# THE CONCEPT OF IDENTITY OF GENES BY DESCENT

OSCAR KEMPTHORNE
IOWA STATE UNIVERSITY

## 1. Introduction

The idea that the underlying mechanism of inheritance is the transmission according to elementary laws of probability of particulate units was, of course, due to the genius of Mendel. If one of the parents of an individual only has an $A$ gene and an offspring has an $A$ gene, then the $A$ gene of the offspring is a copy of the $A$ gene of the particular parent, and the two $A$ genes are identical by descent. Mendel knew this, even though he did not use precisely these words to express the fact. Early in the rediscovery of Mendel's work, mathematicians or statisticians or biometricians (whatever one wishes to call the individuals) were impressed by the mathematical interest of the Mendelian system. It was realized early, of course, that inbreeding, that is, the mating of related individuals was an important tool for the understanding of genetic phenomena. It, therefore, became necessary to develop some of the theory of inbreeding. Pioneers in this work were Robbins and Jennings. But the great advance was made by Sewall Wright [7], who, it appears to me, singlehandedly developed the method of path coefficients to enable the answering of many important questions of inbreeding. The magnitude of Wright's contribution stuns my imagination. It is true that much of his work was algebraic computation and this any one could do. But the erection of a framework for the computations was a great intellectual feat. It would not, however, be stretching the history to say that to most of the world the method of path coefficients was a mystique, obviously very powerful in the hands of its inventor, but quite obscure to many others. Perhaps the cornerstone of Wright's work was the coefficient of correlation of uniting gametes or the coefficient of inbreeding $F$ and Wright developed knowledge of the progress of $F$ with various systems of inbreeding.

It appears to be the great contribution of Malécot [5] to put the Wright ideas into a form which was more readily understood and communicable, by introducing the ideas of genes being alike by descent and by considering inbreeding purely as a matter of probability of genes being identical by descent. To fix the ideas, let a diploid individual $X$ have two genes $a$, $b$ at a locus. Then,

Journal Paper No. J-5618 of the Iowa Agricultural and Home Economics Station, Ames, Iowa. Project 890. Part of the work reported was done while the author was Visiting Professor of Genetics and Statistics at Stanford University with the support of an NIH Grant GM10452 in 1964. Part was done in connection with NSF Grant 19218.

the coefficient of inbreeding $F$ is the probability that genes $a$, $b$ have arisen by the copying process of reproduction from a single gene in the ancestry. At the same time Malécot defined the "coefficient de parente" (which I translated as "coefficient of parentage") of two individuals $X$ and $Y$ as the probability that a random gene of $X$ and a random gene of $Y$ at the same locus are identical by descent. The inherent simplicity of these ideas is conveyed in the simple formulae for diploids,

$$F_{X\times Y} = r_{XY},$$

$$r_{A(B\times C)} = \frac{1}{2}(r_{AB} + r_{AC}). \tag{1.1}$$

It is then a relatively simple matter to develop the progress of $F$ over generations of inbreeding. The great bulk of the elementary results is given in my book [4]. Of course, the ensuing result is a function of the status of the base population from which the probabilities are calculated, and this is true whatever mode of calculating probabilities is followed. The criticisms by Fisher [1] of Wright's $F$ are, therefore, clearly unfounded.

Malécot [5] applied these ideas also to the progress of a finite population and to quantitative inheritance. Later I [3] showed that the covariance of two relatives $X$ and $Y$, neither of which is inbred, in a random mating population without linkage, with respect to a metrical trait, which contains a random environmental contribution, is equal to

$$2r_{XY}\sigma_A^2 + u_{XY}\sigma_D^2 + \sum_{r+s=3}^{n} (2r_{XY})^r (u_{XY})^s \sigma_{A^rD^s}^2, \tag{1.2}$$

where $r_{XY}$ is given above and $u_{XY}$ is the probability that the genes of $X$ are identical by descent to the genes of $Y$ at a locus, and $\sigma_A^2$, $\sigma_D^2$, and so forth, are components of genotypic variance. More recently Harris [2] has generalized this result to the case when the individuals $X$ and $Y$ are inbred, that is, have nonzero coefficient of inbreeding, and his ideas were a stimulus to some of the material I present later.

It seems to be inherent in the concept of identity by descent that the processes of genetics for which it has utility are processes which do not depend on the actual identity of the genes. Thus, each gene is to be subjected to the same probability processes as every other gene. In contrast to this, if the different genes have different selective values (however this be defined), their fates in the evolutionary process will depend on their actual names. So the idea of identity by descent has so far been applied only to the processes of a genetic population with "neutral" genes. I would not wish to imply by this, that I believe there are any such "neutral" genes. It was thought at one time that there were genes that were "neutral" as regards "selective fitness," but every such case has been shown later to be fallacious. Looking back, one may well wonder how the idea was ever judged reasonable, but perhaps I am using hindsight which is always better than foresight.

One should not, on account of the above, consider that a theory of the progress of a population with "neutral" genes is irrelevant, because such a theory gives a base of reference by which to evaluate the lack of neutrality of genes.

The applications of the concept of identity of a pair of genes to regular systems of inbreeding are reviewed extensively in my book [4]. The aim of the present paper is to present some basic ideas of the application to the progress of finite populations, and to describe some extensions of the basic idea to the status of more than two genes.

