\bar{q})$. Alors l'équation (9) où $w = q(1-q)/2N$, devient l'équation de Gauss :

$$(9') \qquad q(1-q)K'' + [1 - 2q + 4Nk(q-q)]K' + 4N\lambda K = 0.$$

Les paramètres de Gauss sont ici : α et β racines de (11) :

$$(11) \qquad \alpha^2 + (4Nk-1)\alpha - 4N\lambda = 0 \quad \text{et} \quad \gamma = 1 - 4Nk\bar{q}.$$

En appelant $F(\alpha, \beta, \gamma, q)$ la série hypergéométrique, la solution générale de (9') est :

$$C_1F(\alpha, \beta, \gamma, q) + C_2q^{4Nk\bar{q}}F(\alpha', \beta', \gamma', q)$$

avec :

$$\alpha' = \alpha + 1 - \gamma, \qquad \beta' = \beta + 1 - \gamma, \qquad \gamma' = 2 - \gamma.$$

Les solutions qui s'annulent pour $q = 0$ correspondent à $C_1 = 0$. Il y aura donc des « solutions propres » s'annulant à la fois pour $q = 0$ et $q = 1$ à condition que soit nul $F(\alpha', \beta', \gamma', 1)$ qui d'après la théorie de l'équation de Gauss est égal à $\dfrac{\Gamma(\gamma')\Gamma(\gamma' - \alpha' - \beta')}{\Gamma(\gamma' - \alpha')\Gamma(\gamma' - \beta')}$; ce qui exige que α ou β soit égal à un entier $n \geqslant 1$, c'est-à-dire que l'équation (11) ait une racine entière positive n, ce qui donne pour λ les « valeurs propres » $\lambda_n = n^2/4N + n(k - 1/4N)$, valeurs qui croissent de $\lambda_1 = k$ jusqu'à $+\infty$.

Les solutions propres normées correspondantes sont les fonctions hypergéométriques :

$$K_n(q) = h_nF(n + 4Nk\bar{q}, 1 - n - 4Nk + 4Nk\bar{q}, 1 + 4Nk\bar{q}, q).$$

Les constantes h_n étant choisies de façon que :

$$\int_0^1 \frac{[K_n(q)]^2}{q^{4Nk\bar{q}}(1-q)^{4Nk(1-\bar{q})}}\, dq = 1.$$

Les coefficients A_n sont donnés par :

$$A_n = \int_0^1 \frac{R_0(q)K_n(q)}{q^{4Nk\bar{q}}(1-q)^{4Nk(1-\bar{q})}}\, dq.$$

Et l'écart est fourni par la formule (10).

Puisque $\lambda_1 = k$, l'ordre de grandeur de la décroissance de cet écart sera en général celui de e^{-kt} ; le nombre t de générations nécessaires pour se rapprocher appréciablement de l'état d'équilibre asymptotique sera donc de l'ordre de grandeur de $1/k$. Nous avons vu p. 49 que quand $\delta(q)$ a la forme générale donnée p. 39 mais que la distribution reste au cours du temps suffisamment concentrée autour de la valeur $\bar{q}$, on prend :

$$k = -\delta'(\bar{q}) = u + v - (1 - 2\bar{q})t - w\bar{q}(2 - 3\bar{q}),$$

k est donc de l'ordre de grandeur de la plus grande (en valeur absolue) des quantités u, v, t, w. Si toutes ces quantités sont petites, $1/k$ est grand et le nombre de générations nécessaires pour atteindre sensiblement l'état d'équilibre est considérable. Une population naturelle ne peut donc alors être présumée avoir atteint l'état d'équilibre que si les conditions y sont restées les mêmes pendant une très longue période.

La méthode précédente ne s'applique plus dans le cas où il n'y a ni mutations ni migrations : $u = v = 0$; car alors K est nul, et la densité de probabilité asymptotique $\varphi(q)dq$ est nulle partout entre 0 et 1, toute la probabilité est concentrée aux extrémités $q = 0$ et $q = 1$; la manière dont cet état asymptotique est atteint peut être étudiée par une méthode différente (13).

