AN ALGEBRA FOR THEORETICAL GENETICS

By

Claude Elwood Shannon

B.S., University of Michigan

1936

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

From The

Massachusetts Institute of Technology

1940

Signature of Author.....

Department of Mathematics, April 15, 1940

Signature of Professor

in Charge of Research.

Signature of Chairman of Department

Committee on Graduate Students....

Table of Contents

I. Introduction

In this paper an attempt will be made to develop an algebra especially suited to problems in the dynamics of Mendelian populations. Many of the results presented here are old in the theory of genetics, but are included because the method of proof is novel, and usually simpler and more general than those used previously.

For the benefit of readers who are not familiar with modern genetics theory, we will first give a brief factual summary of those parts of it which are necessary for our work. Although all parts of the theory have not been incontestably established, still it is possible for our purposes, to act as though they were, since the results obtained are known to be the same as if the simple representation which we give were true. Hereafter we shall speak therefore as though the genes actually exist and as though our simple representation of hereditary phenomena were really true, since so far as we are concerned, this might just as well be so. We will omit from consideration mutations and phenomena in the sex chromosomes.

Hereditary traits are transmitted by small elements called genes. These genes are carried in rodlike bodies known as chromosomes, a large number of genes lying side by side along the length of a chromosome. Chromosomes occur in pairs and an individual obtains one chromosome of each pair from his mother and the other from his fa-

ther.

By the genetic constitution of an individual we mean the kind and location of the genes which he possesses. If we represent the different genes by letters, then we may write a genetic formula for an individual. Thus considering two chromosome pairs and four gene positions in each chromosome, an individual might have the formula:

$$\begin{matrix} A_1\ B_1\ C_3\ D_5 & E_4\ F_1\ G_6\ H_1 \\ A_3\ B_1\ C_4\ D_3 & E_4\ F_2\ G_6\ H_2 \end{matrix} \qquad (1)$$

Here the series $A_1\ B_1\ C_3\ D_5$ represents one chromosome, with $A_3\ B_1\ C_4\ D_3$ the corresponding one of the first pair. A_1, B_1, C_3, D_5, A_3, B_1, C_4, D_3 are the genes lying in the positions under consideration. $E_4\ F_1\ G_6\ H_1$ and $E_4\ F_2\ G_6\ H_2$ are the two chromosomes of the second pair. We will sometimes write a genetic formula in one line. Thus (1) would be written:

$$A_1\ A_3\ B_1\ B_1\ C_3\ C_4\ D_5\ D_3\ E_4\ E_4\ F_1\ F_2\ G_6\ G_6\ H_1\ H_2$$

alternate letters being taken from the top and bottom lines of (1).

There is no essential ordering of chromosomes in a pair. That is to say that the top and bottom lines of the formula for a chromosome pair may be inverted and still represent the same individual. Thus the formula (1) is identical, for example with the following:

$A_3\ B_1\ C_4\ D_3 \qquad E_4\ F_1\ G_6\ H_1$
$A_1\ B_1\ C_3\ D_5 \qquad E_4\ F_2\ G_6\ H_2$

in which we have inverted the first pair.

Certain simple traits are controlled by only one pair of genes lying at analogous points in corresponding chromosomes. Two such corresponding points in a chromosome pair are known as a gene locus, and the different genes which may occupy one locus are known as allelomorphs or more shortly as alleles. In our example (1) the positions occupied by genes C_3 and C_4 constitute a locus. We shall adopt the convention that allelomorphic genes shall have the same base letter with different subscripts. Thus C_1, C_2, C_3, C_4, C_5 represent five alleles. A C gene can only occur in the locus corresponding to C genes.

The appearance of an individual depends only on the kinds of genes, not on their positions. Thus an individual with the formula

$A_1\ B_1\ C_4\ D_3 \qquad E_4\ F_1\ G_6\ H_2$
$A_3\ B_1\ C_3\ D_5 \qquad E_4\ F_2\ G_6\ H_1$

would appear (insofar as the characteristics controlled by these genes are concerned) the same as (1). He would, however, breed differently as will appear later. Two such individuals are said to be phenotypically the same with respect to these characteristics. They are genotypically different; they have different genetic formulae with respect

to these loci. Such a situation can occur in a different way. In garden peas there are two alleles which control the size of the plant. These genes we may represent by A_1 and A_2. If a plant has two A_1 genes, it will be tall. If it has two A_2 genes, it will be a dwarf. A plant with one A_1 gene and one A_2 gene is tall, since the gene for tallness (A_1) is, as we say, dominant over the recessive gene (A_2) for shortness. Thus, A_1A_1 plants and A_1A_2(or A_2A_1) plants are phenotypically the same but genotypically different with respect to tallness.

As we stated above, an individual receives one chromosome of each pair from the corresponding pair possessed by his mother and the other from that of his father. Let us now consider a pair of chromosomes possessed by a parent. In case a phenomenon known as cross-over does not occur in the chromosome pair under consideration, an offspring receives an entire chromosome selected at random from these two. We say that the genes in the chromosome are linked together meaning that they tend to be transmitted as a body. Genes located close together in the same chromosome are closely linked; the greater the distance between them, the weaker the linkage. Let us suppose that an individual has the genetic formula represented as follows

$$A_1\ B_2\ C_3\ D_3\ E_4\ F_1\ G_1$$
$$A_2\ B_2\ C_2\ D_1\ E_6\ F_1\ G_2$$

for a pair of corresponding chromosomes. Now, as we have said, in case crossover does not occur, an offspring of this individual will receive either the series

$$A_1\ B_2\ C_3\ D_3\ E_4\ F_1\ G_1 \text{ or } A_2\ B_2\ C_2\ D_1\ E_6\ F_1\ G_2$$

and he is equally likely to receive either of these. However, it may happen that a crossover occurs between these chromosomes. If this crossover occurred, for instance, between the C and D loci, he would receive either

$$A_1\ B_2\ C_3\ D_1\ E_6\ F_1\ G_2 \text{ or } A_2\ B_2\ C_2\ D_3\ E_4\ F_1\ G_1.$$

There is a definite probability that a crossover will occur between any two gene loci. Determining the relative positions of genes in a chromosome according to such a probability scale is known as <u>mapping</u> the chromosome. This has been carried out quite extensively for Drosophila and to a lesser extent for some other plants and animals. The <u>map distance</u> between two loci a and b may be defined as follows. Let x measure the actual physical distance along the chromosome and let $p(x)$ be the probability that a crossover occurs between the points x and $x+dx$, providing it is known that no other crossover occurs near to the point x. This last restriction is necessary due to a phenomenon known as <u>interference</u> in which a crossover at one point hinders nearby crossovers. The map distance is then given by

$$\int_a^b p(x)\,dx.$$

The recombination value of two loci is the probability of an odd number of crossovers between these loci. For small distances the probability of more than one crossover is a second order term and the map distance is nearly equal to the recombination value, and both approximate the probability of one crossover between the loci.

If the two genes in a certain locus are identical, the individual is said to be homozygous in this factor. Otherwise he is heterozygous. The individual (1) is thus homozygous in the B, E, and G factors and heterozygous in all others.