## 2. An elementary use of the concept

Suppose we start with a diploid population of size $N$, having $N_1$ $A$ genes and $N_2$ $a$ genes, with $N_1 + N_2 = 2N$. Suppose that there is no selection, mutation, immigration or emigration. Then, the fate of genes is determined solely by the elementary probability processes of Mendelism. Suppose also that each generation arose by mating of individuals of the preceding generation, these matings possibly being based on consanguinity, but random within any consanguinity restrictions. Suppose also that we can calculate on the basis of the mating system, the probability that two genes of an individual are identical by descent and find this to be $F$, say. Then the probability status of this individual is given by the array

$$(2.1)\quad F\left\{\frac{N_1}{2N}AA + \frac{N_2}{2N}aa\right\} + (1-F)\left\{\frac{1}{2N(2N-1)}N_1(N_1-1)AA + N_2(N_2-1)aa + 2N_1N_2Aa\right\}.$$

That this is so follows by elementary arguments because the two events: (i) the genes of an individual being alike by descent and (ii) the genes of an individual being unalike by descent, are mutually exclusive, with probabilities $F$ and $1 - F$, respectively. In the absence of any selection, the conditional probability that a single gene from which a pair are both descended is an $A$ gene is equal to $N_1/2N$, and that it is an $a$ gene is $N_2/2N$. If the genes are copies of two different genes in the base population, the probability that both are $A$ genes is $N_1(N_1 - 1)/2N(2N - 1)$, and so on.

This argument gives then the probability status of a random individual with coefficient of inbreeding $F$. It is to be noted that it cannot be used to give the probability status of two individuals. In fact the argument gives only the status of two genes united in the particular individual.

If, of course, $N_1$ and $N_2$ are both indefinitely large with $N_1/2N = p$ and $N_2/2N = q$, the probability status of an individual is

$$(2.2)\qquad FpAA + qaa + (1-F)p^2AA + 2pqAa + q^2aa.$$

The generalization to $k$ types of genes is quite obvious.

### 3. A finite dioecious population

Suppose we have a finite population of $N$ individuals, and that each member of a succeeding generation arises by random mating. Let the individuals in generation $t$ be

$$A^t_{11}A^t_{12},\ A^t_{21}A^t_{22},\ \cdots,\ A^t_{N1}A^t_{N2}. \tag{3.1}$$

A possible mating

$$A^t_{i1}A^t_{i2} \times A^t_{i'1}A^t_{i'2} \tag{3.2}$$

gives the offspring array

$$\frac{1}{4}\left[A^t_{i1}A^t_{i'1} + A^t_{i1}A^t_{i'2} + A^t_{i2}A^t_{i'1} + A^t_{i2}A^t_{i'2}\right], \tag{3.3}$$

so that the inbreeding coefficient $F_{t+1}$ of any individual in generation $(t + 1)$ is

$$\frac{1}{N(N-1)} \sum_{\substack{i,i' \\ i \neq i'}} P\{A^t_{ij} = A^t_{i'j'}\}, \qquad j,j' \text{ free}, \tag{3.4}$$

which is the average coefficient of parentage of two individuals in generation $t$. So

$$F_{t+1} = R_t. \tag{3.5}$$

An obvious argument gives

$$R_{t+1} = \frac{1}{N^2}\left[N\,\frac{1}{2}\,(1 + F_t) + N(N-1)R_t\right], \tag{3.6}$$

and the well known conclusion

$$F_{t+2} = \frac{1}{2N} + \left(1 - \frac{1}{N}\right)F_{t+1} + \frac{1}{2N}\,F_t \tag{3.7}$$

follows. This argument seems superior to that presented in my book [4].

### 4. A finite haploid population with random viability

A common way of handling the two conflicting aspects of finiteness of population and variability of number of offspring per individual is to follow a conditioning argument. One may suppose, for instance, that the individuals of the finite population produce offspring independently according to a Poisson distribution, and then consider the probabilities of numbers of each type conditional on the total offspring population being equal to the population size $N$. This process is mathematically workable and is a way of introducing reproductive competition. I find some obscurity in understanding the effect of this conditioning on the distribution of actual number of progeny in the population, though presumably this can be worked out.

An alternative easy approach with regard to homozygosity can be worked out by means of the concept of identity of genes by descent in the following way.

Consider the population at generation $t$ with individuals

$$A_1^t, A_2^t, \cdots, A_N^t. \tag{4.1}$$

Take a distribution of a nonnegative random variable $X$. Let $N$ random values be $x_1, x_2, \cdots, x_N$, and let

$$p_i = \frac{x_i}{\sum x}. \tag{4.2}$$

Let the probability that an individual in generation $t + 1$ arises from individual $A_1^t$ be $p_1$, from individual $A_2^t$ be $p_2$ and so on. Label the individuals in generation $t + 1$ as

$$A_1^{t+1}, A_2^{t+1}, \cdots, A_N^{t+1}. \tag{4.3}$$

Then, with independent drawing for each individual of generation $(t + 1)$, we have

