Ann. Génét. Sél. anim., 1973, **5** (3), 333-361.

GÉNÉTIQUE DES POPULATIONS DIPLOÏDES NATURELLES DANS LE CAS D'UN SEUL LOCUS

III. — Parenté, mutations et migration

G. MALÉCOT

Mathématiques appliquées,
Université de Lyon I, BP 37
69 Villeurbanne Charpennes

RÉSUMÉ

La biologie moléculaire met en évidence de nombreux polymorphismes dus à la « structure fine » des protéines, et donc du segment d'ADN qui constitue le gène correspondant ; toute substitution — même non décelable actuellement — d'un acide aminé à un autre est une mutation, de probabilité extrêmement faible pour chaque codon, et chaque telle « mutation moléculaire » produit un nouveau gène, un nouvel allèle, en général neutre (sur le plan de la sélection) par rapport à l'ancien. Le multiallélisme est donc fréquent même dans une population limitée, et chaque allèle n'y figure que de façon transitoire, jusqu'à ce que de nouveaux allèles s'y substituent. La notion de « fréquence d'équilibre » introduite dans la partie I (chap. II) perd toute signification à long terme. Par contre, le coefficient de parenté défini dans la partie II (chap. I) conserve la même définition : probabilité pour que les deux loci tirés soient « identiques », c'est-à-dire dérivent sans aucune mutation d'un même locus d'un ancêtre commun ; la différence est maintenant que deux loci « non identiques » ne sont plus « indépendants en probabilité » (comme dans la partie II où chaque « bêta-mutation » était un « passage à l'indépendance ») mais sont différents (puisque chaque « mutation moléculaire » introduit un nouveau codon, ou un nouvel allèle, non encore représenté).

Les calculs de la partie II, dans le cas de migration homogène, sont ici répétés par de nouvelles méthodes qui décomposent le coefficient de parenté φ en somme de contributions des chaînes de parenté des divers ordres (§ V et VII). La parenté entre générations différentes est étudiée au § IX. Le § X établit les expressions asymptotiques pour de grands intervalles d'espace et de temps, et justifie alors l'introduction de la « distance génétique » $D = -\text{Log}\,\varphi$ par le fait que sa partie principale est — toujours dans le cas de migration homogène, avec taux de mutation constant — proportionnelle à la distance géographique et à la distance dans le temps.

Le § XI, (B) suggère deux méthodes d'études de la migration non homogène : en la ramenant par une transformation géométrique à une migration homogène ; ou bien en dressant une nouvelle carte géographique où les distances mutuelles sont proportionnelles aux logarithmes des coefficients de parenté déterminés expérimentalement.

I. — INTRODUCTION

La génétique humaine met en évidence de très nombreux polymorphismes, en découvrant que bien des cistrons conditionnant les protéines que l'on détecte dans les liquides organiques sont susceptibles chacun de nombreux états alléliques, différant les uns des autres le plus souvent par des substitutions au niveau de quelques nucléotides (ces substitutions se retrouvant, dans la plupart des cas, au niveau des acides aminés de la protéine : l'exemple de l'hémoglobine est célèbre).

Par ailleurs, la sélection en faveur de certaines combinaisons de protéines (donc de certaines combinaisons de gènes) est un phénomène macroscopique qui (la lenteur de l'évolution horotélique le montre, cf. KIMURA, 1962) ne joue qu'assez rarement par rapport à la vitesse de l'évolution moléculaire, et elle n'introduit des changements profonds (« révolution génétique », MAYR, 1963) qu'à partir de petits groupes d'individus assez isolés du reste de l'espèce (« effets de fondation » dans des colonies assez isolées). Il est donc permis de supposer que la plupart des substitutions produites par mutation dans la molécule d'ADN sont, en tout cas au début, « neutres », c'est-à-dire sans valeur sélective : ce n'est alors que le hasard, la « dérive génétique », qui décide de leur extinction ou au contraire de leur extension à une grande partie de l'espèce. Le hasard joue d'autant plus que l'espèce est plus fragmentée en colonies n'ayant que de faibles échanges naturels par migration (1).

Ces remarques incitent à attribuer quelque importance à notre modèle (voir partie II, MALÉCOT, 1972) de « mutations neutres » survenant, à chaque locus et dans chaque génération, avec des probabilités données, dans une population fragmentée en « colonies » définies chacune par un nombre d'individus ou « effectif » N et par un « site » géographique x de coordonnées données (dans une aire assimilée à un plan, ou bien, à titre de simplification mathématique, sur une droite, un cercle, etc.).

II. — RAPPELS SUR LA MIGRATION

Puisque la migration joue un rôle explicatif essentiel les calculs devront faire intervenir les « taux de migration », c'est-à-dire (rappelons-le) les probabilités l_{xz} pour qu'un gamète tiré au hasard parmi ceux (« gamètes utiles ») qui donnent naissance à des individus dans le site x, soit produit par un géniteur né lui-même dans le site z.

Pour un site de reproduction donné (x donné) les l_{xz} doivent être définis pour tous les sites d'origine z (et parmi eux, pour le site x lui-même : l_{xx} est la probabilité

(1) Si la migration était nulle, il y aurait subdivision en espèces différentes. S. WRIGHT (1965) a souligné que des taux de migration modérés entre groupes peu nombreux sont la circonstance la plus favorable à une différenciation locale sans rupture du lien spécifique, donc à l'extension ultérieure de toute combinaison locale qui se révélerait avantageuse pour l'ensemble de l'espèce, ou en tout cas pour un domaine écologique étendu (c'est là un aspect de la « préadaptation » de CUÉNOT). Il faut noter qu'au contraire, si le brassage est trop grand entre tous les individus de l'espèce du fait d'une très grande mobilité, la sélection joue en général un rôle stabilisateur, en écartant les variants extrêmes.

pour que le gamète soit « autochtone ») ; leur somme sur tous les sites z possibles est une probabilité égale à 1, ce que nous noterons $\sum_z l_{xz} = 1$; il est sous-entendu que s'il s'agit de migration dans le plan (ou même dans R^n, avec d'évidentes modifications de notations), le signe $\sum_z$ désigne une sommation sur l'ensemble des couples de valeurs des coordonnées z_1 et z_2 du « point » z (si l'on veut envisager une migration continue dans le plan ou l'espace, il suffit d'entendre cette somme comme une intégrale multiple).

Note mathématique

Il est commode de définir dès maintenant la « fonction génératrice de migration » [1] :

$$L(\alpha) = \sum_z l_{xz}\, \alpha^{z-x} \qquad (1)$$

Dans un problème « unidimensionnel » (droite, cercle, etc.) α désigne un complexe de module 1 appartenant au « cercle trigonométrique » C du plan complexe, ce qui assure que la somme ou série qui figure dans (1) est uniformément convergente sur C.

Dans un problème plan, il faut introduire deux complexes indépendants de modules 1, α_1 et α_2, appartenant respectivement à deux « cercles trigonométriques C_1 et C_2. α^z sera alors une notation abrégée pour le produit $\alpha_1^{z_1-x_1}\alpha_2^{z_2-x_2}$, z_1 et z_2 (respectivement x_1 et x_2) étant les deux coordonnées du point z (respectivement x). Mais nous continuerons à désigner par $L(\alpha)$ le résultat de la sommation sur z_1 et z_2 qui figure dans le deuxième membre de (1), (résultat qui est indépendant de x_1 et x_2 dans le cas de « migration homogène », voir § III B).

Si l'on adopte un schéma de « migration continue », il est préférable de définir l'intégrale multiple correspondant au deuxième membre de (1) à partir du changement de variables $\alpha_1 = e^{iv_1}$, $\alpha_2 = e^{iv_2}$ (MALÉCOT, 1948, 1967 *b*, 1969) : l'intégration multiple fournit alors la « transformée de Fourier » de la distribution des l_{xz}, qui doit être regardée comme une fonction $\overline{L}(v_1, v_2)$ des réels quelconques v_1 et v_2.

Dans tous les cas l' « inversion » (passage de L ou $\overline{L}$ aux l_{xz}) s'effectue par la « formule de Fourier réciproque » : le « cas discret » et le « cas continu » ne diffèrent que par les intervalles d'intégration par rapport à v_1 et v_2, intervalles qui sont chacun $[-\pi, +\pi]$ dans le cas discret, et $[-\infty, +\infty]$ dans le cas continu.

Toutes les formules ainsi esquissées s'appliqueront aussi à toutes les fonctions génératrices qui seront définies ultérieurement (§ VIII).

III. — LA MATRICE DE MIGRATION

a) *Dans le cas unidimensionnel discret,* l'ensemble des l_{xz} peut être regroupé en une « matrice » L [2] dont les lignes sont indexées par le « site d'arrivée » x et les colonnes par le « site de départ » z. C'est une « matrice stochastique » puisque $l_{xz} \geqslant 0$ et $\sum_z l_{xz} = 1$

Le module maximum 1 des valeurs propres n'est réalisé que pour la valeur propre 1 (exclusion des « modèles oscillants ») si l'on suppose que tous les éléments diagonaux l_{xx} sont > 0 (c'est-à-dire que, en tout site x, les gamètes utiles « autochtones » ont une probabilité non nulle). En outre, la valeur propre 1 sera simple si l'on suppose que la matrice L est indécomposable, c'est-à-dire que la migration [3] de chaque site z vers chaque site x a une probabilité non nulle (sans quoi il y aurait partage de la popu-

[1] En général fonction de x, sauf dans le cas de migration homogène (voir III-B).
[2] Finie dans le cas du cercle, infinie dans le cas de la droite illimitée.
[3] Sur un certain nombre de générations.

lation totale en plusieurs populations sans communication, qu'il suffirait alors d'étudier séparément).

b) *Cas de la migration homogène.* — Lorsque l_{xz} ne dépend que de la « distance » $z - x$, L est une matrice « de Toeplitz », ou « cyclique » : ses valeurs propres se calculent alors aisément (dans le cas fini, avec r sites), en remarquant que les composantes des vecteurs propres sont les puissances des racines d'ordre r de l'unité. Les combinaisons linéaires qui résultent de la diagonalisation deviennent, quand $r \to \infty$, des intégrales effectuées le long du cercle trigonométrique C (appendice I).

c) *Dans le cas de la migration sur un intervalle borné de la droite,* il est permis de supposer que l_{xz} ne dépend que de $(x - z)$ *sauf à proximité des bornes* ; dans les cas simples qui se ramènent à la marche au hasard sur un segment avec absorption aux bornes, il est encore facile de diagonaliser (voir appendice I). En fait, le principal problème de diagonalisation porte sur la matrice symétrique $L\tilde{L}$, précédemment introduite par nous (MALÉCOT, 1950, 1951).

d) Il est en général aisé d'étendre aux problèmes bidimensionnels de migration discrète, sauf lorsque les intégrales d'inversion portent sur des fonctions ayant des points critiques et non plus des pôles. Mais on peut souvent alors se ramener à des intégrales elliptiques (MALÉCOT, 1950 ; KIMURA et WEISS, 1964 ; MALÉCOT, 1971). Cette étude sera rappelée plus loin.

e) La principale critique faite aux matrices de migration porte sur leur variabilité dans le temps qui, dans bien des espèces animales, est une conséquence nécessaire des changements géographiques ou écologiques. Mais il serait aisé de développer, à partir des équations de récurrence linéaire qui seront écrites plus loin, les conséquences d'une variation des l_{xz} suivant des fonctions analytiques simples (fonctions sinusoïdales, qui font apparaître des vibrations forcées, asymptotiquement sinusoïdales ; fonctions polynomiales, qui font apparaître une tendance séculaire) ; on peut aussi introduire pour les l_{xz} une variation aléatoire au cours du temps (et éventuellement dans l'espace) avec des espérances fixes et des variances fixes (1).

f) Je ne vois à ces extensions qu'une difficulté majeure : tenir compte du cas où le « déplacement moyen » déduit des l_{xz} varie (par moyenne sur les z, et éventuellement par moyenne sur le temps) en grandeur avec le point d'arrivée x ; ce cas se rencontre pratiquement dans le cas de « déséquilibre démographique » entre les sites (« peuplement convergent » de « places vides », ou émigration divergente à partir de zones surpeuplées) ; les modèles alors nécessaires doivent comporter de nombreuses hypothèses de caractère démographique.