III. — INFLUENCE DE LA MIGRATION

L'hypothèse par laquelle S. Wright traduit la migration ne s'appliquerait bien qu'à la population d'une île recevant des immigrants d'une population continentale nombreuse et à composition constante. Un schéma plus près de la réalité courante, tenant compte de l'interaction, par migration, des groupes les uns sur les autres sera le suivant : soit une population répartie sur une aire A et présentant au point P de coordonnées x et y, une densité $\delta(P)$. Nous supposerons que chaque individu, entre sa naissance et sa phase reproductrice, présente une probabilité connue $f(P, Q)dS_Q$ d'émigrer depuis le point P jusque dans une aire élémentaire dS_Q centrée au point Q $\left(\int\int_A f(P, Q)dS_Q = 1\right)$. D'après la formule de Bayes, chaque parent d'un individu naissant en un point Q aura la probabilité connue :

$$g(P, Q)dS_P = \delta(P)f(P, Q)dS_P \Big/ \int\int \delta(P)f(P, Q)dS_P$$

d'être né dans une aire dS_P centrée au point P $\left(\int\int_A g(P, Q)dS_P = 1\right)$.

Soit $X_n(C)$ la variable aléatoire représentant l'état d'un locus mendélien d'un individu de la $n^{ième}$ génération né au point C. *A priori*, $X_n(C)$ prendra les valeurs 1 ou 0, correspondant aux états allèles a ou A, avec des probabilités q et $p = 1 - q$, dépendant du point C et du rang n de la génération, et les X_n relatives à 2 points C différents seront en liaison stochastique. Les aléatoires $X_{n+1}(D)$ relatives à la génération suivante ont des probabilités bien déterminées par la connaissance des valeurs des $X_n(C)$, par suite, d'après la théorie des probabilités en chaîne, les probabilités des $X_n(C)$ et leurs liaisons tendront à la longue vers un *état stationnaire*, indépendant du rang n de la génération. *C'est cet état stationnaire que nous nous proposons d'étudier.*

Si u et v sont les probabilités de mutation, à chaque génération, de a en A et de A en a, l'espérance mathématique de l'aléatoire X' relative à un locus d'un enfant provenant d'un parent déterminé sera :

$$\mathfrak{M}(X') = (1 - u)X + v(1 - X),$$

X étant la valeur de l'aléatoire chez le parent. Ceci peut s'écrire :

$$\mathfrak{M}(X') = (1 - k)X + kc,$$

en appelant c la quantité $\overline{q} = v/(u + v)$ de p. 41 et k la quantité $v + u$ traduisant la *pression de mutation* (1). Puisqu'il n'y a pas d'autre liaison stochastique entre des enfants que celle qui provient de la liaison éventuelle entre leurs parents, les moments conjoints $\mathfrak{M}[X'(C)X'(D)]$ d'un certain nombre d'aléatoires X' de la $(n + 1)^{ième}$ génération seront des fonctionnelles linéaires des produits des quantités X de leurs parents, si ceux-ci sont connus ; ou, s'ils sont inconnus, des espérances mathématiques de ces produits, c'est-à-dire des moments conjoints de la $n^{ième}$ génération. En égalant les moments conjoints des 2 générations, on obtiendra des équations intégrales linéaires pour déterminer ces moments. Nous nous bornerons à indiquer le calcul pour les moments d'ordre 1 et 2.

L'espérance mathématique E(Q) de X(Q) sera donnée par :

$$\mathfrak{M}[X(Q)] = \int\int_A \{ (1 - k)\mathfrak{M}[X(P)] + kc \} g(P, Q) dS_P$$

(1) On sait que, s'il y a une *pression de sélection* constante, en faveur de l'hétérozygote, elle se traduit en moyenne pour un grand nombre d'individus par une formule de même forme (k et c ayant naturellement d'autres valeurs). Le calcul que nous allons faire sera donc en 1re approximation applicable à ce dernier cas (mais il exclut naturellement la « sélection géographique » dépendant du lieu).

c'est-à-dire :

$$E(Q) = \int\int_A (1-k)E(P)g(P, Q)dS_P + kc$$

équation dont l'unique solution, si $k = u + v > 0$, est :

$$E(P) = C^{te} = c = v/(u+v).$$

L'espérance mathématique c est donc indépendante de la position géographique. Nous poserons dorénavant $X - c = Y$, on aura $\mathfrak{M}(Y) = 0$, et d'une génération à l'autre $\mathfrak{M}(Y') = (1-k)Y$.