$$P\{A_u^{t+1} = A_v^{t+1}\} = \sum_i p_i^2 + \sum_{i,i'} p_i p_{i'} P\{A_i^t = A_{i'}^t\}, \tag{4.4}$$

where the $P$ are probabilities of genes being identical by descent. The left side is the same for all pairs of individuals after the first generation, so we can write

$$F_{t+1} = \sum_i p_i^2 + [1 - \sum p_i^2] F_t \tag{4.5}$$

or with $P = 1 - F$, so that $P$ is the panmictic index of Wright,

$$P_{t+1} = (1 - \sum p_i^2) P_t. \tag{4.6}$$

Hence, if $\mu_2'$ is the second moment about the origin of the "viabilities," the $p_i$, we have

$$P_{t+1} = (1 - N\mu_2') P_t, \tag{4.7}$$

a very simple recurrence equation. In the special case when all $p_i$ equal $1/N$, the equation is

$$P_{t+1} = \left(1 - \frac{1}{N}\right) P_t \tag{4.8}$$

which is very well known.

## 5. A simplified two niche problem

I am indebted to E. Pollak for the following example and some ideas about it. The progress of a population divided into two parts, or niches, between which there is random migration is of some general genetic interest. Suppose the generations do not overlap, and that in niche 1 in generation $t$, the genes are

$$A_{11}^t, A_{12}^t, \cdots, A_{1N_1}^t, \tag{5.1}$$

and in niche 2 the genes are

$$A_{21}^t, A_{22}^t, \cdots, A_{2N_2}^t. \tag{5.2}$$

Suppose a gene for niche 1 in generation $(t + 1)$ arises with probability $(1 - p)$ from niche 1, and with probability $p$ from niche 2, and similarly for a gene in niche 2. Define,

$F_1^{(t)}$ = probability that two genes in niche 1 in generation $t$ are identical by descent,
$F_2^{(t)}$ = same for two genes in niche 2,
$R^t$ = probability that a gene in niche 1 and a gene in niche 2 in generation $t$ are identical by descent.

Then, with no selection of any sort, the probability array of a gene in niche 1 in generation $t + 1$ is

$$(5.3) \qquad (1-p)\left[\frac{1}{N_1}A_{11}^t + \frac{1}{N}A_{12}^t + \cdots + \frac{1}{N}A_{1N_1}^t\right] + p\left[\frac{1}{N_2}A_{21}^t + \frac{1}{N_2}A_{22}^t + \cdots + \frac{1}{N_2}A_{2N_2}^t\right].$$

Hence, $P\{A_{1u}^{t+1} = A_{1v}^{t+1}\}$, the probability that two genes, labeled for convenience as $u$ and $v$, in niche 1 in generation $(t + 1)$ are identical by descent is equal to

$$(5.4) \qquad \frac{(1-p)^2}{N_1^2}N_1 + (1-p)^2\frac{N_1(N_1-1)}{N_1^2}F_1^t + \frac{p^2}{N_2^2}N_2 + p^2\frac{N_2(N_2-1)}{N_2^2}F_2^t + 2(1-p)pR^t.$$

A similar expression with the subscripts 1 and 2 interchanged holds for a pair of genes in niche 2. Also,

$$(5.5) \qquad \begin{aligned} R^{t+1} &= P\{A_{1u}^{t+1} = A_{2v}^{t+1}\} \\ &= \frac{(1-p)p}{N_1^2}N_1 + \frac{(1-p)p}{N_1^2}N_1(N_1-1)F_1^t + \frac{(1-p)p}{N_2^2}N_2 \\ &\quad + \frac{(1-p)p}{N_2^2}N_2(N_2-1)F_2^t + \frac{p^2+(1-p)^2}{N_1N_2}N_1N_2R^t. \end{aligned}$$

If now, as is advisable in all calculations of this type, we let $P = 1 - F$, $S = 1 - R$, we have

$$(5.6) \qquad \begin{aligned} P_1^{t+1} &= \frac{(1-p)^2(N_1-1)}{N_1}P_1^t + \frac{p^2(N_2-1)}{N_2}P_2^t + 2p(1-p)S^t, \\ P_2^{t+1} &= \frac{p^2(N_1-1)}{N_1}P_1^t + \frac{(1-p)^2(N_2-1)}{N_2}P_2^t + 2p(1-p)S^t, \\ S^{t+1} &= \frac{p(1-p)(N_1-1)}{N_1}P_1^t + \frac{p(1-p)(N_2-1)}{N_2}P_2^t + [p^2+(1-p)^2]S^t. \end{aligned}$$

So we have a simple recurrence relation for the three quantities $P_1$, $P_2$ and $S$, which can be solved for any choice of $N_1$, $N_2$ and $p$.