IV. — LA MUTATION

Les précisions, apportées par la Génétique humaine en particulier, sur la nature biochimique des mutations et sur leur détection, permettent (KIMURA et CROW, 1964, 1968 *a*) de modifier les anciens modèles (2) en supposant que chaque locus (chaque

(1) C'est ce qui a été fait, mais dans le cas de la sélection seulement, par KIMURA (1954) pour le temps, et par nous-même (MALÉCOT, 1959, 1960, 1965 *b*) pour l'espace.

(2) Suivant les indications données en partie II, p. 387, note (3).

« cistron », est potentiellement occupable par de très nombreux allèles différant les uns des autres (cf. Introduction) par quelques nucléotides seulement, ces différences apparaissant par des « mutations moléculaires » qui transforment « au hasard » un des nucléotides de l'ADN d'un individu (chaque transformation consistant à passer d'une des bases A, T, G, C à une autre) ; comme il y a approximativement 500 à 1 000 nucléotides par cistron, il est permis de regarder comme très improbable (même sur les 10^6 cistrons de l'Homme, et même sur une durée de 5×10^8 années comme celle de l'hémoglobine des Vertébrés) que des mutations moléculaires affectant deux individus différents conduisent au même cistron, pourvu que les cistrons soient assez finement analysés (directement, ou par les enzymes qu'ils produisent). De ce point de vue, chaque mutation survenant dans un cistron donné produit un allèle nouveau (CAVALLI-SFORZA et BODMER, 1971). Nous avions suggéré (MALÉCOT, 1951) comme commodité mathématique seulement, de regarder chaque mutation comme donnant naissance à un allèle nouveau. Il n'est donc plus question de parler d'état d'équilibre, même statistique, entre les fréquences respectives d'allèles déterminés : chaque allèle introduit par une mutation peut, au cours du temps, s'éteindre, fluctuer, ou s'étendre provisoirement à toute la population (mais seulement jusqu'à apparition par mutation de nouveaux allèles). On ne peut plus alors parler (KIMURA et CROW, 1964, 1968 *b*) que d'un flux mouvant d'allèles supposés neutres (voir Introduction), sans qu'il y ait pratiquement de réapparition d'un allèle précédemment disparu.

Tous les calculs que nous avons développés sur la parenté sont alors valables à condition de modifier la définition du coefficient de mutation k que nous avions introduit dans la partie II [1] ; plusieurs cas seront à distinguer, suivant que l'on pourra analyser complètement les nucléotides du cistron, ou les acides aminés de la chaîne polypeptidique (comme c'est le cas pour l'hémoglobine) ou seulement globalement un certain nombre d'états alléliques du cistron ou de la protéine : à chacun de ces niveaux d'observation, nous désignerons par k la probabilité pour que, dans le génome d'un individu quelconque, à partir d'une nucléotide donnée, ou à partir d'un état allélique donné d'un cistron donné, une mutation se produise, conduisant à un état différent de tous les autres états déjà existants ; cette notion de « mutaiton non récurrente » est logiquement différente de celle de mutation vers un état indépendant en probabilité de l'état de départ, qui implique une possibilité de retour à cet état. Mais, s'il y a beaucoup d'allèles, la différence entre les deux cas est faible [2].

Il sera commode de supposer en outre que k est indépendant de l'allèle de départ, ou, ce qui revient approximativement au même, de prendre pour k une moyenne calculée sur les divers états de départ possibles.

Lorsqu'il s'agit des nombreuses nucléotides d'un même cistron (par exemple celui qui engendre une des deux chaînes α ou β de l'hémoglobine), il est permis actuellement de regarder k comme approximativement constant d'une nucléotide à une autre : c'est ce que suggèrent les études de ZUCKERKANDL et PAULING (1965) et les données de DAYHOFF et ECK 1968), (sur les chaînes de l'hémoglobine, de la Lamproie à l'Homme

[1] Dans le cas de 2 allèles seulement avec « mutations récurrentes », k est le taux fictif de « bêta-mutation » somme des deux taux de mutation réels réciproques u et v. Dans l'état stationnaire d'équilibre statistique, k est la probabilité de passer, dans le locus considéré et quel que soit le gène A ou a qui l'occupe, à un gène (a ou A) *indépendant* en probabilité.

[2] D'après un résultat de KIMURA (1967) les résultats asymptotiques sont les mêmes lorsque le nombre d'allèles observés simultanément est grand par rapport à $1 + 4\,N\!k$.

la comparaison point par point des quelques 140 sites d'acides aminés (rappelons que chacun correspond à un « codon », « triplet » de 3 nucléotides consécutives) est en bon accord avec l'hypothèse que *chaque* nucléotide aurait présenté (pendant $5 \cdot 10^8$ années!) une probabilité de mutation $k_1 \# 1/2\ 10^{-9}$ *par an.*

Étant donné que la probabilité de survenue de *l'un ou l'autre* d'événements (même compatibles) est sensiblement la somme de leurs probabilités individuelles lorsque leur survenue simultanée a une probabilité très petite par rapport à celles-ci, la probabilité de mutation dans un acide aminé déterminé serait sensiblement $3k_1$, s'il n'existait des mutations de nucléotides « synonymes » c'est-à-dire fournissant le même acide aminé, ce qui réduit $3k_1$ d'environ 30 p. 100 (KIMURA, 1967) ; la probabilité de mutation d'un acide aminé est sensiblement $k_2 = 2k_1 \# 10^{-9}$ par an.

Si l'on considère un cistron formé par exemple de 1 500 nucléotides, donc de 500 amino-acides, et en admettant que toutes les mutations d'amino-acides y soient détectables et définissant donc autant d'allèles différents, le taux de mutation d'un cistron donné vers n'importe lequel de ses allèles sera sensiblement $k = 500 \times k_2 = 1\,000\, k_1 \# 1/2\ 10^{-6}$. Ce taux doit être mutiplié par la longueur d'une génération pour obtenir l'habituel taux de mutation par locus par génération, ce qui concorde avec l'ordre de grandeur donné par CAVALLI-SFORZA et BODMER (1971). Cela n'est pas en désaccord avec les taux de mutations plus grands observés pour certaines mutations à effets phénotypiques, car ce sont les plus fréquentes qu'on a le plus tendance à observer.

Remarque : KIMURA (1968, 1971), utilisant un résultat que nous avons établi en 1948 (cf aussi p. 32 et 34 de la traduction anglaise, MALÉCOT, 1969) : la probabilité de la fixation finale d'un allèle neutre est égale à sa fréquence initiale, en conclut que, dans une population de N individus comprenant donc 2N cistrons (ou 2N acides aminés) homologues, la fréquence initiale d'une mutation qui apparaît dans un des cistrons (ou dans un des acides aminés), soit 1/2 N, est aussi sa probabilité de fixation finale sur l'ensemble des N individus.

Dans l'état stationnaire du flux de mutations, le nombre total par locus de mutations fixées, sur l'ensemble des N individus, dans tout le génome (ou dans tous les acides aminés d'une protéine) est le produit du nombre total d'emplacements nouvellement affectés par unité de temps, 2Nk (ou $2Nk_2$), par la probabilité de fixation finale 1/2 N : le taux de fixation finale ou « taux de substitution », est donc égal au taux d'apparition ; c'est ce théorème de KIMURA qui lui permet de calculer, à partir des taux de fixation comparés dans l'hémoglobine des vertébrés, les taux d'apparition k_2 et k_1 des mutations d'acides aminés et de nucléotides. J'indiquerai plus loin, par une voie toute différente cette relation entre taux d'apparition et taux de fixation (NEI, 1971).

V. — LES CHAINES DE PARENTÉ

L'analyse de parenté que nous avons établie à partir de 1942 peut être plus clairement exposée en introduisant les chaînes de parenté gamétiques (plus commodes que les chaînes de parenté zygotiques que, pour comparer aux statistiques sur les cousinages, nous avions introduites en 1950 et 1967 *b*).

Soient deux individus I et J appartenant à deux sites donnés x et w, dans la génération F_n. Tirons au hasard, chez I et chez J, deux locus homologues ; autrement dit tirons au hasard, pour chacun, un des deux gamètes respectifs qui leur ont donné naissance et dont les origines aléatoires seront nommées z et u (fig. 1) ; il peut arriver qu'ils apportent le même locus d'un parent (père ou mère) commun, ce qui ne peut se produire que si z et u coïncident, si de plus on tire en ce site le même parent (événement de probabilité $\frac{1}{N}$ si les N individus dans ce site sont des parents équiprobables) et si de plus on tire le même locus parmi les 2 loci homologues de ce parent $\left(\text{événement de probabilité } \frac{1}{2}\right)$.

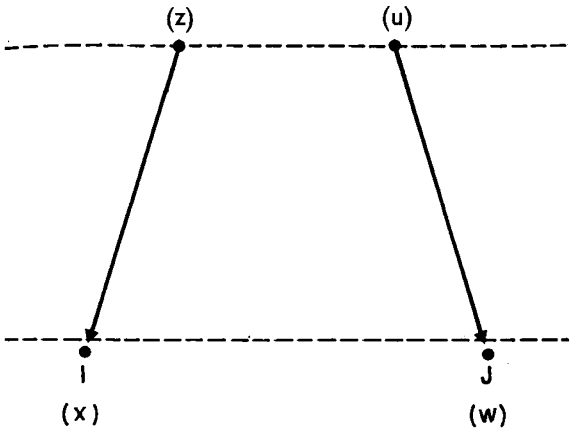

FIG. 1

La probabilité pour que I et J soient unis par une « chaîne de parenté gamétique d'ordre 2 » est donc :

$$\pi_1(x, w) = \sum_z l_{xz}\, l_{wz}/2N \qquad (1)$$

En remplaçant $\frac{1}{2N}$ par $\frac{1}{N}$ on obtiendrait la probabilité d'une « chaîne zygotique d'ordre 2 » (probabilité d'être demi-frères). Mais cela introduirait des complications inutiles.

En toute généralité, (fig. 2) nous appellerons *probabilité d'une parenté d'ordre 2p*, soit $\pi_p(x, w)$, la probabilité pour que, en remontant, à partir des sites x et w de F_n, les ascendances de 2 locus (de 2 gamètes), on leur trouve pour origine un locus commun pour la première fois dans la génération F_{n-p}, chez un « ancêtre commun d'ordre p » en remontant ce que nous appellerons une « chaîne de parenté gamétique d'ordre $2p$ » réunion de deux « chaînes d'ascendance d'ordre p ».

Les probabilités des chaînes *des divers ordres* $p \in N^+$ sont celles d'événements *s'excluant* mutuellement, ce qui nous permettra plus loin de les ajouter.

Il est aisé d'obtenir pour les π_p une récurrence sur p : en effet une chaîne d'ordre $2p > 2$ est une chaîne d'ordre $2p - 2$ reliant les loci *supposés distincts* des parents P

et Q dont I et J ont tiré respectivement les 2 gamètes considérés. Donc, de la probabilité conjointe $l_{xz}l_{wu}$ de provenance de ces gamètes, il faut retrancher les probabilités sommées dans la formule (1), puis multiplier par la probabilité que P et Q soient unis

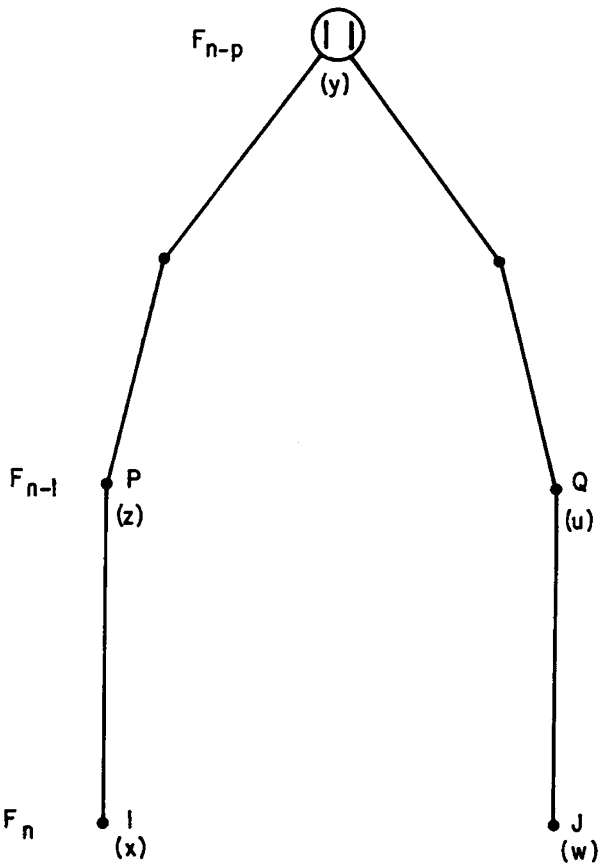

FIG. 2

par une chaîne d'ordre $2p - 2$, puis sommer sur tous les couples (z, u) de sites possibles pour P et Q dans F_{n-1}, soit :

$$\pi_p(x, w) = \sum_z \sum_{u \neq z} l_{xz}l_{wu}\ \pi_{p-1}(z, u) + \sum_z l_{xz}l_{wu}\left(1 - \frac{1}{2N}\right)\pi_{p-1}(z, z)$$

$$\pi_p(x, w) = \sum_z \sum_u l_{xz}l_{wu}\ \pi_{p-1}(z, u) - \sum_z l_{xz}l_{wz}\,\pi_{p-1}(z, z)/2N \qquad (2)$$