La fluctuation de X, ou de Y, sera :

$$s^2 = \mathfrak{M}(Y^2) = \mathfrak{M}(X^2) - c^2 = c(1-c).$$

Le 1er moment conjoint de 2 aléatoires Y(C) et Y(D) de la même génération sera désigné par $\mathfrak{M}[Y(C)Y(D)] = s^2\varphi(C, D)$; $\varphi(C, D)$ sera donc le *coefficient de corrélation* de ces 2 aléatoires; nous appellerons $\varphi(C, C)$ sa limite, évidemment < 1, quand D devient infiniment voisin de C, tout en en *restant distinct.*

2 loci $Y_{n+1}(C)$ et $Y_{n+1}(D)$ de 2 individus de la $(n+1)^{ième}$ génération nés en C et D auront la probabilité $g(E, C)g(F, D)dS_E dS_F$ de provenir de parents nés en E et F, et la probabilité $g(E, C)g(F, D)dS_E^2$ de provenir de parents nés tous 2 dans un même voisinage dS_E du même point E; dans ce dernier cas, ils auront la probabilité $1/[2\delta(E)dS_E]$ de provenir d'un même locus d'un même parent, et la probabilité $1 - 1/[2\delta(E)dS_E]$ de provenir de loci infiniment voisins mais différents. On a donc quand les lieux de naissance des parents sont connus :

$$\mathfrak{M}[Y_{n+1}(C)Y_{n+1}(D)] = (1-k)^2 Y_n(E)Y_n(F)$$

et quand ils sont inconnus :

$$\mathfrak{M}[Y_{n+1}(C)Y_{n+1}(D)] = (1-k)^2 \int\int_A \int\int_A \mathfrak{M}[Y_n(E)Y_n(F)]g(E, C)g(F, D)dS_E dS_F.$$

$\mathfrak{M}[Y_n(E)Y_n(F)]$ devant être pris égal à $s^2\varphi_n(E, F)$ si les éléments d'aire dS_E et dS_F sont distincts, mais s'ils sont confondus à :

$$[1/2\delta(E)dS_E]\mathfrak{M}[Y_n(E)^2] + [1 - 1/2\delta(E)dS_E]s^2\varphi_n(E, E)$$

c'est-à-dire à :

$$s^2\varphi_n(E, E) + [1/2\delta(E)dS_E]\, s^2[1 - \varphi_n(E, E)];$$

on a donc en divisant par s^2 :

$$(1)\quad \begin{cases} \varphi_{n+1}(C, D) = (1-k)^2 \int\int_A \int\int_A \varphi_n(E, F) g(E, C) g(F, D) d\mathrm{S}_E d\mathrm{S}_F \\ \qquad + (1-k)^2 \int\int_A \frac{1-\varphi_n(E, E)}{2\delta(E)} g(E, C) g(E, D) d\mathrm{S}_E. \end{cases}$$

Dans l'état stationnaire $\varphi_{n+1} \equiv \varphi_n \equiv \varphi(C, D)$ serait donné, si $\varphi(E, E)$ était connu, par une équation de Fredholm qui admet $1-k=1$ pour valeur singulière de module minimum, et qui, par suite, a, si $k>0$, une seule solution, continue si $g(E, C)$ n'admet que des discontinuités de la 1re sorte. La méthode d'approximations successives est applicable et fournit pour cette solution :

$$(2)\quad \begin{cases} \varphi(C, D) = \int\int_A \frac{1-\varphi(E, E)}{2\delta(E)} [(1-k)^2 g(E, C) g(E, D) \\ \qquad + (1-k)^4 g_1(E, C) g_1(E, D) + \dots \\ \qquad + (1-k)^{2n+2} g_n(E, C) g_n(E, D) + \dots] d\mathrm{S}_E \end{cases}$$

en posant :

$$g_1(E, C) = \int\int_A g(E, F) g(F, C) d\mathrm{S}_F, \dots,$$

$$g_n(E, C) = \int\int_A g_{n-1}(E, F) g(F, C) d\mathrm{S}_F.$$

En faisant ensuite $E \equiv C$, on obtient une 2e équation de Fredholm pour déterminer $\varphi(E, E)$:

$$(3)\quad \begin{cases} \varphi(C, C) = \int\int_A \frac{1-\varphi(E, E)}{2\delta(E)} [(1-k)^2 g^2(E, C) + \dots \\ \qquad + (1-k)^{2n+2} g_n^2(E, C) + \dots] d\mathrm{S}_E \end{cases}$$