We consider here the solution of a special case only. We shall calculate the

values for $P$, $R$, $S$ with reference to a base population for which we take $P_1$, $P_2$ and $S$ to be unity, of course, and for which $N_1$ equals $N_2$. Obviously, under these circumstances $P_1^t = P_2^t = P^t$, say, so the recurrence equations reduce to two only, namely,

$$P^{t+1} = [(1-p)^2 + p^2]\left(1 - \frac{1}{N}\right)P^t + 2p(1-p)S^t,$$

(5.7)

$$S^{t+1} = 2p(1-p)\left(1 - \frac{1}{N}\right)P^t + [p^2 + (1-p)^2]S^t.$$

The characteristic equation is

$$\begin{vmatrix} [(1-p)^2 + p^2]\left(1 - \frac{1}{N}\right) - \lambda & 2p(1-p) \\ 2p(1-p)\left(1 - \frac{1}{N}\right) & (1-p)^2 + p^2 - \lambda \end{vmatrix} = 0 \tag{5.8}$$

or

(5.9)

$$\lambda^2 - \lambda\left(2 - \frac{1}{N}\right)[(1-p)^2 + p^2] + \left(1 - \frac{1}{N}\right)\{[(1-p)^2 + p^2]^2 - 4p^2(1-p)^2\} = 0$$

$$= \lambda^2 - \lambda\left(2 - \frac{1}{N}\right)[1 - 2p(1-p)] + \left(1 - \frac{1}{N}\right)[1 - 4p(1-p)] = 0.$$

The roots $\lambda_1$ and $\lambda_2$ are, therefore,

(5.10)

$$\left(1 - \frac{1}{2N}\right)[1 - 2p(1-p)]$$

$$\pm\left\{\left(1 - \frac{1}{2N}\right)^2[1 - 4p(1-p) + 4p^2(1-p)^2] - \left(1 - \frac{1}{N}\right)[1 - 4p(1-p)]\right\}^{1/2}$$

$$= \left(1 - \frac{1}{2N}\right)[1 - 2p(1-p)]$$

$$\pm\left\{\left(1 - \frac{1}{N} + \frac{1}{4N^2}\right)[1 - 4p(1-p) + 4p^2(1-p)^2] - \left(1 - \frac{1}{N}\right)[1 - 4p(1-p)]\right\}^{1/2}$$

$$= \left(1 - \frac{1}{2N}\right)[1 - 2p(1-p)]$$

$$\pm\left\{\left(1 - \frac{1}{N}\right)4p^2(1-p)^2 + \frac{1}{4N^2}[1 - 4p(1-p) + 4p^2(1-p)^2]\right\}^{1/2}.$$

$P^t$ and $S^t$ are, therefore, of the form

$$a\lambda_1^t + b\lambda_2^t. \tag{5.11}$$

A somewhat simple answer is obtained if one assumes that $p$ equals $c/N$ and is small, so that a small number of genes migrate from one niche to the other to produce the next generation. In this case the larger root is

$$\lambda_1 = \left(1 - \frac{1}{2N}\right)\left[1 - \frac{2c}{N} + O\left(\frac{1}{N^2}\right)\right] + \left[\frac{4c^2}{N^2} + \frac{1}{4N^2} + O\left(\frac{1}{N^3}\right)\right]^{1/2}, \tag{5.12}$$

which is approximately equal to

$$1 - \frac{1}{2N}[1 + 4c - (1 + 16c^2)^{1/2}]. \tag{5.13}$$

This result is related to the results of Moran [6]. Other results in connection with population subdivision have been obtained with this type of argument by my colleague E. Pollak.

## 6. The consideration of triples, quadruples, and so on, of genes

Just as one can consider the probability that a pair of genes are identical by descent, one can consider the status of a triple of genes, or a set of four genes and so on. Harris [2], in his generalization of the work on covariance of two relatives, $X$ and $Y$ with genes say $x_s$, $x_d$, for $X$, and $y_s$, $y_d$ for $Y$, the subscripts $s$ and $d$ denoting genes received from sire and dam, respectively, found it necessary to consider all the possible configurations with regard to identity by descent of the four genes. The possibilities are as follows in which vertical bars separate groups of genes identical by descent

| | |
|---|---|
| $x_s\ x_d\ y_s\ y_d$ | $x_s\ x_d \mid y_s \mid y_d$ |
| | $x_s\ y_s \mid x_d \mid y_d$ |
| $x_s\ x_d\ y_s \mid y_d$ | $x_s\ y_d \mid x_d \mid y_s$ |
| $x_s\ x_d\ y_d \mid y_s$ | $x_d\ y_s \mid x_s \mid y_d$ |
| $x_s\ y_s\ y_d \mid x_d$ | $x_d\ y_d \mid x_s \mid y_s$ |
| $x_d\ y_s\ y_d \mid x_s$ | $y_s\ y_d \mid x_s \mid x_d$ |
| $x_s\ x_d \mid y_s\ y_d$ | $x_s \mid x_d \mid y_s \mid y_d.$ |
| $x_s\ y_s \mid x_d\ y_d$ | |
| $x_s\ y_d \mid x_d\ y_s$ | |

So, taking account of origin of the genes in an individual, there are 15 different possibilities, and a complete accounting of the probability status of two individuals would require the determination of the probabilities of the 15 different possibilities.

The calculus of probabilities of two genes being identical by descent is very easy, as I have already mentioned. The general calculus of probabilities with regard to four genes will not be easy. A beginning was made by Harris [2], and further work is in progress.