A simple example will perhaps help to clarify these notions. Suppose that two gene loci are under consideration. There are three allelomorphic genes forthe first locus, A_1, A_2, A_3 ; the second locus has four alleles, B_1, B_2, B_3, B_4. The recombinationvalue for these two loci is 1/4. An individual with the genetic formula:

$$\begin{matrix} A_1 & B_4 \\ A_3 & B_2 \end{matrix}$$

is mated with an individual having the formula:

$$\begin{matrix} A_2 & B_4 \\ A_2 & B_1 \end{matrix}$$

What is the probability that an offspring of this mating will have the formula

$$\begin{matrix} A_3 & B_4 \\ A_2 & B_1 \end{matrix}$$

Stated another way, what fraction of the offspring population should be expected to have this formula?

Evidently an offspring must obtain the $A_3\ B_4$ chromosome from the first parent. The probability that he will get an A_3 gene from this parent is 1/2 since A_3 and A_1 are equally likely. If he gets this A_3 gene, the probability that he will also get a B_4 gene from this parent is 1/4, the recombination value, since A_3 and B_4 are in opposite chromosomes. Thus, the probability that both events occur is $1/2 \cdot 1/4 = 1/8$. Now our offspring must obtain A_2 and B_1 from the second parent. He will certainly obtain an A_2 since both genes in this locus are of this type. The chance that he obtains a B_1 is 1/2, since B_1 and B_3 are equally likely. The probability of the combination is therefore also 1/2. Our final answer is, since the events are independent, $1/8 \cdot 1/2 = 1/16$. If we had asked what fraction would be of the type

$$\begin{matrix} A_3 & B_2 \\ A_2 & B_2 \end{matrix}$$

then in place of multiplying by the recombination value

1/4, we would multiply by 1-1/4=3/4 since this is the probability that a crossover <u>does not</u> occur between the loci.

II. Notation

To non-mathematicians we point out that it is a commonplace of modern algebra for the symbols to represent concepts other than numbers, and frequently therefore not to obey all the laws governing numbers. Such is the case in vector algebra, the theory of groups, rings, matrix algebra, in symbolic logic, tensor analysis, etc. In the particular algebra we construct for genetics theory the symbols represent Mendelian populations, and stand for a large group of numbers which describe the genetic constitution of the population. Addition and multiplication are defined to mean simple combination and cross breeding respectively, and it is shown that nearly all the laws of ordinary numerical algebra hold here. One interesting exception is the associative law of multiplication. It is not in general true that

$$(\lambda \times \mu) \times \nu = \lambda \times (\mu \times \nu)$$

Much of the power and elegance of any mathematical theory depends on use of a suitably compact and suggestive notation, which nevertheless completely describes the concepts involved. We will employ an index notation somewhat similar to that of the tensor calculus, which has proven so useful in differential geometry and in relativity

theory. Because the notation employed is so basic to our work we will first explain the meaning of indexed symbols.

Consider, for example, the symbol

$$\lambda_{j\,k}^{h\,i} \qquad (2)$$

Here λ is the base letter and h, i, j, and k are indices. Each index has a certain specific range of variation and the different indices may vary independently and even have different ranges of variation. In our work two indices in the same vertical column, such as h and j in (2), will always have the same range, but vary independently over this range. Thus h and j might have the range of values 1, 2, and 3 while i and k have the range 1, 2, 3, 4...9.

When the indices of (1) take on specific values, e.g. h = 1, j = 3, i = 5, k = 5, the symbol

$$\lambda_{3\,5}^{1\,5}$$

represents a number. Symbol (2) then stands for a whole group of numbers, one for each combination of values of the indices; however, it should not be thought of as a group of separate numbers, but rather as a single entity having components whose values are the different numbers of the array.

When we think of an indexed symbol as representing a whole array of numbers and the indices as variables which assume any of the values in their ranges we say the indices are <u>live</u> or <u>variable</u>. Occasionally, however, it is desirable to think of $\lambda_{j\kappa}^{hi}$ (say) as representing a certain specific one of the components. Thus we may set $h=1$, $i=3$, $j=2$, $k=3$. We say then that we have <u>fixed</u> or <u>killed</u> the indices at these values; they become <u>dead</u> indices. Also we sometimes wish to think of the indices as fixed at some value which is perfectly arbitrary. Without any change of notation we use $\lambda_{j\kappa}^{hi}$ to represent an arbitrary component rather than the whole set of components. In such a case fixing the indices is purely subjective.

In an equation, although indices represented by different letters may vary independently, a specific letter e.g. h, must not take on different values in different places. Thus the <u>sum</u> of two indexed symbols

$$\lambda_{j\kappa}^{hi} + \mu_{j\kappa}^{hi} \qquad (3)$$

is an indexed symbol, say $\nu_{j\kappa}^{hi}$, whose components are the sums of the corresponding components of $\lambda_{j\kappa}^{hi}$ and $\mu_{j\kappa}^{hi}$. For example

$$\nu_{11}^{11} = \lambda_{11}^{11} + \mu_{11}^{11}$$

$$\nu_{35}^{12} = \lambda_{35}^{12} + \mu_{35}^{12}.$$

On the other hand, if

$$\nu^{k\,i}_{j\,\kappa} = \lambda^{k\,i}_{j\,\kappa} + \mu^{l\,\kappa}_{j\,i}$$

then $\nu'^{\,2}_{3\,5} = \lambda'^{\,2}_{3\,5} + \mu'^{\,5}_{3\,2}$ etc.

An equation in indexed symbols stands therefore for a large number of ordinary equations, one for each combination of values of the variable indices.

Ordinary multiplication of indexed symbols will be indicated by juxtaposition, e.g. $R^i_j \lambda^i_\kappa$, $R^k_i \lambda^l_i$, $\lambda^k_i \mu^j_\kappa$ Ordinary multiplication means numerical multiplication of the components indicated, and results therefore in another indexed symbol. The multiplications above would result respectively in symbols with indices as follows:

$$\tau^i_{j\kappa}, \quad \rho^k_i, \quad \sigma^{k\,j}_{i\,\kappa}$$

where

$$\tau'_{23} = R'_2 \lambda'_3, \quad \tau^2_{33} = R^2_3 \lambda^2_3$$
$$\rho^1_2 = R^1_2 \lambda'_2, \quad \rho^3_3 = R^3_3 \lambda^3_3$$
$$\sigma'^{\,2}_{43} = \lambda'_4 \mu^2_3$$

etc. There are always as many variable indices in a product as there are different variable indices in the factors. Thus $R^i_j \lambda^i_\kappa$ has three independent variable indices i, j, k and hence the product $\tau^i_{j\kappa}$ has three indices.