équation qui a en général une seule solution $\varphi(E, E)$. En la reportant dans (2), on obtient $\varphi(C, D)$.

Cas particulier de la migration « homogène et isotrope ».

Supposons que l'aire occupée par la population puisse être regardée comme illimitée, que la densité $\delta(E)$ soit constante (dans l'espace et dans le temps), et que $f(P, Q)$ ne dépende que de la distance $PQ = r$; alors $g(P, Q)$ lui est égal, nous poserons $g(P, Q) = g(r)$ (fonction

d'une seule variable et non de plus de 4), et de même $g_n(P, Q) = g_n(r)$. (3) donne alors :

$$\varphi(C, C) = \int\int_A \frac{1-\varphi(E, E)}{2\delta}[(1-k)^2 g^2(EC) + \ldots + (1-k)^{2n+2} g_n^2(EC) + \ldots]dS_E$$

équation intégrale dont la résolution par approximations successives donne $\varphi(C, C) = C^{te} = \varphi_0$, et on a nécessairement :

$$\varphi_0 = H\frac{1-\varphi_0}{2\delta},$$

d'où :

$$(3') \qquad \varphi_0 = \frac{H}{2\delta + H}$$

en posant :

$$H = \int\int_A [(1-k)^2 g^2(r) + \ldots + (1-k)^{2n+2} g_n^2(r) + \ldots]dS$$

telle est la valeur du *coefficient de corrélation entre 2 loci infiniment voisins.*

Alors l'équation (2) montre que φ (C, D) ne dépend que de la distance CD : soit $\varphi = \varphi(CD)$:

$$(4) \quad \left\{ \varphi(CD) = \frac{1-\varphi_0}{2\delta}\int\int_A [(1-k)^2 g(EC)g(ED) + \ldots + (1-k)^{2n+2} g_n(EC)g_n(ED) + \ldots]dS_E \right.$$

tel est le *coefficient de corrélation entre 2 loci dont la distance est* CD. On peut algébriser les « produits de composition » qui figurent dans (4), en considérant les transformées de Fourier :

$$F(u, v) = \int\int e^{iux+ivy} g\left(\sqrt{x^2+y^2}\right)dxdy$$

et :

$$K(u, v) = \int\int e^{iux+ivy} \varphi\left(\sqrt{x^2+y^2}\right)dxdy,$$

car on a :

$$K(u, v) = \frac{1-\varphi_0}{2\delta}\left[\sum_{p=1}^{\infty}(1-k)^{2p}F^{2p}\right] = \frac{1-\varphi_0}{2\delta}(1-k)^2F^2/[1-(1-k)^2F^2]$$

[formule d'ailleurs immédiate si on applique directement la transformation de Fourier à (1)].

K est ainsi exprimé en fonction de $F(u, v)$ qui est connu, d'où par inversion de la transformation de Fourier à 2 variables :

$$(5) \qquad \varphi = \frac{1-\varphi_0}{8\pi^2\delta}\int\int e^{-iux-ivy}\frac{(1-k)^2F^2}{1-(1-k)^2F^2}\,du\,dv$$

(en faisant $x = y = 0$ on retrouve l'équation linéaire pour φ_0).