I believe it is interesting to relate the above to Fisher's working out of the progress under full-sib inbreeding. If for ease of typography and reading we denote the two genes of the male by $ab$ and of the female by $cd$, and if we use

the same letter out of $a$, $b$, $c$, $d$ to denote genes identical by descent, the possible configurations are

$$\begin{array}{ll} aa \times aa & (1) \\ aa \times ab & (4) \\ aa \times ba & (4) \\ ab \times aa & (4) \\ ba \times aa & (4) \\ aa \times bb & (3) \\ ab \times ab & (2) \\ ab \times ba & (2) \\ aa \times bc & (5) \\ ab \times ac & (6) \\ ab \times ca & (6) \\ ba \times ac & (6) \\ ab \times cb & (6) \\ bc \times aa & (5) \\ ab \times cd & (7). \end{array}$$

As regards configurations for an autosomal locus, the 15 different possibilities reduce to the 7 mating types for which Fisher did an extensive analysis. So, as is after all obvious, Fisher's work based on the study of likeness of four genes is in the same spirit as Wright's study of the likeness of two genes, and it is clearly foolish to use one piece of work to belittle the other.

## 7. The progress of a finite monoecious population

Let the individuals in generation $t$ be denoted by

$$A^t_{11}A^t_{12},\ A^t_{21}A^t_{22},\ \cdots,\ A^t_{N1}A^t_{N2}. \tag{7.1}$$

Then with random mating, including selfing with the appropriate frequency, every gene in generation $t+1$ is an independent member from the probability array

$$A^{t+1}_u = \frac{1}{2N}\sum_{ij} A^t_{ij} = \frac{1}{M}\sum_{ij} A^t_{ij}, \qquad M = 2N. \tag{7.2}$$

Let

$P_{2,t} = F_t$ denote $P\{A^t_{ij} = A^t_{i'j'}\}$, which will be the same for all pairs $(ij) \neq (i'j')$,

$P_{3,t}$ denote $P\{A^t_{ij} = A^t_{i'j'} = A^t_{i''j''}\}$ which will be the same for all triples which are unequal,

$P_{4,t}$ denote $P\{A^t_{ij} = A^t_{i'j'} = A^t_{i''j''} = A^t_{i'''j'''}\}$, the same for all unequal quadruples.

It is, of course, well known that

$$P_{2,t+1} = \frac{1}{M} + \left(1 - \frac{1}{M}\right) P_{2,t}. \tag{7.3}$$

For triples of genes, there are $M^3$ possible choices; in $M$ of the choices the same gene will occur; in $3M(M-1)$ cases, one gene will occur twice and the other once; and in $M(M-1)(M-2)$ cases, three different genes of the previous generation will occur. So

$$P_{3,t+1} = \frac{1}{M^3}[M + 3M(M-1)P_{2,t} + M(M-1)(M-2)P_{3,t}] \tag{7.4}$$

or

$$P_{3,t+1} = \frac{1}{M^2} + 3\frac{(M-1)}{M^2}P_{2,t} + \frac{(M-1)(M-2)}{M^2}P_{3,t}. \tag{7.5}$$

In an obvious way,

$$P_{4,t+1} = \frac{1}{M^3} + 7\frac{(M-1)}{M^3}P_{2,t} + 6\frac{(M-1)(M-2)}{M^3}P_{3,t} + \frac{(M-1)(M-2)(M-3)}{M^3}P_{4,t}. \tag{7.6}$$

If now, we let $P^* = 1 - P$, we have the simple recurrence relations

(7.7)

$$\begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \end{pmatrix}_{t+1} = \begin{bmatrix} 1 - \frac{1}{M} & 0 & 0 \\ 3\frac{(M-1)}{M^2} & \frac{(M-1)(M-2)}{M^2} & 0 \\ 7\frac{(M-1)}{M^3} & 6\frac{(M-1)(M-2)}{M^3} & \frac{(M-1)(M-2)(M-3)}{M^3} \end{bmatrix} \begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \end{pmatrix}_t,$$

where the subscript represent the generation. The roots are obviously

$$1 - \frac{1}{M}, \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right), \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right)\left(1 - \frac{3}{M}\right) \tag{7.8}$$

or if we let

$$\delta_i = 1 - \frac{i}{M}, \qquad i = 1, 2, \cdots, M-1, \tag{7.9}$$

the roots are $\delta_1$, $\delta_1\delta_2$, $\delta_1\delta_2\delta_3$. The matrix of the recurrence equations then becomes

$$G = \begin{bmatrix} \delta_1 & 0 & 0 \\ \frac{3\delta_1}{M} & \delta_1\delta_2 & 0 \\ \frac{7\delta_1}{M^2} & \frac{6\delta_1\delta_2}{M} & \delta_1\delta_2\delta_3 \end{bmatrix}, \tag{7.10}$$

which has a particularly simple form.