An important operation in indexed symbols is summation on one or more indices. This is so common in our work that

we indicate it by replacing the index in question by a large dot. Thus suppose the index h has a range of variation of 1 to 3 and i, a range 1 to 5, then

$$\lambda_{jk}^{\cdot i} = \sum_{h=1}^{3} \lambda_{jk}^{hi} = \lambda_{jk}^{1i} + \lambda_{jk}^{2i} + \lambda_{jk}^{3i}$$

$$\lambda_{jk}^{\cdot\cdot} = \sum_{h=1}^{3}\sum_{i=1}^{5} \lambda_{jk}^{hi} = \lambda_{jk}^{11} + \lambda_{jk}^{12} + \lambda_{jk}^{13} + \lambda_{jk}^{14} + \lambda_{jk}^{15}$$
$$+ \lambda_{jk}^{21} + \lambda_{jk}^{22} + \lambda_{jk}^{23} + \lambda_{jk}^{24} + \lambda_{jk}^{25}$$
$$+ \lambda_{jk}^{31} + \lambda_{jk}^{32} + \lambda_{jk}^{33} + \lambda_{jk}^{34} + \lambda_{jk}^{35}$$

Most of our indexed symbols will represent populations. Suppose we are considering two different Mendelian factors. Let the first have two alleles, A_1 and A_2, and suppose the second factor has three; B_1, B_2, and B_3. Then any population may be divided into 21 genetically different groups, having the genetic formulae

(1) $A_1\ A_1\ B_1\ B_1$
(2) $A_1\ A_1\ B_1\ B_2$
(3) $A_1\ A_1\ B_1\ B_3$
(4) $A_1\ A_1\ B_2\ B_2$
(5) $A_1\ A_1\ B_2\ B_3$
(6) $A_1\ A_1\ B_3\ B_3$
(7) $A_2\ A_2\ B_1\ B_1$
(8) $A_2\ A_2\ B_1\ B_2$
(9) $A_2\ A_2\ B_1\ B_3$
(10) $A_2\ A_2\ B_2\ B_2$
(11) $A_2\ A_2\ B_2\ B_3$
(12) $A_2\ A_2\ B_3\ B_3$
(13) $A_1\ A_2\ B_1\ B_1$
(14) $A_1\ A_2\ B_1\ B_2$
(15) $A_1\ A_2\ B_1\ B_3$
(16) $A_1\ A_2\ B_2\ B_2$
(17) $A_1\ A_2\ B_2\ B_3$
(18) $A_1\ A_2\ B_3\ B_3$
(19) $A_1\ A_2\ B_2\ B_1$
(20) $A_1\ A_2\ B_3\ B_1$
(21) $A_1\ A_2\ B_3\ B_2$

This population would be represented by the symbol λ_{jk}^{hi}.

The indices h and j correspond to the first locus and since there are two alleles for this factor they each have a range of variation of 1 to 2. The second factor has three alleles and correspondingly i and k range over the values 1, 2, 3. Now, thinking of h, i, j, and k as fixed or dead we define the components of λ_{jk}^{hi} in the following manner. If h = j and i = k then $\lambda_{jk}^{hi} = \lambda_{hi}^{hi} =$ the fraction of the population with the genetic formula $A_h A_h B_i B_i$. Thus λ_{13}^{13} is the fraction of the population of the type $A_1 A_1 B_3 B_3$. If $h \neq j$, or $i \neq k$, or both, then λ_{jk}^{hi} represents <u>one half</u> the fraction of the population having the formula $A_h A_j B_i B_k$ or what is the same thing $A_j A_h B_k B_i$. Thus λ_{23}^{13} and λ_{12}^{13} are <u>one half</u> the fractions having the respective formulas $A_1 A_2 B_3 B_3$ and $A_2 A_1 B_3 B_2$.

We shall use Greek letters as base letters for populations, and in general, then, the symbol

$$\lambda_{j_1 j_2 \cdots j_s}^{i_1 i_2 \cdots i_s} \qquad (4)$$

represents a population in which <u>s</u> gene loci are under consideration. The first column ofindices, i_1 and j_1, corresponds to the first factor under consideration, the second column to the next factor, etc. Each factor may have an arbitrary number of alleles, and the linkage between any two may be of any value including 50% or random assortment. In case $i_1 = j_1, i_2 = j_2, \ldots, i_s = j_s$ then

$\lambda^{i_1 i_2 \dots i_s}_{i_1 i_2 \dots i_s}$ is the fraction of the population having the formula $(A_{i_1}A_{i_1}B_{i_2}B_{i_2}\dots S_{i_s}S_{i_s})$. If these equalities are not all true then $\lambda^{i_1 i_2 \dots i_s}_{j_1 j_2 \dots j_s}$ is 1/2 the fraction having the formulae $(A_{i_1}A_{j_1}B_{i_2}B_{j_2}\dots S_{i_s}S_{j_s})$. It is helpful in using this notation to note the close connection between the two rows of letters in the indices and the two rows of genes in the chromosomes. The analogue is more than superficial, for we will later show how crossing over, say between the second and third loci, is connected, in this notation, with the symbol $\lambda^{i\,i\,\cdot}_{\cdot\,\cdot\,j}$.

III. Fundamental Theorems

There are two fundamental manipulation laws which we present as theorems because of their importance, although both are almost obvious.

Theorem I.

$$\lambda^{i_1 i_2 \cdots i_s}_{j_1 j_2 \cdots j_s} = \lambda^{j_1 j_2 \cdots j_s}_{i_1 i_2 \cdots i_s} \qquad (5)$$

That is, inverting the upper and lower rows of indices of a population gives an identical population. This is evident from the meaning of the symbols, since a genetic formula $(A_{i_1}A_{j_1}B_{i_2}B_{j_2}\cdots S_{i_s}S_{j_s})$ is identical with the formula $(A_{j_1}A_{i_1}B_{j_2}B_{i_2}\cdots S_{j_s}S_{i_s})$. This inversion of indices may be carried out independent of the location of the gene loci in question. However, if it is known that certain of the loci are in one chromosome pair, and none of the others are in this pair, further identities will hold. Namely, we may invert the indices corresponding to this chromosome pair and leave the others intact without changing the meaning of the symbol. Thus, if in the population $\lambda^{h\,i\,j}_{k\,l\,m}$, the first two loci are in one chromosome pair, and the third in another, we have:

$$\lambda^{h\,i\,j}_{k\,l\,m} = \lambda^{h\,i\,m}_{k\,l\,j} = \lambda^{k\,l\,m}_{h\,i\,j} = \lambda^{k\,l\,j}_{h\,i\,m}.$$

Theorem II.

$$\lambda^{\cdot\cdot\cdots\cdot}_{\cdot\cdot\cdots\cdot} = 1 \qquad (6)$$

That is summation of a population on all indices gives the result one. Obviously the sum of all the fractional parts of the population is unity. Now those parts which we have divided by 2 come into the summation (6) twice, corresponding to an inversion of the upper and lower indices. Thus with the population λ^{hi}_{jk} the term λ^{ii}_{kk} appears twice in the summation $\lambda^{\cdot\cdot}_{\cdot\cdot}$, once as λ^{ii}_{kk} and once as λ^{kk}_{ii}, which are equal by Theorem I.

The significance of summation on a fewer number of indices is also of considerable importance. Consider the population λ^{hi}_{jk}. Summation on k gives $\lambda^{hi}_{j\cdot}$, a symbol with three variable or "live" indices. The reader may verify that if $h \neq j$, $\lambda^{hj}_{j\cdot}$ represents one half the fraction of λ^{hi}_{jk} which have the formula $A_h A_j B_i$ - where the - may be any gene. If $h = j$ then $\lambda^{hi}_{h\cdot}$ represents 1/2 the fraction having the formula $A_h A_j B_i$ - plus one half the fraction of the type $A_h A_h B_i B_i$.