(On vérifie aisément que le terme soustractif est plus petit que le terme additif, lui-même par récurrence < 1 ; d'ailleurs nous savons déjà que la somme $\sum_{p=1}^{\infty} \pi_p$ doit être comprise entre 0 et 1 puisque c'est encore une probabilité totale). Quelques exemples numériques m'ont montré que cette probabilité, lorsque N est grand et la

migration faible, décroît très lentement quand p augmente : par exemple, s'il n'y a aucune migration : $l_{xz} = \delta_{xz}$ de KRONECKER, on a :

$$\pi_p = \left(1 - \frac{1}{2N}\right)\pi_{p-1} = \left(1 - \frac{1}{2N}\right)^{p-1}\pi_1 = \left(1 - \frac{1}{2N}\right)^{p-1}/2N$$

Une méthode de formulation de la solution générale de (2) consiste à introduire les $l_{xy}^{(p)} = \sum_z l_{xz} l_{zy}^{(p-1)}$ (avec $l^{(1)} = l$) qui sont, pour le mathématicien, des « produits de composition de Volterra », et pour le démographe, les probabilités pour qu'un gamète tiré dans le site x provienne d'un ancêtre d'ordre p occupant le site y (dans le cas particulier où l_{xz} ne dépend que de la distance vectorielle $z - x$, cas de ce que nous appellerons la « migration homogène », cette loi de composition est une convolution, $l_{xy}^{(p)}$ ne dépend que de $y - x$, et une transformation de Fourier résout complètement les problèmes, comme nous le verrons).

VI. — GÉNÉRALISATIONS

Les formules (1) et (2) sont susceptibles d'innombrables généralisations, qui lèvent une partie des objections pratiques faites à l'usage des matrices de migration.

1. Si la « matrice de migration » définie par les l_{xz} varie au cours des générations suivant une formulation simple (« oscillatoire » ou « séculaire », ou « aléatoire »), la récurrence (2) se résout par des produits de composition à facteurs fonction du temps.

2. Si les taux de migration l_{xz} sont différents pour les ovules (ou pour les femelles) et pour les gamètes mâles, on montre aisément que, s'il s'agit d'un locus mendélien autosomal, la récurrence (2) subsiste à condition de prendre pour l_{xz} la *moyenne arithmétique* des taux de migration correspondants pour les mâles et pour les femelles (chaque gamète tiré au sort parmi les deux qui donnent naissance à un individu donné a une chance sur deux d'être mâle ou femelle) ; par contre, la formule (1) correspond aux demi-frères qui, dans le cas « dioïque » (sexes séparés) ont le même père (avec probabilité $\frac{1}{N_1}$, si N_1 est le nombre de mâles) ou la même mère (avec probabilité $\frac{1}{N_2}$, si N_2 est le nombre de femelles). Si les taux de migration des gamètes mâles et femelles sont distincts, soit l'_{xz} et l''_{xz}, (1) doit être remplacé par

$$4\,\pi_1(x,w) = \sum_z l'_{xz} l'_{wz}/2N_1 + \sum_z l''_{xz} l''_{wz}/2N_2 \qquad (1')$$

D'ailleurs le N de la formule (2) doit aussi être remplacé par $1/\left[\frac{1}{4N_1} + \frac{1}{4N_2}\right]$, double de la « moyenne harmonique », comme S. WRIGHT (1931) l'a montré depuis longtemps.

3. Les formules s'étendent immédiatement au chromosome Y, porté par les mâles seulement ; il suffit de préciser que l_{zz} est le taux de migration des gamètes mâles et de remplacer dans (1) et (2) $1/2N$ par $1/N_1$ (probabilité pour que deux gamètes mâles issus d'un même site z aient tiré chez le même père, leur chromosome Y). On a remarqué, depuis DARWIN (1875), que les noms de familles « de type occidental » s'héritent comme le chromosome Y. Ce qui explique l'usage de plus en plus répandu, chez les statisticiens américains qui étudient les noms de famille, de la théorie des chaînes de parenté. Le chromosome X donnerait lieu à des formules faciles à établir (les deux sexes interviendraient).

4. Les formules sont applicables, bien au-delà de la « génétique mendélienne » à toute « génétique particulaire » dans laquelle des plasmagènes, chloroplastes, etc. seraient transmis (en dehors de rares modifications par mutations) par les ovules, donc par les femelles. Il suffit de prendre pour l_{xx} le taux de migration des ovules ou des femelles, et de remplacer 2N par le nombre N_2 de celles-ci (avec une correction qui reste à préciser, dans le cas où chaque ovule comporterait plusieurs plasmagènes non identiques entre eux).

5. L'effectif N (ou les effectifs mâles et femelles N_1 et N_2) n'interviennent dans les formules (1) et (2) que par la probabilité $1/N$ pour que deux gamètes issus du même site y proviennent du même individu alors qu'il y en a N : cela suppose implicitement que le tirage chez un individu d'un gamète utile ne modifie pas sa probabiblité de produire d'autres gamètes utiles, donc que sa loi de fécondité (en gamètes utiles) est poissonnienne ; en fait il arrive souvent que le nombre d'enfants par individu ait une variance V supérieure à sa moyenne $\bar{k}$ (ce dont une loi de Pascal rend compte). Nous avons montré (1951 p. 82) qu'il suffit de remplacer dans les formules, N par le « nombre effectif » N_e défini par

$$1/N_e = \frac{\bar{k}(\bar{k}-1) + V}{N\,\bar{k}^2}$$

Cette généralisation d'un calcul initial de S. WRIGHT (1938) a été retrouvée, après MALÉCOT (1950, 1951) par un grand nombre d'auteurs, en particulier CROW et KIMURA (1963) qui ont certes apporté au dénominateur une correction soustractive de l'ordre de $\bar{k}$.

Beaucoup d'autres articles ont été écrits sur le nombre effectif N_e et ses conséquences statistiques. Chez l'Homme, il semble que N_e/N soit voisin de 0,7 ou 0,8 (CROW, 1954) si N est le nombre d'adultes reproducteurs.

6. Les calculs aboutissant à (1) et (2) ont supposé que, lorsque les deux gamètes ne provenaient pas du même locus mais des deux locus homologues d'un même individu, la parenté de ceux-ci était la même que s'ils provenaient de deux individus différents de même site ; or il arrive souvent que les deux locus homologues d'un même individu, et les deux gamètes qui leur ont appartenu, aient de plus grande probabilité de parenté du fait que les deux parents de cet individu ne sont souvent choisis en raison de liens de parenté proches : c'est la « consanguinité préférentielle dans les croisements naturels » qui crée une distorsion par rapport à la « loi de croisement au hasard » ou « panmixie » par exemple en écartant l'inceste, en favorisant les cousins, etc.

Nous avons montré (MALÉCOT, 1951, 1971) qu'il suffit alors de remplacer N par $N/(1+\alpha)$, α désignant le coefficient de corrélation entre les gamètes qui s'unissent, par rapport à l'ensemble des gamètes (« réservoir gamétique ») théoriquement utilisables dans le site x [1]. En particulier l' « homogamie gamétique totale » correspond à $\alpha = 1$, c'est-à-dire à une division par 2 de l'effectif réel. Toutes les autres corrections pour homogamie ou consanguinité préférentielle sont moindres [2].

7. La formule (2) peut bien entendu être écrite en supposant que N soit une fonction (oscillatoire, monotone, ou aléatoire) de l'emplacement z et de numéro $p-1$ de la génération ; c'est évidemment la variation de $\frac{1}{N}$ qui définit la variation des derniers coefficients de la récurrence linéaire (2). Cet effet de « moyenne harmonique » avait déjà été indiqué, dans de nombreux cas particuliers, par S. WRIGHT (1931).

8. Notre modèle suppose une évolution par générations séparées (c'est le cas, par exemple, des plantes annuelles, ou des animaux à éclosion annuelle).

Nous avons traité ailleurs (MALÉCOT, 1965 *a*, 1965 *b*) le cas d'une « reproduction continue », les 2N loci étant remplacés « un par un » dans la population, suivant une loi d' « attente exponentielle » ; ce qui revient à supposer que la loi de fécondité de chaque individu est une loi d'attente exponentielle de la première naissance à venir (sans « mémoire » des naissances passées).

Nous avons montré que les résultats asymptotiques restent inchangés (à condition de prendre pour « durée d'une génération » le temps moyen nécessaire pour que les 2N loci soient tous remplacés).

Étant donné que les modèles de « reproduction continue » ayant un intérêt pratique sont intermédiaires entre ce dernier modèle et le modèle à générations séparées, on doit s'attendre à une grande généralité des résultats que nous allons établir ci-après pour des générations séparées. Les calculs pourraient d'ailleurs être faits en remplaçant la loi exponentielle de fécondité par une loi Γ plus générale ou « loi de χ^2 », suivant la suggestion faite par KENDALL (1952) pour la « durée de génération » d'une bactérie.

Plus précisément, on supposera que la probabilité pour que deux individus I et J, nés aux dates 0 et $\tau \geqslant 0$ dans les sites x et w aient des mères (respectivement des pères) nés dans le même emplacement z et dans le même intervalle de temps petit $[-u, -u+\Delta u]$, est :

$$l_{xz}l_{wz}\, b(u)\, b(u+\tau)\, (\Delta u)^2$$

$b(u)$ étant défini à partir des probabilités de survie et de fécondité à l'âge u.

Si l'on désigne par $N\,\Delta u$ le nombre de femelles (respectivement de mâles) qui naissent dans le site z pendant l'intervalle de temps Δu et si on les regarde comme des mères (respectivement des pères) équi probables, la probabilité pour que I et J soient unis par une chaîne de parenté maternelle (respectivement paternelle) d'ordre $1+1$ sera :

$$\pi_1(x, w, \tau) = \sum_z l_{xz}l_{wz} \int_0^{+\infty} b(u)\, b(u+\tau)\, du/N$$

(1) Cf. S. WRIGHT (1965) et MALÉCOT (1969).
(2) MALÉCOT (1951).

L'intégrale pourrait d'ailleurs être remplacée par une somme si l'on ne distinguait qu'un nombre fini de classes d'âge et les « cohortes » correspondantes.

Le calcul numérique peut être fait en adoptant pour $b(u)$ une loi Γ, ou pour loi discrète (par classes d'âges) la loi de Pascal (équivalente à la loi continue Γ). (Voir appendice II.)

VII. — LES COEFFICIENTS DE PARENTÉ

a) A titre d'introduction, nous définirons d'abord le « coefficient de parenté large » de deux individus quelconques I et J de sites fixés x et w, comme la probabilité totale que deux loci homologues tirés au sort chez l'un et chez l'autre, ou — ce qui est équivalent — la probabilité totale pour que les deux gamètes « apportant » deux locus dans les deux sites x et w, les tiennent d'un même locus d'un ancêtre commun d'ordre p quelconque ; la formule des probabilités totales fournit immédiatement

$$\sum_{p=1}^{p=+\infty} \pi_p(x, w)$$

comme valeur du « coefficient de parenté large » ; cette série, par construction, converge vers une limite $\leqslant 1$; si on la tronque (en la réduisant à ses p_0 premiers termes), on diminue beaucoup la valeur numérique : or, si l'on ne dispose que des statistiques de mariages entre cousins jusqu'au 6e degré inclus, on ne peut remonter que jusqu'à $p_0 = 3$; nous verrons plus loin que ce n'est que dans le cas de fortes mutations et migrations que la « série tronquée » peut fournir une approximation valable.