The relation between $P_2^*$, $P_3^*$, $P_4^*$ in successive generations is then given succinctly by

$$\theta_{t+1} = \begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \end{pmatrix}_{t+1} = G \begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \end{pmatrix}_t = G\theta_t. \tag{7.11}$$

To determine the vector $\theta$ for any generation we find a matrix $C$ such that

$$CGC^{-1} = \text{diag}\,(\delta_1,\ \delta_1\delta_2,\ \delta_1\delta_2\delta_3). \tag{7.12}$$

With some elementary manipulation, it turns out that

$$C^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{3}{M(1-\delta_2)} & 1 & 0 \\ \dfrac{7+11\delta_2}{M^2(1-\delta_2)(1-\delta_2\delta_3)} & \dfrac{6}{M(1-\delta_3)} & 1 \end{bmatrix} \tag{7.13}$$

and

$$C = \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{-3}{M(1-\delta_2)} & 1 & 0 \\ \dfrac{11+7\delta_3}{M^2(1-\delta_3)(1-\delta_2\delta_3)} & \dfrac{-6}{M(1-\delta_3)} & 1 \end{bmatrix}. \tag{7.14}$$

It follows that

$$G^t = C^{-1} \begin{bmatrix} \delta_1^t & 0 & 0 \\ 0 & (\delta_1\delta_2)^t & 0 \\ 0 & 0 & (\delta_1\delta_2\delta_3)^t \end{bmatrix} C, \tag{7.15}$$

and in fact,

$$\begin{aligned} G^t = \delta_1^t & \begin{bmatrix} 1 & 0 & 0 \\ \dfrac{3}{M(1-\delta_2)} & 0 & 0 \\ \dfrac{7+11\delta_2}{M^2(1-\delta_2)(1-\delta_2\delta_3)} & 0 & 0 \end{bmatrix} \\ + (\delta_1\delta_2)^t & \begin{bmatrix} 0 & 0 & 0 \\ \dfrac{-3}{M(1-\delta_2)} & 1 & 0 \\ \dfrac{-18}{M^2(1-\delta_2)(1-\delta_3)} & \dfrac{6}{M(1-\delta_3)} & 0 \end{bmatrix} \\ + (\delta_1\delta_2\delta_3)^t & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \dfrac{11+7\delta_3}{M^2(1-\delta_3)(1-\delta_2\delta_3)} & \dfrac{-6}{M(1-\delta_3)} & 1 \end{bmatrix} \end{aligned} \tag{7.16}$$

or

$$G^t = \begin{bmatrix} \delta_1^t & 0 & 0 \\ \dfrac{3\delta_1^t - 3\delta_1^t\delta_2^t}{M(1-\delta_2)} & \delta_1^t\delta_2^t & 0 \\ \dfrac{(7+11\delta_2)\delta_1^t}{M^2(1-\delta_2)(1-\delta_2\delta_3)} & & \\ \dfrac{-18\delta_1^t\delta_2^t}{M^2(1-\delta_2)(1-\delta_3)} & \dfrac{6\delta_1^t\delta_2^t - 6\delta_1^t\delta_2^t\delta_3^t}{M(1-\delta_3)} & \delta_1^t\delta_2^t\delta_3^t \\ \dfrac{+(11+7\delta_3)\delta_1^t\delta_2^t\delta_3^t}{M^2(1-\delta_3)(1-\delta_2\delta_3)} & & \end{bmatrix}. \tag{7.17}$$

If then we measure the progress of likeness by descent of genes, we take the base population to consist of different genes and the initial values of $P_2^*$, $P_3^*$, $P_4^*$ to be unity. Therefore,

$$\begin{aligned} P_{2,t}^* &= \delta_1^t, \\ P_{3,t}^* &= \frac{3}{M(1-\delta_2)}\delta_1^t + \left[1 - \frac{3}{M(1-\delta_2)}\right]\delta_1^t\delta_2^t \\ &= \frac{3}{2}\delta_1^t - \frac{1}{2}\delta_1^t\delta_2^t, \qquad (7.18) \\ P_{4,t}^* &= \frac{7+11\delta_2}{M^2(1-\delta_2)(1-\delta_2\delta_3)}\delta_1^t + \left[\frac{6}{M(1-\delta_3)} - \frac{18}{M^2(1-\delta_2)(1-\delta_3)}\right]\delta_1^t\delta_2^t \\ &\quad + \left[1 - \frac{6}{M(1-\delta_3)} + \frac{11+7\delta_3}{M^2(1-\delta_3)(1-\delta_2\delta_3)}\right]\delta_1^t\delta_2^t\delta_3^t \\ &= \left(\frac{9-11/M}{5-6/M}\right)\delta_1^t - \delta_1^t\delta_2^t + \left(\frac{1-1/M}{5-6/M}\right)\delta_1^t\delta_2^t\delta_3^t. \end{aligned}$$

A pair of genes can have one of the two states: identical by descent with probability $P_2$ (called $F$ earlier in this paper); and nonidentical by descent $P_{1/1}$ ($P$ earlier). For a triple of genes the possibilities are all three alike with probability $P_3$, two alike and one different $P_{2/1}$, all different $P_{1/1/1}$. It is clear that

$$P_2 = P_3 + \frac{1}{3}P_{2/1}, \tag{7.19}$$

so

$$P_{2/1} = 3(P_2 - P_3) = 3(P_3^* - P_2^*), \tag{7.20}$$

and

$$\begin{aligned} P_{1/1/1} &= 1 - P_3 - P_{2/1} \\ &= 1 + 2P_3 - 3P_2 \\ &= 3P_2^* - 2P_3^*. \end{aligned} \tag{7.21}$$

In the case of a quadruple of genes, things are not as simple, and the probabilities we need for complete description are