Summing on both k and i we have $\lambda^{h}_{j\cdot\cdot}$ and this may be shown to have exactly the same meaning as λ^{h}_{j} where h and j refer to the same gene locus in each case. That is, summing on a pair of vertical indices is equivalent to eliminating this locus from consideration. Summing on two horizontal indices, j and k, gives $\lambda^{hi}_{\cdot\cdot}$, a two index

symbol, whose components are the fractions of all chromosomes in the population in which genes A_h and B_i both appear. Likewise $\lambda^{h\,\cdot}_{\cdot\,k}$ represents the fraction of all chromosome pairs in which A_h is in one and B_k in the other.

Summation on three indices $\lambda^{h\,\cdot}_{\cdot\,\cdot}$ gives a one index symbol and its components are the gene frequencies of A_1, A_2...in the general population. That is $\lambda^{1\,\cdot}_{\cdot\,\cdot}$ is the sum of the fraction of homozygotes in A_1 and half the heterozygotes having one A_1 gene.

The reader will easily generalize these statements for more complex cases. For easy reference we summarize the above remarks in the following proposition.

Theorem III.

1. If $i_1 = j_1$, $i_2 = j_2$, . . . , $i_{s-1} = j_{s-1}$ then $\lambda^{i_1 i_2 \cdots i_s}_{j_1 j_2 \cdots \cdot}$ is one half the fraction of the population of the type $A_{i_1}A_{i_1}\cdots S_{i_s}-$ plus half the fraction of the type $A_{i_1}A_{i_1}\cdots S_{i_s}S_{i_s}$. If the conditions above do not hold then it represents one half the population of the type $A_{i_1}A_{j_1}B_{i_2}B_{j_2}\cdots S_{i_s}-$.

2. Summation on two vertical indices is equivalent to eliminating the corresponding locus from consideration.

3. Summing on one index in each column gives the fractions of the chromosome pairs in which alleles with the remaining indices appear.

4. The fraction of $\lambda^{i_1 \cdots i_s}_{j_1 \cdots j_s}$ having at least one A_h gene is given by

$$2\lambda^{h\,\cdot\cdots\cdot}_{\cdot\,\cdot\cdots\cdot} - \lambda^{h\,\cdot\cdots\cdot}_{h\,\cdot\cdots\cdot}$$

In order to present a rigorous mathematical development it is convenient to consider symbols whose components are not all positive real numbers lying between zero and one. Of course, such a symbol cannot represent an actual group of individuals, but in some cases it is possible to solve problems using these symbols and get an actual population for the final answer. We shall speak of the symbol as a population in either case, but if the symbol <u>does</u> correspond to a possible group of individuals we shall say the population is <u>realizable</u>. If the components are all real numbers we will say the population is <u>real</u>. The use of unrealizable populations also adds elegance and generality to some of our later theorems.

We now introduce an operation between two populations to be known as cross multiplication and written:

$$\lambda^{hi}_{jk} \times \mu^{hi}_{jk}$$

This will be defined in such a way that the cross product of two realizable populations represents the expected offspring if the population in question are cross mated at random. We will first give the definition for the case of two linked factors. Let the probability of <u>zero</u> or an <u>even</u> number of crossovers between the two factors be $p_0 = 1-p_1$, with p_1, then, the probability of an odd number (i.e the recombination value). We define the

cross product of λ^{hi}_{jk}, μ^{hi}_{jk} as follows

$$\nu^{hi}_{jk} = \lambda^{hi}_{jk} \times \mu^{hi}_{jk}$$
$$= \frac{1}{2}\left[p_0 \lambda^{hi}_{\cdot\cdot} + p_1 \lambda^{h\cdot}_{\cdot i}\right]\left[p_0 \mu^{jk}_{\cdot\cdot} + p_1 \mu^{j\cdot}_{\cdot k}\right]$$
$$+ \frac{1}{2}\left[p_0 \lambda^{jk}_{\cdot\cdot} + p_1 \lambda^{j\cdot}_{\cdot k}\right]\left[p_0 \mu^{hi}_{\cdot\cdot} + p_1 \mu^{h\cdot}_{\cdot i}\right] \quad (8)$$

To prove that this represents the expected offspring population we use the idea of gene-pair frequencies, a generalization of the idea of gene frequencies. We wish to determine the probability that an offspring of mating λ with μ will have the genetic formula $A_h A_j B_i B_k$. One way an offspring may obtain this formula is to get $A_h B_i$ from λ and $A_j B_k$ from μ. This corresponds to the term

$$\left[p_0 \lambda^{hi}_{\cdot\cdot} + p_1 \lambda^{h\cdot}_{\cdot i}\right]\left[p_0 \mu^{jk}_{\cdot\cdot} + p_1 \mu^{j\cdot}_{\cdot k}\right] \quad (9)$$

in the equation (8). Each of the terms of (9) is a gene-pair frequency. The first is the frequency of the gene pair $A_h B_i$ in the population λ. The second is the frequency of the gene-pair $A_j B_k$ in μ. The product is then the probability of obtaining an offspring $A_h A_j B_i B_k$ by the method described.

Gene-pair frequency here means the frequency with which this pair of genes occurs together in the same chromosome <u>after</u> crossovers have taken place. The term

$$\left[p_0 \lambda^{hi}_{\cdot\cdot} + p_1 \lambda^{h\cdot}_{\cdot i}\right] \quad (10)$$

actually represents this because, from Theorem III, $\lambda^{h\,i}_{\cdot\,\cdot}$ represents the frequency with which A_h and B_i appear together in the same chromosome before crossover, and p_0 is the probability that there is no crossover between these factors (or at most an even number) so that $p_0 \lambda^{h\,i}_{\cdot\,\cdot}$ is the probability of getting A_h and B_i together after crossovers if they start together. Likewise $p_1 \lambda^{h\,\cdot}_{\cdot\,i}$ is the probability of A_h and B_i ending together when they start in opposite chromosomes of a pair. Since these two possibilities are mutually exclusive, and collectively exhaustive, their sum (10) represents the gene-pair frequency of $A_h B_i$ in λ after crossovers. Similarly the second term of (9) is the gene-pair frequency of $A_j B_k$ in μ after crossover and the product of these two is the probability of an offspring getting the pair $A_h B_i$ from λ and $A_j B_k$ from μ. The only other way for an offspring to get the formula $A_h A_j B_i B_k$ is to get $A_j B_k$ from λ and $A_h B_i$ from μ. This corresponds in exactly the same way to the second term of (8), and we may add the probabilities since the events are mutually exclusive. All that remains to be explained in (8) is the factor 1/2. In case the equations $h = j$, $i = k$ do not both hold then for the components of the offspring population we want half of the fraction of individuals of this type in order to fit our previous definition of a population symbol, and hence this factor. If both these equalities are true then both terms

of (8) are identical and we may add getting

$$\omega^{hi}_{hi} = [p_0 \lambda^{hi}_{\cdot\cdot} + p_1 \lambda^{h\cdot}_{\cdot i}][p_0 \mu^{hi}_{\cdot\cdot} + p_1 \mu^{h\cdot}_{\cdot i}]$$

which is what we get by the derivation above in this case, since the two "different" possibilities become identical under this restriction and therefore (8) holds for all values of the indices.