Par contre, si l'on remonte indéfiniment, jusqu'à « p infini », la somme est égale à 1 dans le cas où le nombre total d'individus dans tous les sites communiquant effectivement par migration reste borné supérieurement.

Nous avons démontré (MALÉCOT, 1948 et p. 39 de la traduction anglaise) dans le cas d'un site complètement isolé comprenant $N(n)$ individus dans la génération F_n, que le coefficient de parenté large tend vers 1 quand $n \to +\infty$ à condition que $N(n)$ ne croisse pas plus qu'une fonction linéaire de n. Sinon, la limite est < 1. Ce raisonnement rejoint celui fait par S. WRIGHT (1921) et JACQUARD (1970) pour les « systèmes de croisements réguliers » : le nombre total d'individus à utiliser reste fini dans le cas de croisements systématiques entre « doubles cousins », croît linéairement entre « simples cousins germains », et augmente plus que linéairement dans le cas de « cousins issus de germains », où la limite est < 1 (1/53, JACQUARD, 1970).

Quelle est la signification d'un coefficient de parenté égal à 1 ? Il correspond à la presque certitude que deux individus I et J tirés au hasard aient au moins un ancêtre commun : ce qui est, d'évidence même, assuré si on remonte indéfiniment en arrière dans une population d'effectif total borné, où le nombre d'ancêtres distincts d'ordre p ne peut augmenter indéfiniment.

b) Beaucoup plus significatif pour l'évolution à long terme est le « coefficient de parenté (strict) » que nous avons défini dans la partie II, p. 386 comme la

probabilité(1) de descendre *sans aucune mutation intermédiaire* d'un même locus d'un ancêtre commun (ou probabilité d'« identité » de 2 locus homologues).

Comme dans la partie II complétée par les remarques ci-dessus (§ IV) nous noterons k la probabilité, sur chaque chaînon parent-enfant, d'une « mutation vers un état allélique indépendant » ou bien « vers un état allélique différent » et $1 - k$ la probabilité de « maintien de l'identité de l'état allélique » (2).

La probabilité de maintien constant de l'identité le long de p chaînons (c'est-à-dire le long d' « d'une chaîne d'ascendance » d'ordre p) étant $(1 - k)^p$, *le coefficient de parenté* (strict) entre deux individus I et J de sites x et w dans F_n est :

$$\varphi(x, w) = \sum_{p=1}^{\infty} (1 - k)^{2p} \pi_p(x, w) \tag{3}$$

série évidemment convergente ($0 \leqslant k \leqslant 1$) et de somme d'autant plus petite que k est plus grand. C'est là la décomposition annoncé dans la partie II, p. 397.

Nous avons souvent considéré le « coefficient de parenté tronqué »

$$\varphi_n(x, w) = \sum_{p=1}^{p=n} (1 - k)^{2p} \pi_p(x, w) \tag{4}$$

qui ne décompte que les chaînes remontant de F_n à une génération initiale » F_0 (plus ou moins arbitraire). Cette valeur serait correcte si tous les individus de F_0 n'étaient unis entre eux par aucune chaîne de parenté (par exemple s'ils étaient prélevés au hasard dans une population panmictique infinie : ce pourrait être le cas pour certains élevages de drosophiles, quoique, en toute rigueur, les drosophiles sont toutes apparentées, comme le sont d'ailleurs tous les individus d'une même espèce).

Il est évident que tenir compte complètement de tous les liens de parenté existant dans F_0 revient à « remonter jusqu'à l'infini [formule (3)] ; mais on peut remarquer que, si n est suffisamment grand pour que $(1 - k)^{2n}/k$ soit négligeable par rapport à φ_n, φ_n coïncide sensiblement avec φ, l'effet de la « distribution initiale de parenté dans F_0 » est effacé, L'évolution de φ_n vers sa limite φ est certes lente si k est petit (3) : le calcul à partir de F_0 montre clairement cette lenteur qui est camouflée quand on remonte à l'infini, c'est-à-dire quand on se place d'emblée dans l'état stationnaire, comme le font parfois les économétriciens.

Nous avons, dans la partie II, établi la récurrence (sur n et sur les sites) qui régit φ_n (cette récurrence est liée à celle de π_n [formule (2)] puisque $(1 - k)^{2n}\pi_n$ est d'après (4) la « différence première de » φ_n) : il y a probabilité $l_{xz}l_{wz}/2N$ pour que les deux gamètes tirés dans les sites x et w apportent le même locus d'un même individu situé en z dans la génération précédente F_{n-1} ; auquel cas le coefficient de parenté strict est $(1 - k)^2$; dans tous les autres, il est égal à

$$(1 - k)^2 \varphi_{n-1}(z, u)$$

(1) Pour les deux loci homologues.

(2) Avec des modifications évidentes s'il s'agit, non d'un cistron, mais d'une nucléotide ou d'un codon : k est alors remplacé par k_1 ou k_2.

(3) MALÉCOT (1948, 1954, 1972) ; MARUYAMA (1970).

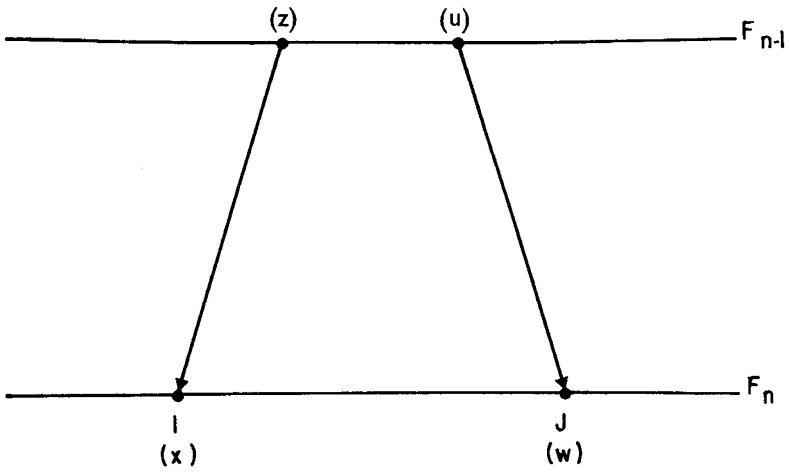

FIG. 3

D'où [cf., avec d'autres notations, partie II, p. 395, formule (1.1')] :

$$\varphi_n(x, w) = (1 - k)^2 \sum_z \sum_u l_{xz}\, l_{wu}\, \varphi_{n-1}(z, u) + (1 - k)^2 \sum_z l_{xz}\, l_{wz} \frac{1 - \varphi_{n-1}(z, z)}{2N} \qquad (5)$$

les l_{xz} et N étant relatifs à la génération F_n et aux sites x et z. Toutes les variations et corrections signalées au VI restent bien entendu applicables.

La limite φ de φ_n est la solution de l'équation linéaire (6) déduite de (5) en remplaçant φ_n et φ_{n-1} par φ :

$$\varphi(x, w) = (1 - k)^2 \left[\sum_z \sum_u l_{xz} l_{wu}\, \varphi(z, u) + \sum_z l_{xz} l_{wz} \frac{1 - \varphi(z, z)}{2N} \right]. \qquad (6)$$

La solution est unique, car l'opérateur linéaire figurant (à un terme additif près indépendant de φ) dans le deuxième membre de (5) a une norme < 1, (sauf dans le cas très particulier où à la fois $k = 0$ et $N = \infty$).

Les formules (5) et (6) peuvent d'ailleurs être résolues explicitement par récurrence, en utilisant les produits de composition $l_{xy}^{(p)}$ définis au § V, et en remontant jusqu'à F_0 dans le cas de (5) et jusqu'à l'infini dans le cas de (6) MALÉCOT, 1948, 1969), ce qui donne dans ce dernier cas :

$$\varphi(x, w) = \sum_{p=1}^{+\infty} \sum_z (1 - k)^{2p}\, l_{xz}^{(p)}\, l_{wz}^{(p)} \frac{1 - \varphi(z, z)}{2N} \ (1) \qquad (7)$$

Remarque : Dans les travaux sur la consanguinité (où le pionnier fut S. WRIGHT (1921), on traite séparément le problème des chaînes de parenté unissant deux individus I et J qui sont des « conjoints », c'est-à-dire sont effectivement les deux parents d'un individu K de la génération suivante : S. WRIGHT et MALÉCOT ont appelé « coefficient de consanguinité », f_K, de l'individu K, le coefficient de parenté de ses parents I et J [au facteur $(1 - k)^2$ près] ; mais il ne se déduit de $\varphi(x, w)$ que si ses

(1) Partie II, p. 395, formule (1.2).

parents I et J sont tirés au sort en x et w (« panmixie ») ; quand il est différent par suite de croisements préférentiels, il est aisé à calculer. Mais nous avons vu au VI que tous les effets statistiques se résument dans le remplacement de N par $N/(1+\alpha)$ et sont donc faibles [1] : les « structures de parenté » chères aux ethnologues, c'est-à-dire les règles restrictives concernant les mariages, ont très peu d'effets statistiques ; au surplus CROW et MENGE (1965), ALLEN (1965), MORTON (1966, 1971) ont montré, dans le cas de communautés isolées comme dans le cas de clans, que α est faible.

VIII. — CAS DE LA MIGRATION HOMOGÈNE : TRANSFORMÉES DE FOURIER

La résolution en termes finis de (5), (6), (7) devient immédiate si l'on suppose que la migration soit homogène, c'est-à-dire que l_{xz} ne soit fonction que de la distance $z - x = y$, soit $l_{xz} = l(y)$, et que l'effectif N soit indépendant sur site z (et du numéro de la génération). Introduisons alors la « fonction génératrice de de migration », indépendante de x, définie par la formule (1) (§ II).

$$L(\alpha) = \sum_y l(y)\, \alpha^y \tag{8}$$

[$|\alpha| = 1$ pour 1 dimension, $|\alpha_1| = |\alpha_2| = 1$ pour 2 dimensions]

Les produits de composition figurant dans (7) étant, maintenant que la migration est homogène, des convolutions, $l_{zx}^{(p)}$ ne dépend que de $z - x$, et sa fonction génératrice est $L^p(\alpha)$: en prenant pour $\varphi(z, z)$ une constante que nous noterons en bref φ_0, (7) fournit pour $\varphi(x, w)$ une fonction de la seule distance vectorielle $w - x$ et on vérifie aisément que cette fonction là vérifie (6), *dont elle est donc la solution unique*, que nous noterons dorénavant $\varphi(w - x) = \varphi(d)$ en désignant par d la distance vectorielle $w - x$.

La transformée de Fourier éventuelle $K(\alpha)$ de cette fonction $\varphi(d)$ s'obtient aisément (MALÉCOT, 1948, 1948 *b*, 1972) en multipliant les deux membres de (7) par $\alpha^{w-x} = \alpha^{z-x}\,(1/\alpha)^{z-w}$, puis sommant $\frac{1}{2}$ à w en 1er lieu et $\frac{1}{2}$ à z en 2e lieu (en vertu de la convergence uniforme pour $|\alpha| = 1$). On obtient ainsi :

$$K(\alpha) = \sum_w \Sigma\, \varphi(w - x)\alpha^{w-x} = \sum_{p=1}^{\infty} (1 - k)^{2p}\, L^p(\alpha)\, L^p(1/\alpha)\, (1 - \varphi_0)/2N \tag{9}$$

Série dont la convergence est assurée, puisque $|L(\alpha)| \leqslant \sum_z l_{xz} = 1$, à condition que $k > 0$.

Pour étudier les chaînes de parenté des divers ordres, on peut traiter de façon analogue des équations (1), (2), (3), (4) ; on peut même y introduire la « transformée de Laplace par rapport au temps ».

$$P(w - x, \beta) = \sum_{p=1}^{\infty} \beta^p\, \pi_p(x, w) \tag{10}$$

[1] Pour le calcul de f_K à partir de φ et de α, voir S. WRIGHT (1943), 1965) et MALÉCOT (1969 *b*).

qui converge uniformément [en vertu de (3) où l'on remplace $(1 - k)^2$ par β] lorsque $|\beta| \leqslant 1$.