(7.22) $$P_4,\ P_{3/1},\ P_{2/2},\ P_{2/1/1},\ \text{and}\ P_{1/1/1/1}.$$

It is clear that

(7.23) $$P_3 = P_4 + \frac{1}{4}P_{3/1},$$

so we know $P_{3/1}$ for any generation. The others are obtainable only with supplementary computations. We find that

$$P_{2/2}^{t+1} = \frac{3\delta_1}{M^2}P_{1/1}^t + \frac{2\delta_1\delta_2}{M}P_{2/1}^t + \delta_1\delta_2\delta_3 P_{2/2}^t,$$

(7.24)

$$P_{2/1/1}^{t+1} = \frac{6\delta_1\delta_2}{M}P_{1/1/1}^t + \delta_1\delta_2\delta_3 P_{2/1/1}^t.$$

Hence,

$$P_{2/2}^{t+1} = \frac{3\delta_1}{M^2}P_2^* + \frac{6\delta_1\delta_2}{M}(P_3^* - P_2^*) + \delta_1\delta_2\delta_3 P_{2/2}^t,$$

(7.25)

$$P_{2/1/1} = \frac{6\delta_1\delta_2}{M}(3P_2^* - 2P_3^*) + \delta_1\delta_2\delta_3 P_{2/1/1}^t.$$

It follows that the complete specification of the status of two, three, or four genes is given by the equation

(7.26)

$$\begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \\ P_{2/2} \\ P_{2/1/1} \end{pmatrix}_{t+1} = \begin{bmatrix} \delta_1 & 0 & 0 & 0 & 0 \\ \frac{3\delta_1}{M} & \delta_1\delta_2 & 0 & 0 & 0 \\ \frac{7\delta_1}{M^2} & \frac{6\delta_1\delta_2}{M} & \delta_1\delta_2\delta_3 & 0 & 0 \\ \frac{3\delta_1}{M^2} & \frac{-6\delta_1\delta_2}{M} & \frac{6\delta_1\delta_2}{M} & 0 & \delta_1\delta_2\delta_3 \\ & \frac{18\delta_1\delta_2}{M} & \frac{-12\delta_1\delta_2}{M} & 0 & 0 & \delta_1\delta_2\delta_3 \end{bmatrix} \begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \\ P_{2/2} \\ P_{2/1/1} \end{pmatrix}_t.$$

We note that consideration of the other possible configurations of four genes causes repetition of the root $\delta_1\delta_2\delta_3$. A little algebra will yield the $t$th iterate of the coefficient matrix.

The generalization of the above to the case of $k$plets of genes is direct, though there are complexities of algebra. Essentially, what is involved is the making of $k$ independent draws from an array of $M$ objects. There are $M^k$ possible results, and if we denote by $\begin{bmatrix} k \\ r \end{bmatrix}$, the number of unordered partitions of $k$ into $r$ parts, the number of results in which the $k$ draws arise from $r$ different objects is

(7.27) $$\begin{bmatrix} k \\ r \end{bmatrix} M(M-1)\cdots(M-r+1).$$

It follows that the recurrence equation for $P_2^*, P_3^*, \cdots, P_k^*$ is

(7.28)

$$\begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \\ \cdot \\ \cdot \\ \cdot \\ P_k^* \end{pmatrix}_{t+1} = \begin{bmatrix} \delta_1 & 0 & 0 & \cdots & 0 \\ \binom{3}{2}\frac{\delta_1}{M} & \delta_1\delta_2 & 0 & \cdots & 0 \\ \binom{4}{2}\frac{\delta_1}{M^2} & \binom{4}{3}\frac{\delta_1\delta_2}{M} & \delta_1\delta_2\delta_3 & \cdots & 0 \\ \cdot & \cdot & \cdot & & 0 \\ \cdot & \cdot & \cdot & & 0 \\ \cdot & \cdot & \cdot & & 0 \\ \binom{k}{2}\frac{\delta_1}{M^{k-2}} & \binom{k}{3}\frac{\delta_1\delta_2}{M^{k-3}} & \binom{k}{4}\frac{\delta_1\delta_2\delta_3}{M^{k-4}} & \cdots & \delta_1\delta_2\cdots\delta_{k-1} \end{bmatrix} \begin{pmatrix} P_2^* \\ P_3^* \\ P_4^* \\ \cdot \\ \cdot \\ \cdot \\ P_k^* \end{pmatrix}_t$$

or

$$\phi_{t+1} = H\phi_t. \tag{7.29}$$

It appears that the matrix $C$ such that

$$CHC^{-1} = \text{diagonal} \tag{7.30}$$

has a nice structure in terms of $\delta_1, \delta_2, \cdots$.

Some aspects of the progress of a finite dioecious population are being developed by the same approach, but I do not yet have the results.