For three factors the defining equation of the cross product is

$$\begin{aligned}\omega^{hij}_{klm} &= \lambda^{hij}_{klm} \times \mu^{hij}_{klm} \\ &= \tfrac{1}{2}\left[p_{00}\lambda^{hij}_{\cdot\cdot\cdot} + p_{01}\lambda^{hi\cdot}_{\cdot\cdot j} + p_{10}\lambda^{h\cdot\cdot}_{\cdot ij} + p_{11}\lambda^{h\cdot j}_{\cdot i\cdot}\right] \\ &\cdot \left[p_{00}\mu^{klm}_{\cdot\cdot\cdot} + p_{01}\mu^{kl\cdot}_{\cdot\cdot m} + p_{10}\mu^{k\cdot m}_{\cdot l\cdot} + p_{11}\mu^{k\cdot\cdot}_{\cdot\cdot\cdot}\right] \qquad (11) \\ &+ \tfrac{1}{2}\left[p_{00}\lambda^{klm}_{\cdot\cdot\cdot} + p_{01}\lambda^{kl\cdot}_{\cdot\cdot m} + p_{10}\lambda^{k\cdot m}_{\cdot l\cdot} + p_{11}\lambda^{k\cdot\cdot}_{\cdot\cdot\cdot}\right] \\ &\cdot \left[p_{00}\mu^{hij}_{\cdot\cdot\cdot} + p_{01}\mu^{hi\cdot}_{\cdot\cdot j} + p_{10}\mu^{h\cdot\cdot}_{\cdot ij} + p_{11}\mu^{h\cdot j}_{\cdot i\cdot}\right]\end{aligned}$$

In this equation p_{00} is the probability of an even number of crossovers between the first two genes and an even number between the second and third. If we wish to consider interference effects we cannot merely write $p_{00} = p_0 q_0$ with p_0 the probability of an even number of crosses between the first two loci and q_0 that for the second and third, since the events are not independent. However defining q_0 as the probability of an even number of crosses

between the second and third factors after it is known that an even number occurred between the first two, would make this valid. Similarly p_{01} is the probability that an even number of crosses occur between the first two factors and an odd number between the second two, etc. The method of formation of the formula is fairly obvious; note first that all permutations of 0, 1 are used on the coefficient p. Also a 1 corresponds to changing from one row of indices to another, while 0 corresponds to staying in the same row.

The proof of formula (11), and indeed the general case, is an easy generalization of the method used for two factors. It merely amounts to showing that a term such as

$$\left[p_{00}\lambda^{h\,i\,j}_{\cdot\,\cdot\,\cdot} + p_{01}\lambda^{h\,i\,\cdot}_{\cdot\,\cdot\,j} + p_{10}\lambda^{h\,\cdot\,\cdot}_{\cdot\,i\,j} + p_{11}\lambda^{h\,\cdot\,j}_{\cdot\,i\,\cdot}\right]$$

is the gene-triplet frequency for the set $A_h\ B_i\ C_j$ after crossovers in population $\lambda^{h\,i\,j}_{k\,l\,m}$. This the reader will readily verify.

For n linked genes the expression for the cross product of two populations will take the form:

$$\nu^{i_1\,i_2\cdots i_s}_{j_1\,j_2\cdots j_s} = \lambda^{i_1\,i_2\cdots i_s}_{j_1\,j_2\cdots j_s} \times \mu^{i_1\,i_2\cdots i_s}_{j_1\,j_2\cdots j_s}$$

$$= \frac{1}{2}\left\{\left[p_{00\ldots0}\,\lambda^{i_1 i_2 \ldots i_s}_{\bullet\,\bullet\,\ldots\,\bullet} + p_{10\ldots0}\,\lambda^{i_1 \bullet \ldots \bullet}_{\bullet\, i_2 \ldots i_s} + \ldots + p_{11\ldots1}\,\lambda^{i_1 \bullet i_3 \ldots}_{\bullet\, i_2 \bullet \ldots}\right]\right.$$

$$\left.\cdot\left[\text{same expression with } \lambda \text{ replaced by } \mu \text{ and } i_1, i_2, \ldots, i_s \text{ by } j, j_2, \ldots, j_s\right]\right\} + \frac{1}{2}\left\{\text{same pair of expressions with } \lambda \text{ and } \mu \text{ interchanged}\right\}.$$

Although we have spoken throughout as though the factors under consideration were linked and in the same chromosome, this is not necessary. Suppose that in equation (11) the first two genes have a recombination value $p_1 = 1-p_0$, and that they are located in a different chromosome from the third. Under these conditions it is easy to see that

$$p_{00} = p_{01} = 1/2p_0$$
$$p_{10} = p_{11} = 1/2p_1$$

Also we have

$$\lambda^{hij}_{klm} = \lambda^{him}_{klj} \quad ; \quad \mu^{hij}_{klm} = \mu^{him}_{klj}$$

From these equations (11) may be reduced to

$$\frac{1}{2}\left[p_0\,\lambda^{hij}_{\bullet\,\bullet\,\bullet} + p_1\,\lambda^{h\,\bullet\,j}_{\bullet\,i\,\bullet}\right]\left[p_0\,\mu^{klm}_{\bullet\,\bullet\,\bullet} + p_1\,\mu^{k\,\bullet\,m}_{\bullet\,l\,\bullet}\right]$$
$$+ \frac{1}{2}\left[p_0\,\lambda^{klm}_{\bullet\,\bullet\,\bullet} + p_1\,\lambda^{k\,\bullet\,m}_{\bullet\,l\,\bullet}\right]\left[p_0\,\mu^{hij}_{\bullet\,\bullet\,\bullet} + p_1\,\mu^{h\,\bullet\,j}_{\bullet\,i\,\bullet}\right]$$

So that the independence of factors merely simplifies the situation.

In case all three factors are independent

$$p_{00} = p_{01} = p_{10} = p_{11} = 1/4$$

and $\lambda^{hij}_{klm} = \lambda^{him}_{klj} = \lambda^{hlm}_{kij}$ etc.

so that (11) reduces to

$$\frac{1}{2}\lambda^{hij}_{\cdots}\mu^{klm}_{\cdots} + \frac{1}{2}\lambda^{klm}_{\cdots}\mu^{hij}_{\cdots}$$

We have proved, then, the fundamental

Theorem IV. The cross product of two realizable populations represents the expected offspring of these populations when cross mated at random.

Much of our work will now be the investigation of special cases of the general formulae given above. We note at once both from the mathematical definition and obvious genetic considerations the following

Theorem V.

$$\lambda^{i_1 \cdots i_s}_{j_1 \cdots j_s} \times \mu^{i_1 \cdots i_s}_{j_1 \cdots j_s} = \mu^{i_1 \cdots i_s}_{j_1 \cdots j_s} \times \lambda^{i_1 \cdots i_s}_{j_1 \cdots j_s}$$

That is to say, cross multiplication is commutative.