On peut pousser le calcul jusqu'au bout en admettant que le déplacement de chaque individu est une *promenade au hasard* suivant le schéma de Polya, c'est-à-dire que la loi $f(r)dS = g(r)dS$ est une loi de Bravais isotrope $G(\sigma^2)dS = (1/2\pi\sigma^2)e^{-r^2/2\sigma^2}dS$, ce qui donne :

$$F(u,v) = e^{-(\sigma^2/2)(u^2+v^2)}, \qquad K(u,v) = \frac{1-\varphi_0}{2\delta}\sum_{p=1}^{\infty}(1-k)^{2p}e^{-(2p\sigma^2/2)(u^2+v^2)}$$

et on en déduit la série (4), plus facile à calculer que (5), d'où :

$$(4') \qquad \varphi(r) = \frac{1-\varphi_0}{2\delta}\sum_{p=1}^{\infty}(1-k)^{2p}G(2p\sigma^2)$$

(cette série est bien uniformément convergente si $k > 0$ puisque

$$G(2p\sigma^2) \leqslant 1/4\pi p\sigma^2).$$

La formule (3) se retrouve en faisant $r = 0$, ce qui amène à calculer :

$$H = \sum_{p=1}^{\infty}(1-k)^{2p}[1/4\pi p\sigma^2] = [1/4\pi\sigma^2]\int_0^{(1-k)^2}\left[\sum_{p=1}^{\infty}x^{p-1}\right]dx$$
$$= -\log[1-(1-k)^2]/4\pi\sigma^2 = -\log(2k-k^2)/4\pi\sigma^2$$

d'où :

$$(3') \qquad \varphi_0 = 1/[1 - 8\pi\sigma^2\delta/\log(2k-k^2)].$$

φ_0 se calcule donc aisément à partir de la pression (d'hétérosis ou de mutation) k et du nombre $\pi\sigma^2\delta$ d'individus dans un cercle de rayon σ (cercle où restent en moyenne 40 % des individus nés en son centre) ; il est d'autant plus voisin de 1 (homogénéité locale) que ces 2 quantités sont plus faibles ; on déduit ensuite de (4') :

$$(4'') \qquad \frac{\varphi(r)}{\varphi_0} = \frac{\displaystyle\sum_{p=1}^{\infty}(1-k)^{2p}[1/4\pi p\sigma^2]e^{-r^2/4p\sigma^2}}{\displaystyle\sum_{p=1}^{\infty}(1-k)^{2p}[1/4\pi p\sigma^2]}$$

ce qui montre que le *rapport de la corrélation à la distance* r *à la corrélation à la distance 0 décroît de 1 à 0 quand* r *croît de 0 à* ∞.

La valeur numérique de ce rapport ne dépend que des 2 quantités k et r/σ, il est donc facile d'en dresser des tables (1), ce qui permettra d'interpréter à l'aide de cette formule les résultats expérimentaux. Cette formule est indépendante de δ, ce qui permet de la conserver dans une population dont la densité varie considérablement au cours des années (vagues de vitalité de Tschetverikov).

Si les individus manifestent une tendance à rester groupés en « colonies » ou « essaims » on en tiendra compte en admettant que chaque individu a des probabilités respectives α et $(1-\alpha)$ d'effectuer un déplacement infiniment petit de fluctuation ε^2 ou une migration à distance de fluctuation σ^2, c'est-à-dire en prenant $g(r) = \alpha G(\varepsilon^2) + (1-\alpha)G(\sigma^2)$, ce qui donne $F(u, v) = \alpha e^{-(\varepsilon^2/2)(u^2+v^2)} + (1-\alpha)e^{-(\sigma^2/2)(u^2+v^2)}$, d'où une formule pour $\varphi(r)$.

La détermination expérimentale de la corrélation en fonction de la distance, en vue de la vérification de la théorie, peut se faire de plusieurs façons.

a) Si on mesure, en un grand nombre de points P_i d'un territoire étendu, la fréquence q_i d'un gène mendélien sans sélection géographique, on prendra pour estimation de c la moyenne générale de ces fréquences, et pour estimation de $\varphi(r)$ la moyenne de toutes les quantités $\dfrac{(q_i - c)(q_j - c)}{c(1-c)}$ calculés à partir de 2 points P_i et P_j dont la distance est r (2).

b) Si on mesure, sur différents individus, un caractère biométrique neutre au point de vue de la sélection, dont l'intensité puisse être

(1) Pour calculer numériquement (4″), on peut développer le numérateur suivant les puissances de r^2, faisant apparaître les séries $\sum_{p=1}^{\infty}(1-k)^{2p}/p^n$, dont la somme est $\int_0^{(1-k)^2} \dfrac{\{\log[(1-k^2)/X]\}^{n-1}}{(n-1)!}\dfrac{dX}{1-X}$; on déduit d'ailleurs de ceci que le numérateur de (4″) est égal à :

$$(-1/4\pi\sigma^2)\int_0^{\log[1-(1-k)^2]} J_0\left[\frac{r}{\sigma}\sqrt{\log\frac{(1-k)^2}{1-e^u}}\right]du,$$

J_0 étant la fonction de Bessel. En faisant $r = 0$, on retrouve bien le dénominateur H.