(4) donne à la limite :

$$\varphi(w - x) = P[w - x, (1 - k)^2] \tag{11}$$

et la résolution de (2) à partir des convolutions $l_{xz}^{(p)}$ montre que

$$P(w - x, \beta) = \sum_{p=1}^{\infty} (1 - k)^{2p} \beta^p \sum_y l_{wy}^{(p)} l_{xy}^{(p)} [1 - P(0, \beta]/2N$$

Or la connaissance de la loi de migration permet de calculer

$$C(w - x, \beta) = \sum_{p=1}^{\infty} (1 - k)^{2p} \beta^p \sum_y l_{wy}^{(p)} l_{xy}^{(p)} \tag{12}$$

qui est (puisque la somme $\sum_y$ est une probabilité) une série entière absolument convergente quand $|\beta| < 1$.

Et l'on en déduit alors :

$$P(z, \beta) = \frac{C(z, \beta)}{2N + C(0, \beta)} \tag{13}$$

L'étude des points singuliers (qui sont forcément les zéros du dénominateur) permet (MALÉCOT, 1973) d'établir le rayon de convergence de la série entière (10) et d'en déduire la (faible) vitesse de tendance vers zéro, quand $n \to \infty$, des probabilités π_n.

Le calcul numérique des probabilités de parenté π_n pour des lois de migration particulières est donné dans l'appendice III. Nous allons pour le moment expliciter les conséquences pratiques de la formule (9), en remarquant :

1° qu'elle fait apparaître la transformée du coefficient de parenté dans F_n comme somme des contributions des générations F_{n-p} précédentes [ce qui était d'ailleurs déjà le cas dans (7)] ; il est envisageable de corriger la formule en supposant que la matrice de migration varie avec l'ordre p.

2° que la sommation dans (9) de la série géométrique donne (pour $k > 0$)

$$K(\alpha) = \frac{1 - \varphi_0}{2N} \frac{(1 - k)^2 L(\alpha) L(1/\alpha)}{1 - (1 - k)^2 L(\alpha) L(1/\alpha)} \tag{14}$$

[formule de partie II, p. 396, *in fine*] (1).

D'après la définition de $K(\alpha)$ dans la formule (9), $\varphi(w - x) = \varphi(d)$, qui définit le coefficient de parenté en fonction de la distance d, se calcule par « inversion de Fourier » (voir § II, note mathématique).

A. — *Dans le cas unidimensionnel discret sur la droite* (fig. 3)

On a :

$$\varphi(d) = \frac{1}{2\pi i} \int_C \alpha^{d-1} K(\alpha)\, d\alpha \tag{15}$$

Remarque : Cette intégrale devra être remplacée, dans le cas discret sur le cercle, par une somme sur toutes les racines $r^{\text{ième}}$ de l'unité (MALÉCOT, 1973).

(1) Il est intéressant de remarquer que, selon nous (MALÉCOT, 1973) le deuxième facteur du deuxième membre de (14) est aussi en remplaçant $(1 - k)^2$ par $\beta(1 - k)^2]$ la fonction génératrice de la fonction (12).

Et, dans le cas continu unidimensionnel ou bidimensionnel, pour une intégrale (simple ou double) de Fourier (avec $\alpha = e^{iv}$, cf. § II, note mathématique, et MALÉCOT, 1967 *b*).

Dans le cas fréquent où $L(\alpha)$ [défini par (8)] se réduit à un polynôme non entier, et par suite $K(\alpha)$ à une fonction rationnelle, l'intégration (15) nécessite le calcul des résidus relatif aux seuls pôles (en nombre fini) intérieurs au cercle C de rayon 1. D'ailleurs $\varphi(-d) = \varphi(d)$ en raison de la symétrie du coefficient de parenté. Il suffit donc de faire le calcul pour $d \geqslant 0$; les seuls pôles [1] sont donc les racines de l'équation :

$$L(\alpha)\, L(1/\alpha) = 1/(1 - k)^2 \tag{16}$$

qui joue un rôle clef dans tous les problèmes.

D'ailleurs k, taux de mutation, doit dans tous les cas être supposé très petit ; étant donné que, d'après (8), $L(1) = 1$, l'équation (16) admet (MALÉCOT, 1969) deux racines inverses infiniment voisines de 1, soient α_1 et $\alpha_2 = 1/\alpha_1$; les développements limités, en fonction de $\alpha - 1$, de $L(\alpha)$ et de $L(1/\alpha)$ donnent [1969, p. 82] :

$$L(\alpha)\, L(1/\alpha) = 1 + \sigma^2(\alpha - 1)^2 + o[(\alpha - 1)^3\sigma^3] \tag{17}$$

σ^2 désignant la variance de la migration sur la droite, variance définie par

$$\sigma^2 = \sum_y y^2 l(y) - \left[\sum_y y l(y)\right]^2 \text{ [2]} \tag{18}$$

le reste dépendant des cumulants d'ordre $\geqslant 3$, comme on le voit en développant le le Log de $L(e^{iv})\, L(e^{-iv})$

(16) et (17) montrent alors que α_1 et $\alpha_2 = 1/\alpha_1$ sont réels, et la racine $\alpha_1 < 1$ peut être écrite $\alpha_1 = 1 - u$, avec $u \sim \sqrt{2k}/\sigma$.

Nous avons montré (MALÉCOT, 1965 *b*) que, dans le cas où $l(1)$ ou $l(-1)$ est > 0, l'équation (16) ne peut avoir d'autres racines de modules voisins de 1 que $\alpha_1 = 1 - u$ et $\alpha_2 = 1/\alpha_1$; par ailleurs, (MALÉCOT, 1969) dans le cas général où les autres racines sont loin d'un extrémum de $L(\alpha)\, L(1/\alpha)$, alors que α_1 et α_2 en sont voisins, le résidu de ces dernières racines est beaucoup plus grand que les autres.

B. — *Dans le cas bidimensionnel discret*

L'intégrale (15) est (cf. note mathématique) une intégrale double étendue aux 2 cercles C_1 et C_2 définis par $|\alpha_1| = 1$ et $|\alpha_2| = 1$. La première inversion (par exemple par rapport à α_1) fait apparaître les pôles ci-dessus indiqués. Mais la deuxième nécessite le calcul d'intégrales elliptiques dans le cas particulier de « migration entre groupes adjacents » étudié par nous (MALÉCOT, 1950, 1971) [3]. Le cas général peut être approximativement ramené à ce cas particulier si l'on regarde la loi de migration dans deux dimensions comme suffisamment définie par la « matrice de covariance » définie (dans le cas centré, pour simplifier les notations) par les $\sum_y l(y) y_i y_j$; on peut alors, par rotation des axes orthonormés, diagonaliser cette

(1) A l'exception du pôle $\alpha = 0$ dans le cas où $d = 0$.

(2) Il est intéressant de constater que l'espérance mathématique de la migration disparaît ; une « translation moyenne » subie également par les migrants issus de tous les sites ne change rien au résultat ; il n'en est plus de même si la translation moyenne dépend du site, ce qui entraîne les difficultés signalées en III-F).

(3) Ce modèle a été plus tard appelé « stepping stone model » par KIMURA, qui a étendu l'étude à 3 dimensions (WEISS et KIMURA, 1965).

matrice en $\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, et par rapport aux nouveaux axes, le développement (17) se généralise en :

$$L(\alpha)\, L(1/\alpha) = 1 + \sigma_1^2(\alpha_1 - 1)^2 + \sigma_2^2(\alpha_2 - 1)^2 + 0_3 \qquad (19)$$

Le modèle particulier que nous avons traité en Partie II avec taux de migration « immédiatement à gauche et à droite » égaux à m et taux « immédiatement en haut et en bas » égaux à m', correspond à $\sigma_1^2 = 2m$ et $\sigma_2^2 = 2m'$.

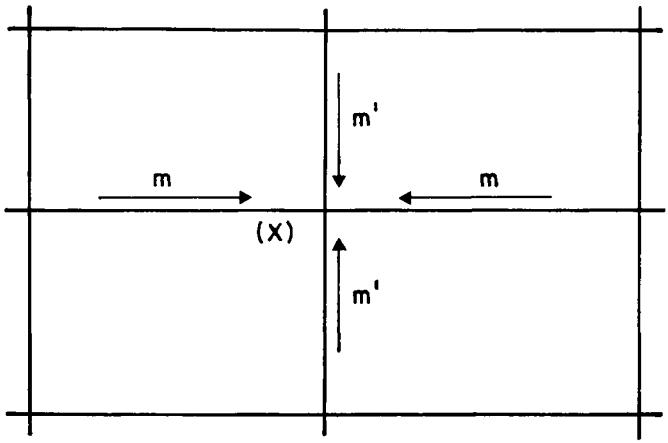

FIG. 4

IX. — PARENTÉ DÉCALÉE DE τ GÉNÉRATIONS

On peut aisément étendre les formules (5), (7), (9), au calcul du coefficient de parenté $\varphi(w - y, \tau)$ entre un individu K situé dans le site y de la génération $F_{n+\tau}$ et un individu J situé dans le site w de la génération F_n.

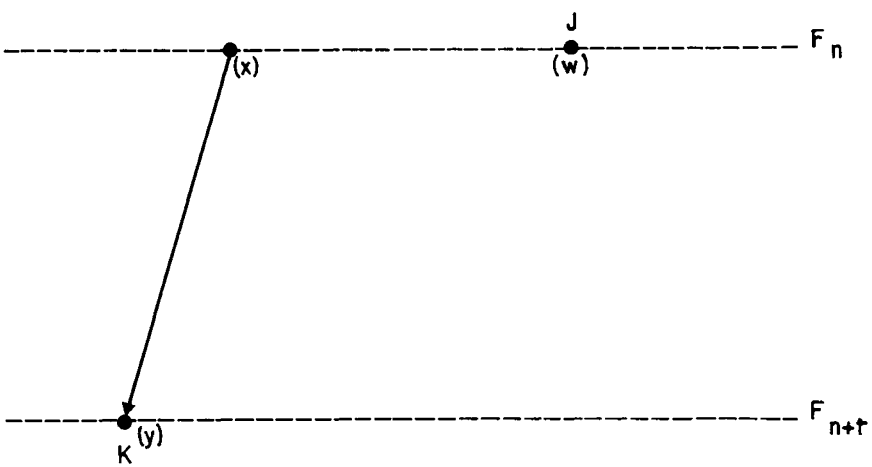

FIG. 5

Un locus tiré chez K a la probabilité $(1-k)^{\tau} l_{yx}^{(\tau)}$ de provenir sans mutation du site x de F_n ; dans le cas où x et w coïncident, ce locus a la probabilité $1/2N$ d'être identique au locus tiré chez J ; hormis ce cas, la probabilité d'identité est $\varphi(w-x)$. Donc

$$\varphi(w-y,\tau) = (1-k)^{\tau}\left[\sum_x l_{yx}^{(\tau)}\,\varphi(w-x) + l_{yw}^{(\tau)}\,(1-\varphi_0)/2N\right]$$

d'où :

$$\sum_d \alpha^d\,\varphi(d,\tau) = L^{\tau}(\alpha)\,[K(\alpha) + (1-\varphi_0)/2N]\,(1-k)^{\tau}$$

d'où, compte tenu de (14) :

$$\varphi(d,\tau) = \frac{1-\varphi_0}{2N}\,\frac{1}{2i\pi}\int_C \frac{\alpha^{d-1}\,L^{\tau}(\alpha)\,(1-k)^{\tau}\,d\alpha}{1-(1-k)^2\,L(\alpha)\,L(1/\alpha)} \tag{20}$$

X. — EXPRESSIONS ASYMPTOTIQUES ET « DISTANCE GÉNÉTIQUE » (DANS LE CAS DE MIGRATION HOMOGÈNE CONSTANTE)

A. — *Variation de la parenté avec la distance dans le cas unidimensionnel*

(14) et (15) donnent :

$$\varphi(d) = \frac{1-\varphi_0}{2N}\,\frac{1}{2i\pi}\int_C \alpha^{d-1}\left[-1 + \frac{1}{1-(1-k)^2\,L(\alpha)\,L(1/\alpha)}\right]d\alpha \tag{21}$$

il résulte de VIII — (A) que, *quand k est petit,* le quotient par $2i\pi$ de l'intégrale a pour partie principale le résidu relatif au pôle $\alpha_1 = 1-u = 1-\sqrt{2k}/\sigma$; or l'équation (17) montre que la dérivée par rapport à α, pour $\alpha = \alpha_1$, de $L(\alpha)\,L(1/\alpha)$ a pour partie principale $2\sigma^2(\alpha_1 - 1) = -2\sigma^2 u = -2\sigma\sqrt{2k}$; d'où

$$\varphi(d) \sim \frac{1-\varphi_0}{2N}\,\frac{\alpha_1^d}{2\sigma\sqrt{2k}}$$

(cf. partie II, p. 397).