Let us now consider the case of a single factor, but with, however, any number of alleles. Then the cross product reduces to

$$\nu^{i}_{j} = \lambda^{i}_{j} \times \mu^{i}_{j} = \frac{1}{2}\left[\lambda^{i}_{\cdot}\mu^{j}_{\cdot} + \lambda^{j}_{\cdot}\mu^{i}_{\cdot}\right] \qquad (14)$$

from which we get the proposition:

Theorem VI. If $\lambda_{\cdot}^{i} = \mu_{\cdot}^{i}$ then $\lambda_{j}^{i} \times \sigma_{j}^{i} = \mu_{j}^{i} \times \sigma_{j}^{i}$ (15)

In other words if two populations have the same gene frequencies for all alleles of a factor, then they will have the same breeding characteristics when cross-mated with another population.

Theorem VII. If $v_{j}^{i} = \lambda_{j}^{i} \times \mu_{j}^{i}$ then $v_{\cdot}^{i} = \frac{1}{2}(\lambda_{\cdot}^{i} + \mu_{\cdot}^{i})$ (16)

This follows immediately from (14) on summing through on the index j and noting by Theorem II that $\lambda_{\cdot}^{\cdot} = \mu_{\cdot}^{\cdot} = 1$. This theorem shows that a gene frequency of a cross product is the arithmetic mean of the corresponding gene frequencies of the factors in the product.

We have already indicated in Part II how indexed symbols are added. When our symbols represent populations we shall consider addition only when the sum of the coefficients is unity. The purpose of this restriction is to keep all terms on an actual fractional basis and thus preserve the validity of our theorems. In general this causes little or no inconvenience, as merely dividing by the sum of the coefficients will always normalize in this sense.

We write

$$\lambda_{jk}^{hi} = R_1 \mu_{jk}^{hi} + R_2 v_{jk}^{hi} + \ldots + R_n \sigma_{jk}^{hi} \qquad (17)$$

where $\sum_{i=1}^{n} R_i = 1$, for the "sum" of the populations

$\mu, \nu, \ldots, \sigma$ in the fractional proportions $R_1, R_2 \ldots R_n$ Note that all terms of a sum must have the same indices (although sometimes an index may be changed in position in the same vertical column). This is part of a useful idea in indexed symbols known as index balance.

Index balance serves as a simple partial check on equations. If the indices do not balance in an equation, the equation is certainly wrong (in fact it is meaningless). The rules governing index balance for our work may be formulated as follows:

1. Each term in a sum must have the same indices.

2. Each side of an equation must have the same indices.

3. A product (ordinary or cross) has indices corresponding to each <u>different</u> <u>live</u> index appearing on any of the factors of the product. (See Part II).

Index balance applies only to live indices. There is no balance, for example, on the dead indices 0 and 1 on p_{01} in equation (11).

Addition of populations (17) is interpreted very simply as the population obtained by combining random samples of $\mu, \nu, \ldots, \sigma$ in the fractional proportions $R_1, R_2, \ldots, R_n$.

<u>Theorem VIII</u>. Cross multiplication is distributive on addition, e.g.

$$\lambda^{hi}_{jk} \times (R_1 \mu^{hi}_{jk} + R_2 \nu^{hi}_{jk}) = R_1 \lambda^{hi}_{jk} \times \mu^{hi}_{jk} + R_2 \lambda^{hi}_{jk} \times \nu^{hi}_{jk} \qquad (18)$$

We shall prove the theorem only for this simple case. The method of proof, however, is perfectly general and will apply with any number of indices and any number of terms in the sum. The left side of the equation is, by the definition (8):

$$\tfrac{1}{2}\left[p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}\right]\left[p_0(R_1\mu^{jk}_{\cdot\cdot} + R_2\nu^{jk}_{\cdot\cdot}) + p_1(R_1\mu^{j}_{\cdot k} + R_2\nu^{j}_{\cdot k})\right]$$

$$+\tfrac{1}{2}\left[p_0\lambda^{jk}_{\cdot\cdot} + p_1\lambda^{j}_{\cdot k}\right]\left[p_0(R_1\mu^{hi}_{\cdot\cdot} + R_2\nu^{hi}_{\cdot\cdot}) + p_1(R_1\mu^{h}_{\cdot i} + R_2\mu^{h}_{\cdot i})\right]$$

$$= R_1\left\{\tfrac{1}{2}\left[p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}\right]\left[p_0\mu^{jk}_{\cdot\cdot} + p_1\mu^{j}_{\cdot k}\right]\right.$$

$$\left. + \tfrac{1}{2}\left[p_0\lambda^{jk}_{\cdot\cdot} + p_1\lambda^{j}_{\cdot k}\right]\left[p_0\mu^{hi}_{\cdot\cdot} + p_1\mu^{h}_{\cdot i}\right]\right\}$$

$$+ R_2\left\{\tfrac{1}{2}\left[p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}\right]\left[p_0\nu^{jk}_{\cdot\cdot} + p_1\nu^{j}_{\cdot k}\right]\right.$$

$$\left. + \tfrac{1}{2}\left[p_0\lambda^{jk}_{\cdot\cdot} + p_1\lambda^{j}_{\cdot k}\right]\left[p_0\nu^{hi}_{\cdot\cdot} + p_1\nu^{h}_{\cdot i}\right]\right\}$$

$$= R_1\,\lambda^{hi}_{jk}\times\mu^{hi}_{jk} + R_2\,\lambda^{hi}_{jk}\times\nu^{hi}_{jk}.$$

The theorem on equilibrium of population after random intermating may be easily proved by the methods we have developed. We shall consider a somewhat more general case than is usually used, in that we allow any number of alleles, and also cross breeding between generations (i.e. the generations need not come in distinct steps nor need each individual mate in its own generation). We prove, then,

Theorem IX.

$$R_1 \lambda_i^k \times \lambda_i^k + R_2 \lambda_i^k \times (\lambda_i^k \times \lambda_i^k) + R_3 (\lambda_i^k \times \lambda_i^k) \times (\lambda_i^k \times \lambda_i^k) + \ldots = \lambda_i^k \times \lambda_i^k = \lambda_\cdot^k \lambda_\cdot^i \qquad (19)$$

The first term corresponds to a component representing direct offspring of our present population λ_i^k the second term represents a fraction obtained by mating this offspring with the parent generation, etc. Consider any term of this expression. In order to have a meaning it must have a factor $(\lambda_i^k \times \lambda_i^k) = \lambda_\cdot^k \lambda_\cdot^i$ (by 14), but this may be replaced by λ_i^k from Theorem VI since $\lambda_\cdot^k \lambda_\cdot^i = \lambda_\cdot^k$ Hence the number of factors in the product may be reduced by one. Continuing in this manner all terms reduce to the form $R_k \lambda_i^k \times \lambda_i^k$ and adding we get the desired result.