(2) Si k tend vers 0 le numérateur et le dénominateur de (4″) tendent vers l'infini, leur différence reste finie (d'après les propriétés de J_0), donc $H \to \infty$, φ_0 et $\varphi \to 1$; la population tend vers l'homogénéité complète, résultat inévitable dans toute population à effectif fini en l'absence de mutations.

regardée comme résultant de l'addition des effets d'un certain nombre d'aléatoires mendéliennes indépendantes X de moyennes M_i (chacune prenant des valeurs s_i et t_i avec des probabilités q_i et p_i fonctions du lieu), la corrélation moyenne entre 2 individus I et I' situés à une distance r sera une estimation de :

$$\frac{\partial\mathfrak{R}[\Sigma(X_i - M_i) \times \Sigma(X'_i - M_i)]}{\partial\mathfrak{R}[\Sigma(X_i - M_i)]^2} = \frac{\Sigma\partial\mathfrak{R}[(X_i - M_i)(X'_i - M_i)]}{\Sigma\partial\mathfrak{R}(X_i - M_i)^2}$$

c'est-à-dire de $\varphi'r$) si on admet que le coefficient de mutation k est le même pour tous les gènes qui interviennent. Mais il faudra tenir compte que la corrélation est affaiblie si une fraction de la variabilité n'est pas génétique.

Autres applications.

1° La panmixie dans une population finie de N individus peut s'étudier en supposant l'aire A occupée égale à l'unité, δ égal à N, et $g(P, Q)$ égal à 1 si P et Q sont dans A et à 0 en dehors. On a alors, pour C et D dans H :

$$\varphi(C, D) = \varphi(C, C) = \frac{1 - \varphi(E, E)}{2\delta}[(1-k)^2 + \ldots + (1-k)^{2n} + \ldots]$$

ce qui donne :

$$\varphi = \text{constante} = \frac{(1-k)^2}{2N[1-(1-k)^2] + (1-k)^2}$$

en faisant $k = u$ et $v = 0$ (pas de mutations de retour), on retrouve approximativement la valeur d'équilibre du coefficient f obtenue page 35. En égalant cette expression à φ_0 donné par (3'), ou (3''), on obtient l'effectif N d'un groupe panmictique équivalent à un groupe d'aire très petite faisant partie d'une population à migration isotrope au hasard ; soit :

$$N[(1-k)^{-2} - 1] = \frac{\delta}{H} = \frac{4\pi\sigma^2\delta}{-\log(2k - k^2)}.$$

Mais cette notion de N équivalent introduite par S. Wright (17) par un raisonnement tout à fait différent, n'a pas la portée qu'il lui attribue, puisqu'elle ne peut rendre compte de la corrélation à distance.

2° On peut essayer de donner un schéma de migration homogène mais anisotrope (dans une population illimitée de densité constante) en supposant que le déplacement d'un individu résulte de 2 déplacements indépendants de lois de probabilité différentes, dans 2 directions rectangulaires, c'est-à-dire que $g(P, Q)dSP = m(x_1 - x_2)n(y_1 - y_2)dx_1dy_1$

en désignant par x_1, y_1, x_2, y_2, les coordonnées de P et Q (m et n étant 2 fonctions d'une variable, dont l'intégrale de $-\infty$ à $+\infty$ est égale à 1) ; la formule (2) devient en posant :

$$m_{p+1}(x_1 - x_2) = \int_{-\infty}^{+\infty} m_p(x_1 - x_3) m(x_3 - x_2) dx_3 \text{ (et de même pour } n)$$