## 8. *k*plets of genes in a finite haploid population with random viabilities

The arguments of previous sections may be combined. If we let $P_{3t}$ be the probability of three genes being identical by descent, we have

$$P_{3t+1} = \sum p_i^3 + 3 \sum_{\substack{i,i' \\ i \neq i'}} p_i^2 p_{i'} P_{2t} + \sum_{\substack{i,i',i'' \\ i \neq i' \neq i''}} p_i p_{i'} p_{i''} P_{3t}. \tag{8.1}$$

Let

$$\begin{aligned} \sum_i p_i^3 &= N\mu_3', \\ \sum_{\substack{i,i' \\ i \neq i'}} p_i^2 p_{i'} &= N^2\mu_{21}', \\ \sum_{i,i',i''} p_i p_{i'} p_{i''} &= N^3\mu_{111}', \end{aligned} \tag{8.2}$$

and

$$P_3^* = 1 - P_3. \tag{8.3}$$

Then,

$$P_{3(t+1)}^* = 3N^2\mu_{21}' P_{2t}^* + N^3\mu_{111}^* P_{3t}^*. \tag{8.4}$$

Also,

$$P_{2(t+1)}^* = (1 - N\mu_2') P_{2t}^*. \tag{8.5}$$

It follows that

$$P^*_{3(t+1)} - \gamma P^*_{2(t+1)} = 3N^2\mu'_{21}P^*_{2t} + N^3\mu_{111}P^*_{3t} - \gamma[1 - N\mu'_2]P^*_{2t} \tag{8.6}$$
$$= N^3\mu'_{111}[P^*_{3t} - \gamma P_{2t}]$$

if

$$N^3\mu'_{111}\gamma = \gamma(1 - N\mu'_2) - 3N^2\mu'_{21} \tag{8.7}$$

or if

$$\gamma = \frac{3N^2\mu'_{21}}{1 - N\mu'_2 - N^3\mu'_{111}}. \tag{8.8}$$

Hence, $P_3 - \gamma P_2$ decreases in the ratio $N^3\mu'_{111}$ in each generation.

The extension to $k$plets of genes is clear, though the algebra becomes very tedious. In general, there is a form

$$P^*_k - \gamma_{k,k-1}P^*_{k-1} - \gamma_{k,k-2}P^*_{k-2} - \cdots - \gamma_{k,2}P^*_2 \tag{8.9}$$

of the probabilities which decreases in each generation on the ratio $N^k\mu'_{11\cdots1}$, where

$$N^k\mu'_{11\cdots1} = \sum_{\substack{i_1,i_2,\cdots,i_k \\ \text{unequal}}} p_{i_1}p_{i_2}\cdots p_{i_k}. \tag{8.10}$$

If all $p_i$ equal $1/N$, the ratio is $N_{(k)}/N^k$ as we already knew.

It may be noted in passing that the incorporation of mutation, at least in one form, presents few difficulties.

## 9. Concluding remarks

The determination of the consequences of Mendelism with random mating whether in a haploid population, a monoecious or dioecious population is a matter of algebraic computation. By so characterizing the work, I do not wish to be derogatory. The computations can become very excessive and very tedious, particularly if one wishes to develop the probability status of the whole population. In one way or another, this involves the determination of the roots and both left and right eigenvectors of the particular probability transition matrix that is a consequence of the whole system. This proves to be very difficult, though Drs. Karlin and McGregor have made notable progress. The purpose of my own presentation is to show that if one is satisfied with a partial picture of the progress of the population, one may well be able to do so by using the concept of identity of genes by descent.

From a genetic viewpoint, I am inclined to think that a working out of limiting distributions using only the largest root, which is frequently of the form $(1 - 1/N)$ are of interest only mathematically. The common case is that the roots are very close as $(1 - 1/N)$, $(1 - 1/N)(1 - 2/N)$, $(1 - 1/N)(1 - 2/N)$ $(1 - 3/N)$, and so on. It would seem that the length of time required for the situation to be describable reasonably in terms of the largest root, would be such that the model can no longer be trusted unless mutation is included. If

the same mutation can occur more than once, I suspect that a solution based on only the largest root can be quite misleading at least for populations of the size that occur in Nature. This situation with regard to the regular inbreeding systems commonly used would appear to be somewhat different. It is evident that many difficult problems in the dynamics of Mendelian populations remain.

It has been brought to my attention that some work on similar lines was done contemporaneously be Michel Gillois in a thesis entitled, "La relation d'identité en génétique," Faculty of Sciences of the University of Paris [8].

## REFERENCES

[1] R. A. FISHER, *The Theory of Inbreeding*, Edinburgh, Oliver and Boyd, 1949.
[2] D. L. HARRIS, "Genotypic covariances between inbred relatives," *Genetics*, Vol. 50 (1964), pp. 1319–1348.
[3] O. KEMPTHORNE, "The correlations between relatives in a random mating population," *Proc. Roy. Soc. Ser. B*, Vol. 143 (1954), pp. 103–113.
[4] ———, *An Introduction to Genetic Statistics*, New York, Wiley, 1957.
[5] G. MALÉCOT, *Les Mathématiques de L'Hérédité*, Paris, Masson, 1948.
[6] P. A. P. MORAN, "The theory of some genetical effects of population subdivision," *Austral. J. Biol. Sci.*, Vol. 12 (1959), pp. 109–116.
[7] S. WRIGHT, "Systems of mating I–V," *Genetics*, Vol. 6 (1921), pp. 11–178.
[8] M. GILLOIS, "La relation d'identité en génétique," unpublished thesis, Faculty of Sciences, University of Paris, 1964.