In particular, if we have "step type" generations this result shows that

$$\lambda_i^k \times \lambda_i^k = (\lambda_i^k \times \lambda_i^k) \times (\lambda_i^k \times \lambda_i^k)$$
$$= [(\lambda_i^k \times \lambda_i^k) \times (\lambda_i^k \times \lambda_i^k)] \times [(\lambda_i^k \times \lambda_i^k) \times (\lambda_i^k \times \lambda_i^k)]$$
$$= \text{etc.}$$

these expressions being the 2nd, 3rd, 4th, etc. offspring generations, and all of these are equal to $\lambda_\cdot^k \lambda_\cdot^i$ It is obvious that a necessary and sufficient

condition for a population λ_i^h to be in this type of equilibrium is that

$$\lambda_i^h = \lambda_\cdot^h \lambda_\cdot^i$$

For two alleles it is well known that this is equivalent to the condition

$$\lambda_1^1 \lambda_2^2 = \lambda_2^1 \lambda_2^1$$

We now show how this, and the generalized result for any number of alleles may be obtained.

Theorem X. The three following statements are all equivalent.

(1) $$\lambda_i^h = \lambda_\cdot^h \lambda_\cdot^i = \lambda_{i\times}^h \lambda_{i}^h$$

(2) $$\lambda_h^h \lambda_i^i = \lambda_i^h \lambda_i^h$$

(3) The matrix $\|\lambda_i^h\|$ is of rank one.

By the matrix $\|\lambda_i^h\|$ is meant the matrix:

$$\begin{matrix} \lambda_1^1 & \lambda_2^1 & \dots & \lambda_s^1 \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_s^2 \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \lambda_1^s & \lambda_2^s & \dots & \lambda_s^s \end{matrix}$$

In the first place (1) implies (2), for if

$$\lambda_{i}^{k} = \lambda_{\cdot}^{k}\,\lambda_{\cdot}^{i}$$

then $$\lambda_{k}^{k} = \lambda_{\cdot}^{k}\,\lambda_{\cdot}^{k}$$

and $$\lambda_{k}^{k}\,\lambda_{i}^{i} = \lambda_{\cdot}^{k}\,\lambda_{\cdot}^{k}\,\lambda_{\cdot}^{i}\,\lambda_{\cdot}^{i}$$
$$= (\lambda_{\cdot}^{k}\,\lambda_{\cdot}^{i})(\lambda_{\cdot}^{k}\,\lambda_{\cdot}^{i}) = \lambda_{i}^{k}\,\lambda_{i}^{k}$$

Also (2) implies (1). Summing (2) on i gives:

$$\lambda_{k}^{k} = \lambda_{\cdot}^{k}\,\lambda_{\cdot}^{k}$$

Hence: $$\lambda_{\cdot}^{k}\,\lambda_{\cdot}^{i} = \sqrt{\lambda_{k}^{k}\,\lambda_{i}^{i}}$$
$$= \sqrt{\lambda_{i}^{k}\,\lambda_{i}^{k}} = \lambda_{i}^{k}$$

Thus (1) and (2) are equivalent.

Condition (3) is equivalent to either of these, for if (1) is true:

$$\lambda_{i}^{k} = \lambda_{\cdot}^{k}\,\lambda_{\cdot}^{i}$$

and the elements of the matrix $\| \lambda_{i}^{k} \|$ can be written as the product of a number depending only on the row by a number depending only on the column. This is a well known condition that the matrix be of rank not greater than one. The rank is actually one since at least one element is different from zero to satisfy $\lambda_{\cdot}^{\cdot} = 1$. Thus (1) implies (3). If (3) is true then

each second order minor of $\| \lambda_i^h \|$ must vanish. In particular we have

$$\begin{vmatrix} \lambda_h^h & \lambda_i^h \\ \lambda_h^i & \lambda_i^i \end{vmatrix} = 0$$

Hence

$$\lambda_h^h \lambda_i^i = \lambda_i^h \lambda_i^h$$

so that (3) implies (2). This shows that all the conditions are equivalent and proves the theorem.

If a population is in equilibrium we have

$$\sum_i \sqrt{\lambda_i^i} = \sum_i \sqrt{\lambda_\cdot^i \lambda_\cdot^i} = \sum_i \lambda_\cdot^i = \lambda_\cdot^\cdot = 1$$

but this is not a sufficient condition for equilibrium, as the example

$$\| \lambda_i^h \| = \left\| \begin{matrix} \frac{1}{9} & \frac{1}{6} & 0 \\ \frac{1}{6} & \frac{1}{9} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{9} \end{matrix} \right\|$$

proves, for here $\sqrt{\frac{1}{9}} + \sqrt{\frac{1}{9}} + \sqrt{\frac{1}{9}} = 1$ while $\begin{vmatrix} \frac{1}{6} & 0 \\ \frac{1}{9} & \frac{1}{6} \end{vmatrix} = \frac{1}{36} \neq 0$ so the population is not in equilibrium.

In case more than one factor is considered the population will, in general, only reach equilibrium (for gene combinations) asymptotically. Suppose we have two linked factors and assume "step type" gene-

rations. The result of random intermating is given by:

Theorem XI. Under random intermating of λ_{jk}^{hi} the nth generation is the population

$$\mu_{jk}^{hi} = \left[p_0^{n-1}(p_0\lambda_{\cdot\cdot}^{hi} + p_1\lambda_{\cdot i}^{h\cdot}) + (1-p_0^{n-1})\lambda_{\cdot\cdot}^{h\cdot}\lambda_{\cdot\cdot}^{\cdot i}\right] \cdot \left[p_0^{n-1}(p_0\lambda_{\cdot\cdot}^{jk} + p_1\lambda_{\cdot k}^{j\cdot}) + (1-p_0^{n-1})\lambda_{\cdot\cdot}^{j\cdot}\lambda_{\cdot\cdot}^{\cdot k}\right] \quad (20)$$

and (assuming $p_0 \neq 1$) approaches asymptotically the population

$$\nu_{jk}^{hi} = \lambda_{\cdot\cdot}^{h\cdot}\lambda_{\cdot\cdot}^{\cdot i}\lambda_{\cdot\cdot}^{j\cdot}\lambda_{\cdot\cdot}^{\cdot k} \quad \text{as } n \to \infty$$

Proof: By definition (8) the first generation is:

$$\sigma_{jk}^{hi} = \lambda_{jk}^{hi} \times \lambda_{jk}^{hi} = [p_0\lambda_{\cdot\cdot}^{hi} + p_1\lambda_{\cdot i}^{h\cdot}][p_0\lambda_{\cdot\cdot}^{jk} + p_1\lambda_{\cdot k}^{j\cdot}]$$

$$= \left[(p_0\lambda_{\cdot\cdot}^{hi} + p_1\lambda_{\cdot i}^{h\cdot}) + (1-p_0^{1-1})\lambda_{\cdot\cdot}^{h\cdot}\lambda_{\cdot\cdot}^{\cdot i}\right] \cdot \left[(p_0\lambda_{\cdot\cdot}^{jk} + p_1\lambda_{\cdot k}^{j\cdot}) + (1-p_0^{1-1})\lambda_{\cdot\cdot}^{j\cdot}\lambda_{\cdot\cdot}^{\cdot k}\right]$$

so the theorem is true for n = 1. We now show that if it is true for the nth generation it will be true for the (n + 1)th generation and thus complete the proof by mathematical induction. Assume, then, that the nth generation is