$$\varphi(P, Q) = \int\int \frac{1 - \varphi(E, E)}{2\delta} \left[\sum_{p=0}^{\infty} (1-k)^{2p+2} m_p(x_3 - x_1) m_p(x_3 - x_2) \right.$$

$$\left. n_p(y_3 - y_1) n_p(y_3 - y_2) \right] dx_3 dy_3.$$

En particulier si on prend pour $g(P, Q)$ la loi de Bravais non isotrope réduite :

$$g(P, Q) = \frac{1}{2\pi\sigma\rho} e^{-\frac{(x_2-x_1)^2}{2\sigma^2} - \frac{(y_2-y_1)^2}{2\rho^2}}$$

on trouve $\varphi(E, E) = C^{te} = \varphi_0$ et on a :

$$\varphi(P, Q) = \frac{1 - \varphi_0}{4\pi\delta\sigma\rho} \sum_{p=1}^{\infty} \frac{(1-k)^{2p}}{2p} e^{-\frac{(x_2-x_1)^2}{2p\sigma^2} - \frac{(y_2-y_1)^2}{2p\rho^2}}$$

(d'où φ_0 en faisant $x_2 = x_1$ et $y_2 = y_1$) ; on en déduit la variation de φ à abscisse constante et à ordonnée constante. On pourrait introduire un schéma analogue avec 3 dimensions pour représenter la variabilité d'une population aquatique suivant les 2 coordonnées de surface et la profondeur.

BIBLIOGRAPHIE

1. Bernstein. — *Z. indukt. Abstamm.*, **56**, 1930, p. 223.
2. L. Blaringhem. — *Principes et formules de l'hérédité mendélienne* (Gauthier-Villars, 1928).
   L. Blaringhem, P. Bertrand, P. Guérin et Th. J Stomps. — *Hérédité, Mutation et Evolution.* L'œuvre de Hugo de Vries au Palais de la Découverte (Masson et Cie, 1937).
3. R. A. Fisher. — *Transactions of the R. S. of Edimburgh*, **52**, 1918 p. 399.
4. R. A. Fisher. — *The Genetical Theory of Natural Selection.* Oxford, Clarendon Press, 1930.
   Ph. L'Héritier. — *Génétique et Evolution.* Actualités Hermann, n° 158, Paris.
5. F. Galton. — *Natural Inheritance* (London, 1889).
6. Haldane. — *Annals of Eugenics*, **9**, 4.

7. Haldane et Philip. — *J. of Genetics*. **36**, p. 197.
8. Hogben. — *Proc. R. S. of Edimburgh*, **53**, p. 239 ; *J. of Genetics*, **26**, p. 417.
W. Johannsen. *Elemente der exackten Erblichkeitslehre*, Jena, 1909, et 2e édition revue, 1913.
9. Malécot. — *Annales de l'Université de Lyon*, A **2**, 1940, p. 25 ; *C. R. des Séances de la Soc. Math.*, 1938, p. 44.
10. Malécot. — *C. R. de l'Acad. des Sciences*, **206**, p. 153 et 404 ; **208**, p. 407 et 552.
11. Malécot. — *Thèse Sc.*, Paris (Imp. Guilhot, 52, Bd Malesherbes), 1939.
12. Malécot. — *C. R. des Séances de l'Acad. des Sc* , **215**, p. 313 ; *Annales de l'Université de Lyon*, Sciences, 1941, p. 45.
13. Malécot. — *C. R. de l'Ac. des Sc.*, **219**, p. 379.
14. Malécot. — *C. R. de l'Ac. des Sc.*, **221**, p 340 ; **222**, p. 841.
15. K. Pearson. — *Phil. Trans.*, **203** A, 1903, p. 53 ; *Proc. R. S. of London*. B **81**, 1909. p. 225.
Snow. — *Proc. R. S. of London*, B **83**, 1911, p. 37.
16. G. Teissier. — *Revue scientifique*, 82e année, fasc. 3, p. 145.
17. S. Wright. — *Genetics*, 1931, *passim* ; *Genetics*, 1946, p. 39 ; *Stat. Genetics in Relation to Evolution*. Paris, Actualités Hermann, 802.

TABLE DES MATIÈRES