D'où en particulier :

$$\varphi_0 = \varphi(0) \sim \frac{1}{1+4N\sigma\sqrt{2k}} \tag{22}$$

$$\varphi(d) \sim \varphi(0)\left(1-\frac{\sqrt{2k}}{\sigma}\right)^d \sim \varphi(0)\,e^{-\sqrt{2k}\frac{d}{\sigma}} \tag{23}$$

La formule (23) nous incite à appeler *distance génétique* la quantité

$$D = -\operatorname{Log}\varphi(d) = -\operatorname{Log}\varphi(0) + \sqrt{2k}\,\frac{d}{\sigma} \tag{24}$$

D augmente proportionnellement à $\frac{d}{\sigma}$, « distance normée par la migration », le coefficient de proportionalité étant $\sqrt{2k}$ que nous pouvons appeler « taux de différenciation géographique pour le locus considéré ». Si l'on effectue des mesures

de $\varphi(d)$, donc de D, sur différents locus, les résultats doivent être proportionnels, D étant d'autant plus grand que le taux de mutation est plus grand.

Il est remarquable que l'effectif N dans chaque site n'intervienne par ailleurs que dans le terme constant de (24).

Remarque : Ces résultats simples découlent bien sûr de l'hypothèse de « migration homogène ».

Nous verrons au § XI comment tenir compte d'une migration non homogène ou du passage d'un petit groupe à l'état d'isolement complet (« effet de fondation », MAYR, 1942).

B. — *Variation de la parenté avec la distance dans le cas bidimensionnel*

La formule d'inversion (21) reste applicable avec les modifications indiquées dans la note mathématique : remplacer $\frac{1}{2\pi i}$ par $\left(\frac{1}{2\pi i}\right)^2$, et l'intégrale simple par une intégrale double le long des cercles $C_1(|\alpha_1| = 1)$ et $C_2(|\alpha_2| = 1)$, avec $\alpha^{d-1} = \alpha_1^{d_1-1}\alpha_2^{d_2-2}$, etc.

Dans le cas « symétrique » où $L(1/\alpha) = L(\alpha)$, on pourrait ((MALÉCOT, 1971) se ramener à l'intégrale « prépondérante »

$$\int_{C_2}\int_{C_1} \frac{\alpha^{d-1}}{1-(1-k)L(\alpha)}\, d\alpha$$

Mais on peut traiter le cas général en utilisant le développement (19) dans le voisinage des valeurs $\alpha_1 = 1$ et $\alpha_2 = 1$ qui (sur les cercles C_1 et C_2) correspondent à un maximum de $L(\alpha)L(1/\alpha)$ donc à un maximum de l'intégrant de (21).

Nous nous bornerons au cas où $d_1 = 0$, et nous poserons $d_2 = p$; il s'agit de calculer en premier lieu l'intégrale

$$\frac{1}{2i\pi}\int_{C_1} \frac{\alpha_1^{-1}\, d\alpha_1}{1-(1-k)^2\,[1+\sigma_1^2(\alpha_1-1)^2+\sigma_2^2(\alpha_2-1)^2+o_3]} \qquad (25)$$

qui admet un pôle p_1 de module < 1, et voisin de 1, défini (cf. VIII, 16) par :

$$1+\sigma_1^2(p_1-1)^2+\sigma_2^2(\alpha_2-1)^2+o_3 = 1/(1-k)^2$$

d'où :

$$\sigma_1^2(p_1-1)^2 = 2k-\sigma_2^2(\alpha_2-1)^2+o\,[(\alpha_2-1)^3\,\sigma_2^3]$$

et le résidu r_1 qui fournit la partie principale de (25) est lui-même (cf. A ci-dessus) équivalent à

$$\sim \frac{1}{-2\,\sigma_1^2(p_1-1)} \sim \frac{1}{2\,\sigma_1\sqrt{2k-\sigma_2^2(\alpha_2-1)^2+o\,[(\alpha_2-1)^3\,\sigma_2^3]}}$$

L'intégration $\frac{1}{2i\pi}\int_{C_2} \alpha_2^{p-1}\, r_1\, d\alpha_2$ fournit alors, d'après (21), la partie principale $\varphi(d)$, que nous noterons dorénavant $\varphi(0, p)$ puisque le vecteur d est présentement le vecteur de composantes $d_1 = 0$ et $d_2 = p$:

$$\varphi(0,p) \sim \frac{1-\varphi_0}{4N\pi i}\int_{C_2} \frac{\alpha_2^{r-1}\, d\alpha_2}{2\,\sigma_1\sqrt{2k-\sigma_2^2(\alpha_2-1)^2+o\,[(\alpha_2-1)^3\,\sigma^3]}}$$

Lorsque $\sqrt{2k}$ est suffisamment petit, le calcul peut être poursuivi sans avoir à pousser plus loin le développement limité sous le radical, en remarquant que l'intégrant est maximum et grand au voisinage de $\alpha_2 = 1$ en raison de la proximité du point critique $C_2 \sim 1 - \frac{\sqrt{2k}}{\sigma_2}$. La partie principale ne dépend que de la dérivée seconde du radical par rapport à $\alpha_2 - 1 = u$, et on ne la modifie pas en y remplaçant

$$(\alpha_2 - 1)^2 \text{ par l'équivalent } \frac{(\alpha_2 - 1)^2}{\alpha_2} = \alpha_2 + \frac{\alpha_2}{1} - 2 \ (^1)$$

et posant

$$\alpha_2 = e^{i\theta},\ \theta \in]-\pi, +\pi],$$

on a :

$$\varphi(0, p) \sim \frac{1 - \varphi(0)}{4N\pi\sigma_1} \int_0^{\pi} \frac{\cos p\theta \, d\theta}{\sqrt{2k + 2\sigma^2(1 - \cos\theta)}}$$

$$\varphi(0, p) \sim \frac{1 - \varphi(0)}{8N\pi\sigma_1} \int_0^{\pi} \frac{\cos p\theta \, d\theta}{\sqrt{k/2 + \sigma^2 \sin^2 \frac{\theta}{2}}} = \frac{1 - \varphi_0}{4N\pi\sigma_1\sigma_2} \int_0^{\frac{\pi}{2}} \frac{\cos 2p\psi \, d\psi}{\sqrt{\sin^2\psi + \frac{k}{2\sigma_2^2}}} \quad (26)$$

ou, en posant $h = k/2\sigma_2^2$, que nous supposerons *petit* (comme dans la partie I, § II) 2) *b*), p. 400) :

$$\varphi(0, p) \sim \frac{1 - \varphi_0}{4N\pi\sigma_1\sigma_2} \int_0^{\frac{\pi}{2}} \frac{\cos 2p\psi \, d\psi}{\sqrt{h + \sin^2\psi}} = \frac{1 - \varphi_0}{4N\pi\sigma_1\sigma_2} I_p \quad (27)$$

Les intégrales elliptiques

$$I_p = \int_0^{\frac{\pi}{2}} \frac{\cos(2p\psi) \, d\psi}{\sqrt{h + \sin^2\psi}} = (-1)^p \int_0^{\frac{\pi}{2}} \frac{\cos(2p\varphi) \, d\varphi}{\sqrt{h + 1 - \sin^2\varphi}}$$

se calculent aisément par récurrence à partir des deux premières qui sont :

$$I_0 = (h + 1)^{-\frac{1}{2}} K(\alpha)$$

et

$$I_1 = -2(h + 1)^{\frac{1}{2}} E(\alpha) + (1 + 2h)(1 + h)^{-\frac{1}{2}} K(\alpha)$$

en utilisant les notations d'Abramowicz (1964) pour les intégrales complètes de première et deuxième espèces $K(\alpha)$ et $E(\alpha)$: l'angle α est défini par $\sin\alpha = (1 + h)^{-1/2}$.

La récurrence, qui s'obtient en intégrant entre 0 et $\frac{\pi}{2}$ la dérivée de $\cos(2p\varphi)\sqrt{h + 1 - \sin^2\varphi}$, est (pour $p \in N^+$) :

$$(2p + 1) I_{p+1} - 4p(2h + 1) I_p + (2p - 1) I_{p-1} = 0 \ (^2) \quad (28)$$

Un développement asymptotique de I_p pour les grandes valeurs de p peut être obtenu en remarquant que la transformation $J_p = \sqrt{p}\ I_p$ fournit pour J_p une

(1) De façon à retrouver le calcul que nous avons fait dans un cas particulier (Malécot, 1971).

(2) Ce résultat est généralisable à toute intégrale hyperelliptique possédant un nombre pair de points critiques à l'intérieur du cercle C_2. (Voir appendice 3).

récurrence à coefficients sensiblement constants (1), soit (avec une erreur *relative* de l'ordre de $p^{-3/2}$) :

$$2J_{p+1} - 4(2h + 1)J_p + 2J_{p-1} = 0$$

Des deux solutions particulières α_1^p et α_1^p de cette récurrence linéaire (α_1 de module < 1 et $\alpha_2 = 1/\alpha$ étant les racines de l'équation $\alpha^2 - 2(2h + 1)\alpha + 1 = 0$), seule α^p est bornée ; or I_p doit être borné quand $p \to +\infty$. On a donc :

$$J_p = \text{cte} \times \alpha_1$$

$$I_p = \text{cte} \times p^{-1/2}\alpha_1^p \qquad \text{avec } \alpha_1 = 2h + 1 - \sqrt{4h + 4h^2}$$

et la formule (27) donne alors, compte tenu de $h = k/2\sigma_2^2$

$$\varphi(0, p) \sim \text{cte} \times \frac{\left(1 - \frac{\sqrt{2k}}{\sigma_2}\right)^p}{\sqrt{p}} \tag{29}$$

qui diffère de la formule unidimensionnelle (23) par une décroissance plus rapide pour p grand. Mais la loi de décroissance *pour les premières valeurs de p* doit être déterminée *par récurrence numérique* à partir de I_0 et de I_1 définis ci-desus. D'ailleurs, pour $p = 0$, la formule (27) donne :

$$\varphi(0,0) = \varphi_0 \sim \frac{1}{1 + 4N\pi\sigma_1\sigma_2/K(\alpha)} \tag{30}$$

formule de structure notablement différente de la formule (22) du cas unidimensionnel: certes $4N\pi\sigma_1\sigma_2$ (ou $4N\sigma$) peut être interprété comme le nombre total d'individus dans l'ellipse (ou l'intervalle) de diamètres $4\sigma_1$ et $4\sigma_2$ (ou 4σ). Mais le facteur $\sqrt{2k}$ est ici remplacé par $\pi/K(\alpha)$, avec $\sin \alpha = (1 + k/2\sigma_2^2)^{-1/2}$: la variation en fonction de k est beaucoup plus lente.

Les valeurs de $\varphi(o)$ et la décroissance de $\varphi(p)$ ont été calculées sur ordinateur à partir de la récurrence (28) : les résultats ont été donnés dans la partie II, p. 402 et 403, et la lecture peut maintenant en être faite à partir de la formule générale $h = k/2\sigma_2^2$ et non du cas particulier $h = k_1/4m$ (h doit toutefois rester petit).

Rappelons que pour les plus petites valeurs ($h \leqslant 0{,}003$), l'approximation (29) est satisfaisante sauf pour les toutes premières valeurs de p, comme le montre la comparaison [pour $\alpha = 89°$ et $\alpha = 87°$, figures 4 *a*) et 4 *b*) de la partie II] de la courbe (en trait plein) représentant $\log I_p$ et de la courbe (en tirets) représentant $\text{Log}\left[(1 - 2\sqrt{h})^p\right] - \frac{1}{2}\log p$ (la comparaison est facilitée par un décalage vertical arbitraire).

Pour les plus grandes valeurs de h (fig. 4 *c* et 4 *d*) une meilleure approximation est fournie par la simple exponentielle $(1 - 2\sqrt{h})^p$, dont le logarithme est représenté en tirets. Comme c'est dans ce domaine que se situaient les calculs de MORTON (1970) cela explique qu'il ait contesté la validité pratique de la formule (29) : elle dépend de la valeur de h. Cela règle partiellement la controverse signalée par MORTON (1970, p. 581).

(1) Cf. partie II, p. 401, et MALÉCOT (1972 *b*).

De nombreux commentaires ont été consacrés (cf. IMAIZUMI, MORTON et HARRIS, 1970) au problème de la « dimensionalité » de la migration dans les populations humaines, lorsque le coefficient de parenté φ peut être expérimentalement estimé en fonction de la distance d. Les données de ROSIN (1956) sur les corrélations entre les fréquences des groupes sanguins A, B, O dans plus de 3 000 communes suisses suivant leurs distances mutuelles d, ont permis d'ajuster, pour les fréquences relatives des groupes A et B, une courbe de corrélation en décroissance rapide comme $\frac{1}{\sqrt{d}} e^{-\sqrt{2k}\frac{d}{\sigma}}$ [dimensionalité 2, formule (29)], et pour la fréquence absolue du groupe O, une décroissance plus lente comme $e^{-\sqrt{2k}\frac{d}{\sigma}}$ [dimensionalité 1, formule (23)]. Ce contraste s'explique si l'on se réfère aux études de YASUDA (1966) d'après lequel la dimensionalité de la migration en Suisse est 2 jusqu' à 100 km de migration, et 1 au-delà de 300 km (1) ; or, la courbe d'estimation de la corrélation pour les gènes A et B ne fait intervenir que les distances inférieures à 100 km (au-delà, la corrélation est négligeable), alors que le gène O manifeste une corrélation encore significative à des distances supérieures (2).

Il est assez remarquable que toutes les études faites sur l'ajustement d'une décroissance de la forme $\frac{1}{d^c} e^{-bd}$ aient fourni pour l'exposant c une valeur comprise entre 0 [formule (23)] et 1/2 [formule (29)], quoique les hypothèses que notre modèle formule quant à la migration ne soient que grossièrement vérifiées : cela atteste la « robustesse » du modèle. La constatation faite par IMAIZUMI et MORTON (1969) que c soit plus souvent voisin de 0 que de 1/2 souligne le fait que, dans les vallées de montagne monoclinales comme dans les plaines présentant des zones climatiques différant suivant la latitude, il doit exister une direction de migration maximum, la migration étant nettement plus faible dans la direction perpendiculaire. Il peut arriver aussi que la décroissance rapide des intégrales elliptiques figurant dans (27) ne permette pas d'observations significatives dans la zone (grandes distances) où leur expression asymptotique (29) s'applique (MORTON, 1970), et que pour les moyennes distances leur approximation par une pure exponentielle [formule (23)] soit meilleure, comme le suggère la figure 4 de la partie II pour les valeurs de $\alpha \leqslant 85°$.

C. — *Variation de la parenté en fonction du décalage de générations*

Revenons à la formule (20) du § IX, dans le cas unidimensionnel de *migration symétrique* : $L(\alpha) = L(1/\alpha)$.

Utilisons la formule asymptotique quand $\tau \to +\infty$ (DIEUDONNÉ, EVGRADOV, 1961) :

$$\int_{-\pi}^{+\pi} g(\theta)\, e^{\tau h(\theta)}\, d\theta \sim \sqrt{2\pi} g(0)\, e^{\tau h(0)} \sqrt{2/(-\tau h''(0))}$$

(1) Ce qui s'explique par la forme allongée du pays. YASUDA a trouvé pour la Suède des résultats analogues (mais avec des distances doubles).

(2) Après correction de l'effet d'un gradient géographique (MORTON, 1966) qui correspond peut-être à l'influence asiatique exercée de l'Est à l'Ouest.

formule ici valable (en posant $\alpha = e^{i\theta}$), puisque $h(\theta) = \log(1 - k) + \log L(e^{i\theta})$ est réel, et maximum pour $\theta = 0$:

$$L(e^{i\theta}) = l(0) + 2\sum_{p>0} l(p)\cos p\theta$$

Comme $h''(0) = -\sigma^2$, on obtient (en tenant compte du terme suivant du développement asymptotique).

$$\varphi(d, \tau) \sim (1 - \varphi_0)\frac{(1-k)^\tau}{4Nk\sigma\sqrt{\pi\tau}}\left[1 - \left(\frac{4d^2 - 1}{2\sigma^2} + \frac{1}{k}\right)/2\tau + \ldots\right]$$

D'où la définition, quand τ est grand, de la *distance génétique* :

$$D = -\log\varphi(d, \tau) = -\tau\log(1-k) + \frac{1}{2}\log\tau$$
$$-\log\frac{1-\varphi_0}{4Nk\sigma\sqrt{\pi}} + \frac{\frac{4d^2-1}{2\sigma^2} + \frac{1}{k}}{2\tau} + \ldots \qquad (31)$$

On voit que l'influence de la distance géographique d n'est appréciable que si $\frac{d}{\sigma}$ est au moins de l'ordre de $1/\sqrt{k}$, c'est-à-dire si d est tel que dans la formule (23) $\varphi(d)$ soit nettement inférieur à $\varphi(0)$. De toutes façons cette influence est faible lorsque $\sqrt{\tau}$ dépasse l'ordre de $\frac{d}{\sigma}$ (c'est-à-dire lorsque d est petit par rapport à l'écart-type $\sigma\sqrt{\tau}$ de la distance aléatoire parcourue par un gamète en τ générations) et alors D est proportionnel à τ, à un terme logarithmique près.

Remarque : Les formules (24) et (31) montrent qu'une distance géographique d et une distance temporelle τ ont le même « pouvoir de différenciation » (pour des gènes neutres de taux de mutation k) si $\sqrt{2k}\,\frac{d}{\sigma} = \tau k$, c'est-à-dire $\frac{d}{\sigma} = \tau\sqrt{\frac{k}{2}}$

Par exemple, pour 20 000 ans (1 000 générations d' *Homo sapiens*) et un taux de mutation $k = 2 \cdot 10^{-6}$, on obtient $\frac{d}{\sigma} = 1$;

Si σ est de l'ordre de quelques dizaines de km (MORTON, 1966), on voit que le « gradient de différenciation » de gènes *neutres* est beaucoup plus grand dans l'espace que dans le temps (compte tenu, bien sûr, des échelles de l'histoire humaine et de la géographie humaine) ; c'est pourquoi, il est si difficile de situer les « races » fossiles par rapport aux nôtres !

Exemple : évolution de l'hémoglobine des vertébrés.

Retenons de la formule (31) que la partie principale de la distance génétique D entre deux espèces apparentées est $k_2\tau$, en adoptant la notation k_2 (§ IV) au lieu de k, pour les acides aminés de l'hémoglobine. On peut d'ailleurs désigner ainsi le taux annuel, à condition que τ soit exprimé en années. Si deux espèces actuelles descendent d'un ancêtre commun qui vivait p années auparavant, et s'il y a eu depuis cet ancêtre deux évolutions parallèles de la protéine par mutations neutres indépendantes, la formule $k_2\tau$, avec $\tau = 2p$ (de l'ordre de quelques millions d'années), fournit

une estimation de la distance génétique D entre ces deux espèces [1], alors qu'une autre estimation est fournie par $-\log \varphi$ qui a pu être mesuré directement, puisque φ, probabilité d'identité de deux acides aminés homologues dans les chaînes β (ou α) de l'hémoglobine des deux espèces peut être estimé par le pourcentage d'acides aminés restés identiques : environ 18 p. 100 entre la Lamproie et l'Homme, ce qui donne $D = -\log \varphi \neq 1{,}3$, pour un décalage de temps $\tau = 2p = 10^{9}$ années. Ce résultat est en bon accord avec la valeur $k_2 = 10^{-9}$ déduite des durées de séparation pour les autres espèces de vertébrés (KIMURA, 1969) : le « taux de différenciation paléontologique » k_2 semble donc être le même pour la plupart des codons de l'hémoglobine.

D. — *Notion de « distance génétique »*

Qu'il s'agisse de distance d dans l'espace [formules (23) et (29)] ou de distance τ dans le temps [formule (31)], nous constatons que, si φ est le coefficient de parenté correspondant, et dans le cas de *migration homogène* et *constante dans le temps*, la fonction positive $D = -\log \varphi$ a toujours comme partie principale de son développement asymptotique (quand d ou τ est grand) une fonction *linéaire* de la distance d ou de la distance τ (les coefficients respectifs de d ou de τ étant $\frac{\sqrt{2k}}{\sigma}$ ou k).

Comme l'estimation du coefficient de parenté φ de deux races géographiques ou de deux espèces va être possible avec une précision croissante (MORTON, 1970 ; NEI, 1971) pour de nombreux locus polymorphiques ou pour de nombreuses protéines, il semble que la définition de la distance génétique par la formule $D = -\log \varphi$ présentera, sur bien d'autres définitions, l'avantage d'être à la fois *rationnelle* (reposant sur un modèle mathématique précis) et *calculable* (par moyenne pondérée des valeurs de D obtenues sur différents locus, ou sur différentes protéines, cette pondération devant tenir compte (NEI, 1971), d'une manière qui reste à étudier, des différentes valeurs de k qui leur correspondent).

Il est commode de conserver la définition de la distance génétique par $D = -\log \varphi$ dans le cas où la migration n'est pas homogène, quoique alors D ne soit plus une fonction linéaire de d : sa signification et son utilisation dans ce cas seront étudiées au § XI.

XI. — UTILISATION DE LOIS DE MIGRATION VARIANT DANS LE TEMPS OU DANS L'ESPACE

A. — *Cas où, par cessation de migration, un site devient complètement isolé*

Si l'on prend pour génération initiale F_0 celle qui correspond à la date d'isolement, le calcul du § IX pour un individu K de site y dans une génération ultérieure F_τ ($\tau > 0$) doit être modifié, parce que K provient nécessairement d'un des N indi-

[1] Si l'on comparait, entre les espèces, des gènes et non des codons, la distance génétique serait pratiquement infinie, à cause de la grandeur de k pour les gènes (§ IV).

vidus occupant dans la génération F_0 ce *même site y*. On a alors, comme expression du coefficient de parenté décalé de τ générations *dans le même site y*,

$$\varphi(0, \tau) = (1 - k)^{\tau} [\varphi_0 + (1 - \varphi_0)/2N]$$

qui montre que, si τ n'est pas très grand et si l'effectif N du groupe isolé est assez petit pour que $\frac{1}{2N}$ soit nettement plus grand que φ_0, $\varphi(0, \tau)$ peut dépasser notablement φ_0 et approcher de $1/2N$. Alors les descendants ressemblent plus étroitement à l'un des fondateurs que les fondateurs ne se ressemblent entre eux (1).

B. — *Cas où la matrice de covariance de la migration bidimensionnelle varie en fonction du site x*

1. Nous supposerons toujours — pour éviter les difficultés indiquées en III-*f*) — que la translation moyenne ne dépende pas du site ; mais la matrice de covariance pourra maintenant en dépendre, c'est-à-dire (voir VIII-B) que les axes principaux et les « variances principales » σ_1^2 et σ_2^2 de la migration dans le site x pourront dépendre de x ; attachons à chaque point x une « indicatrice elliptique » de centre x, de sommets situés aux abscisses $\pm \sigma_1$ et $\pm \sigma_2$ sur les deux axes principaux correspondants ; cette indicatrice attache au point x, supposé repéré par des coordonnées quelconques, une forme quadratique définie positive et donc un ds^2 riemannien dépendant des coordonnées de x et de leurs différentielles.

En adoptant ce ds^2 pour définir la distance élémentaire et (lorsque l'espace riemannien est à connexion euclidienne) la distance finie entre deux points quelconques x et y, et en prenant les courbes coordonnées tangentes aux axes principaux en chaque point, la matrice de migration bidimensionnelle devient alors une matrice constante ; et les calculs faits précédemment s'appliquent : $D = -\log \varphi$ varie proportionnellement à la distance d calculée suivant une géodésique.

2. La théorie précédente suppose une loi de migration continue définie sur un espace continu (2). En pratique le coefficient de parenté φ ne peut être estimé que sur les $\frac{n(n-1)}{2}$ couples de groupes correspondant à n sites géographiques ;les valeurs correspondantes de $D = -\log \varphi$ fournissent les distances génétiques deux à deux de ces n sites, distances qui seront très loin d'être proportionnelles à deux distances géographiques lorsqu'il y a interposition d'obstacles tels que montagnes, mers, déserts, etc. Ces distances génétiques définissent donc un « espace riemannien discret » dont les éléments sont n points, et qui ne peut en général être plongé que dans un espace euclidien à $(n - 1)$ dimensions. Mais, s'il advient que la dimension de cet espace puisse approximativement être ramené à 3, par exemple en analysant le nuage de points dans R^{n-1} suivant 3 composantes principales (3) et en négligeant les autres, il sera possible de représenter l'ensemble des distances génétiques dans R^3 sur une

(1) A titre d'exemple, rappelons (MALÉCOT, 1948) que l'appartenance au groupe sanguin O de presque tous les Amérindiens peut s'expliquer par l'effectif très réduit du « stock de fondation » mongoloïde.

(2) Il faudrait même que l'analyse soit faite « en temps continu » pour donner un sens concret à un déplacement infiniment petit.

(3) Suggestion de CAVALLI-SFORZA et BODMER (1971).

« carte en courbes de niveau » dont la comparaison avec la carte géographique correspondante sera très instructive : quelles seront les hauteurs génétiques du Grand Himalaya et du Petit Himalaya-Terai séparant respectivement les Népalais des Mongoloïdes Tibétains et des Indiens du Nord ? Quel sera l' « élargissement » du Sahara pour tenir compte de la grande distance génétique entre Noirs soudanais et Méditerranéens nord-africains ? Quels sont les gradients de transition en Moyen-Orient ? Quelle est la position des Australiens par rapport aux Veddahs de Ceylan, aux Indonésiens, aux Mélanésiens ? Quelle est l'importance du fossé racial qui sépare les Australiens des Polynésiens, qui sont les principaux occupants du Pacifique central depuis Hawaii jusqu'au nord de la Nouvelle-Zélande, et quelles sont les affinités mongoloïdes de ceux-ci ? (cf. CAVALLI-SFORZA et BODMER, 1971, carte de la page 711).

Ces problèmes pourront être mieux résolus lorsqu'on disposera de nombreux locus occupés chacun par de nombreux allèles dont aucun ne soit prépondérant ; circonstance qui, à la fois, rend plus vraisemblable l'hypothèse de neutralité de chacun de ces gènes, et assure une meilleure estimation du coefficient de parenté. Le calcul du coefficient de parenté à partir d'un petit nombre d'allèles nécessite en effet la connaissance de leurs valeurs d'équilibre, ce qui est parfois impossible, et même dépourvu de sens lorsque chacun d'eux est transitoire.

Il y a lieu de remarquer toutefois que, dans le cas où il n'existe qu'un petit nombre d'allèles dont l'un au moins a une fréquence prépondérante, et en particulier dans le cas de diallélisme, on peut parfois admettre que la fréquence observée fluctue autour d'une valeur d'équilibre due à l'hétérosis (supériorité sélective des hétérozygotes), soit localement, soit en moyenne sur un grand nombre de colonies, comme dans le cas de *Cepaea nemoralis* étudié par LAMOTTE (1951) et CAIN et SHEPPARD. En linéarisant la pression de sélection au voisinage de l'équilibre, et en désignant par k le coefficient de rappel correspondant, on peut montrer (MALÉCOT, 1948, 1969) que le coefficient de parenté (utilisable, comme nous l'avons montré, comme « corrélation gamétique » permettant de déduire les corrélations entre les fréquences dans les différents sites) reste calculable par les formules données précédemment, à condition d'y donner à k la valeur de ce coefficient de rappel et non du taux de mutation ; cela s'applique en particulier à la décroissance, en fonction de la distance, des corrélations entre fréquences du gène *sb* pour *Cepaea nemoralis*.

Reçu pour publication en février 1973.

SUMMARY

THE GENETICS OF DIPLOID NATURAL POPULATIONS FOR A SINGLE LOCUS
III. — KINSHIP, MUTATION, MIGRATION

Evidence from molecular biology indicates the existence of numerous polymorphism in the ' fine structure ' of proteins and therefore in the corresponding segment of DNA. Each substitution of one amino acid for another, even though this may not be detectable, constitutes a mutation having a very small probability for each codon and each such ' molecular mutation ' produces a new gene, a new allele — in general neutral with respect to selection in relation to the one previously existing. Multi-allelism is therefore frequent even in small populations and each allele has only a transitory existence until it is replaced by yet further mutations. In the long term the concept of ' equilibrium frequency ' introduced in part I (chapter II) loses all signi-

ficance. On the other hand the coefficient of kinship defined in part II (chapter I) retains its meaning : the probability that the two sampled loci are ' identical ', *i. e.* they derive without imutation from the same locus in a common ancester. There is now a difference that two 'non identical' loci are no more 'independent' (as in part II where each 'β mutation' was a 'transition mutation' now giving rise to a new codon or a new allele).

The calculations of part II, assuming homogeneous migration, are now treated by new methods which analyse Φ as the sum of contributions of kinship chains of different orders (§ IV and VII). The kinship between different generations is studied in § IX. § X gives the asymptotic expressions for large intervals of distance and time and the introduction of the *genetic distance* $D = - \text{Log } \Phi$ is justified from the fact that, in the case of homogeneous migration with a constant mutation rate, it is proportional to geographical distance and time interval.

§, XI, B suggests two methods by which non-homogeneous migration might be studied : firstly by using a geometrical transformation to convert it to the homogeneous case or secondly in devising a new geographical map where the relative distances are proportional to the logarithms of the kinship coefficients determined experimentally.

RÉFÉRENCES BIBLIOGRAPHIQUES

ABRAMOWICZ, STEGUN, 1964. *Handbook of mathematical functions.* National Bureau of Standards.

ALLEN, 1965. Random and nonrandom inbreeding. *Eugen. Quart.*, **12**, 181-98.

CAVALLI-SFORZA, BODMER, 1971. *Genetics of human population.* Freeman et Co. San Francisco.

CROW, 1954. Breeding structure of populations II. Effective population number. In KEMPTHORNE *et al.*, *Statistics and Mathematics in Biology*, Iowa State College Press, Ames, Iowa, 543-556.

CROW, MENGE, 1965. Measurements of inbreeding from the frequency of marriages between persons of the same surname. *Eugen. Quart.*, **12**, 199-203.

DAYHOFF, ECK, 1968. *Atlas of protein sequence and structure.* Nat. Biomed. Res. Found. Silver Spring, Maryland.

DIEUDONNÉ. *Calcul infinitésimal.* Gauthier-Villars, Paris.

IMAIZUMI, MORTON, HARRIS, 1970. Isolation by distance in artificial populations. *Genetics*, **66**, 569-582.

JACQUARD, 1970. *Structure génétique des populations.* Masson Paris.

KENDALL, 1952. On bacterial law of growth. *Journal of the Royal Statistical Society.* B, **14**, 14.

KIMURA, 1954. Random fluctuation of selection intensities. *Genetics*, **39**, 280-295.

KIMURA, 1962. On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713-719.

KIMURA, 1967. On the evolutionary adjustement of spontaneous mutation rates. *Genetical Res.*, **9**, 23-34.

KIMURA, 1968. Evolutionary rate at the molecular levels. *Nature*, **217**, 624-626.

KIMURA, 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893-903.

KIMURA, 1971. Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biol.*, **2**, 174-208.

KIMURA, CROW, 1963. The measurement of effective population number. *Evolution*, **17**, 279-288.

KIMURA, CROW, 1964. The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725-738.

KIMURA, CROW, 1968 *a*. Evolutionary rate at the molecular level. *Nature*, **217**, 624-626.

KIMURA, CROW, 1768 *b*. Genetic variability maintained in a finite population due to the mutational production of neutral and nearly neutral isoalleles. *Genetical Res.*, **11**, 247-269.

KIMURA, WEISS, 1964. The stepping stone modal of population structure and the decrease of genetic correlation with distance. *Genetics*, **49**, 561-576.

KIMURA, WEISS, 1965. A mathematical analysis of the stepping stone model of genetic correlation. *J. Appl. Prob.*, **2**, 129-149.

LAMOTTE, 1951. Étude des populations naturelles de *Cepaea nemoralis*. *Bull. Biol. de France et de Belgique.* Supplément 35.

MALÉCOT G., 1948. *Les mathématiques de l'hérédité.* Masson, Paris.

MALÉCOT G., 1948 *b*. Les processus stochastiques de la génétique dans le calcul des probabilités et ses applications. *Colloque international du C. N. R. S.*, **13**, 121-126.

MALÉCOT G., 1950. Quelques schémas probabilistes sur la variabilité des populations naturelles. *Ann. Univ. Lyon, Sciences, Section A*, **13**, 37-60.

MALÉCOT G., 1951. Un traitement stochastique des problèmes linéaires (mutation, linlage, migration) en génétique de populations. *Ann. Univ. Lyon, Sciences, Section A*, **14**, 79-117.

MALÉCOT G., 1952. Les processus stochastiques et la méthodes des fonctions génératrices ou caractéristiques. *Publ. Inst. Statist., Paris*, **1**, F 3, 1-16.

MALÉCOT B., 1954. Sur les modèles stochastiques, linéaires, asymptotiquement stationnaires. *Ann. Univ. Lyon, Section A*, **17**, 19-35.

MALÉCOT G., 1959. Les modèles stochastiques en génétique de population. *Publ. Inst. Stat.* (Univ. Paris) **8**, 173-210.

MALÉCOT G., 1960. Sur l'estimation des taux de mutation et de sélection, compte tenu des migrations. *31e session se l'Institut international de Statistique Bruxelles 1958. Bull. Int. Statist. Inst.*, **37**, 3, 1-11.

MALÉCOT G., 1965 *a*. Les covariances dans un milieu en équilibre statistique. *35e session de l'Institut international de Statistique, Belgrade.*

MALÉCOT G., 1965 *b*. Évolution continue des fréquences d'un gène mendélien. *Ann. Inst. Henri-Poincaré*, **11/2**, 137-150.

MALÉCOT G., 1967 *a*. Identical loci and relationship. *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, **4**, 317-332.

MALÉCOT G., 1967 *b*. Conséquences statistiques de la parenté. *36e session de l'Institut international de Statistique, Sydney*, 651-668.

MALÉCOT G., 1969 *a*. *The mathematics of heredity*. Freeman and Co, San Francsico.

MALÉCOT G., 1969 *b*. Consanguinité panmictique et consanguinité systématique. *Ann. Génét. Sél. anim.*, **1**, 237-242.

MALÉCOT G., 1971. Génétique des populations diploïdes naturelles dans le cas d'un seul locus. I. Évolution de la fréquence d'un gène. Étude des variances et des covariances. *Ann. Génét. Sél. anim.*, **3**, 255-280.

MALÉCOT G., 1972 *a*. Génétique des populations naturelles dans le cas d'un seul locus. II. Étude du coefficient de parenté. *Ann. Génét. Sél. anim.*, **4**, 385-409.

MALÉCOT G., 1972 *b*. Excerpta Médica, Amsterdam.

MALÉCOT G., 1973. Heterozygosity and relationship in finite populations. *Theor. Pop. Biol.* (preprint).

MARUYAMA, 1970. On the rate of decrease of heterozygosity in circular stepping stone model of populations. *Theor. Pop. Biol.*, **1**, 101-119.

MAYR, 1963. *Animal species and evolution*. Harvard University Press.

MORTON, CHUNG, MI, 1967. *Genetics of interracial crosses in Hawaii*. Karger Basel.

MORTON, 1970. Isolation by distance in artificial ppulations. *Genetics*, **66**, 569-582.

MORTON, YEE, HARRIS, LEW, 1971. Bioassay of kinship. *Theor. Pop. Biol.*, **2**, 507-524.

NEI, 1971. Identity of genes and genetic distance between populations. *Genetics*, **68**, 47.

NEI, 1972. Genetic distance between populations. *Amer. Not.* **106**, 283.

ROSIN, 1956. Die Verteilung der ABO Blutgruppen in der. Schweiz. *Arch. Klaus-Stift. Vererbforsch.*, **31**, 1-127.

THEIL, 1970. *Principles of economics*. North-Holland.

WRIGHT S., 1921. Systems of mating. *Genetics*, **6**, 111-178.

WRIGHT S., 1931. Evolution in Mendelian populations. *Genetics*, **16**, 97-159.

WRIGHT S., 1938. Size of population and breeding structure in relation to evolution. *Sciences*, **87**, 430-431.

WRIHGT S., 1943. Isolation by distance. *Genetics*, **28**, 114-138.

WRIGHT S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*. **19**, 395-420.

YASUDA, 1966. *The genetical structure of northeastern Brazil*. Ph. D. thesis, University of Hawaii, Honolulu.

ZUCKERKANDL, PAULING, 1965. Evolutionary divergence and convergence in proteins. In *evolving genes and proteins*, V. Bryson and H. J. Vogel (ed.), Academic Press, New York. Pp, 97-166.