$$\mu_{jk}^{hi} = \left[p_0^{n-1}(p_0\lambda_{\cdot\cdot}^{hi} + p_1\lambda_{\cdot i}^{h\cdot}) + (1-p_0^{n-1})\lambda_{\cdot\cdot}^{h\cdot}\lambda_{\cdot\cdot}^{\cdot i}\right] \cdot \left[p_0^{n-1}(p_0\lambda_{\cdot\cdot}^{jk} + p_1\lambda_{\cdot k}^{j\cdot}) + (1-p_0^{n-1})\lambda_{\cdot\cdot}^{j\cdot}\lambda_{\cdot\cdot}^{\cdot k}\right] \quad (21)$$

whence, summing on j and k:

$$\mu^{hi}_{\cdot\cdot} = \left[p_0^{n-1}(p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}) + (1-p_0^{n-1})\lambda^{h}_{\cdot\cdot}\lambda^{\cdot i}_{\cdot}\right]$$
$$\cdot\left[p_0^{n-1}(p_0\lambda^{\cdot\cdot} + p_1\lambda^{\cdot\cdot}) + (1-p_0^{n-1})\lambda^{\cdot\cdot}\lambda^{\cdot\cdot}\right]$$
$$= \left[p_0^{n-1}(p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}) + (1-p_0^{n-1})\lambda^{h}_{\cdot\cdot}\lambda^{\cdot i}_{\cdot}\right]$$

since $\lambda^{\cdot\cdot}_{\cdot\cdot} = 1$ and $p_0 + p_1 = 1$.

Also:

$$\mu^{h}_{\cdot k} = \left[p_0^{n-1}(p_0\lambda^{h}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot\cdot}) + (1-p_0^{n-1})\lambda^{h}_{\cdot\cdot}\lambda^{\cdot\cdot}_{\cdot}\right]$$
$$\cdot\left[p_0^{n-1}(p_0\lambda^{\cdot\cdot}_{\cdot k} + p_1\lambda^{\cdot\cdot}_{\cdot k}) + (1-p_0^{n-1})\lambda^{\cdot\cdot}_{\cdot\cdot}\lambda^{\cdot k}_{\cdot\cdot}\right]$$
$$= \lambda^{h}_{\cdot\cdot}\lambda^{\cdot k}_{\cdot\cdot}$$

Now the $(n+1)$th generation is given by the cross product of the nth generation with itself, i.e.

$$\mu^{hi}_{jk} \times \mu^{hi}_{jk} = \left[p_0\mu^{hi}_{\cdot\cdot} + p_1\mu^{h}_{\cdot i}\right]\left[p_0\mu^{j}_{\cdot k} + p_1\mu^{j}_{\cdot k}\right]$$

$$= \left[p_0\left\{p_0^{n-1}(p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}) + (1-p_0^{n-1})\lambda^{h}_{\cdot\cdot}\lambda^{\cdot i}_{\cdot}\right\} + p_1\lambda^{h}_{\cdot\cdot}\lambda^{\cdot i}_{\cdot}\right]$$

$$\cdot\left[p_0\left\{p_0^{n-1}(p_0\lambda^{jk}_{\cdot\cdot} + p_1\lambda^{j}_{\cdot k}) + (1-p_0^{n-1})\lambda^{j}_{\cdot\cdot}\lambda^{\cdot k}_{\cdot\cdot}\right\} + p_1\lambda^{j}_{\cdot\cdot}\lambda^{\cdot k}_{\cdot\cdot}\right]$$

$$= \left[p_0^{n}(p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}) + (1-p_0^{n})\lambda^{h}_{\cdot\cdot}\lambda^{\cdot i}_{\cdot}\right]$$
$$\cdot\left[p_0^{n}(p_0\lambda^{jk}_{\cdot\cdot} + p_1\lambda^{j}_{\cdot k}) + (1-p_0^{n})\lambda^{j}_{\cdot\cdot}\lambda^{\cdot k}_{\cdot\cdot}\right]$$

which is the same expression as (21) with n replaced by (n+1). This, therefore, completes the proof. The asymptotic value is obvious since if $p_0 \neq 1$ then as $n \to \infty$, $p_0^{n-1} \to 0$ and $(1-p_0^{n-1}) \to 1$ so that (21) reduces to

$$\lambda^{h\cdot}_{\cdot\cdot}\ \lambda^{\cdot i}_{\cdot\cdot}\ \lambda^{j\cdot}_{\cdot\cdot}\ \lambda^{\cdot k}_{\cdot\cdot} \tag{22}$$

An obvious corollary is that a necessary and sufficient condition for a population C^{hi}_{jk} to be stable under random intermating is that it satisfy the conditions

$$C^{hi}_{jk} = C^{h\cdot}_{\cdot\cdot}\ C^{\cdot i}_{\cdot\cdot}\ C^{j\cdot}_{\cdot\cdot}\ C^{\cdot k}_{\cdot\cdot}$$

or that either $p_0 = 0$ or $p_0 = 1$. If $p_0 = 0$ the expression (21) reduces to its equilibrium value

$$\lambda^{h\cdot}_{\cdot\cdot}\ \lambda^{\cdot i}_{\cdot\cdot}\ \lambda^{j\cdot}_{\cdot\cdot}\ \lambda^{\cdot k}_{\cdot\cdot}$$

at the first generation. If $p_0 = 1$ we have perfect linkage and the expression becomes $\lambda^{hi}_{\cdot\cdot}\ \lambda^{jk}_{\cdot\cdot}$ as it should, since it then acts like a single factor.

We note that the speed with which equilibrium is approached depends entirely on the value of p_0. If p_0 is small the approach will be very rapid, less so as p_0 becomes larger, or the linkage closer between the factors.

It is interesting to see that with a given popula-

tion λ^{hi}_{jk}, the equilibrium will be approached more rapidly if there is very weak linkage ($p_0 < 1/2$) than if the factors are completely independent, either in different chromosomes or in the same chromosome with $p_0 = 1/2$.

Incidentally, if there is no interference between crosses a recombination value $p_0 < 1/2$ is mathematically impossible. Suppose the map distance (measured in morgans) between two loci is d. Let this distance be divided into a large number, n, of sections, each of map distance d/n. Then if n is large d/n is small and the probability of a cross in any one section is approximately d/n and approaches this value as $n \to \infty$. The probability of exactly s crosses between the two loci is given by

$$\lim_{n \to \infty} {}_nC_s \left(\frac{d}{n}\right)^s \left(1 - \frac{d}{n}\right)^{n-s} \qquad (24)$$

Where ${}_nC_s = \frac{n(n-1)\ldots(n-s+1)}{s!}$ is the number of ways we can pick out the s sections where the crosses may occur, $\left(\frac{d}{n}\right)^s$ is the probability that these crosses do occur and $\left(1 - \frac{d}{n}\right)^{n-s}$ is the probability that crosses do not occur in the other (n-s) sections. This limit may be written as follows:

$$\lim_{n \to \infty} \frac{n(n-1)(n-2)\ldots(n-s+1)}{n \cdot n \cdot n \ldots n} \cdot \frac{d^s}{s!} \left(1 - \frac{d}{n}\right)^n \cdot \left(1 - \frac{d}{n}\right)^{-s}$$

The first factor approaches the limit 1, the second is a constant $\frac{d^s}{s!}$, the third is (setting $x = \frac{n}{d}$):

$$\lim_{n\to\infty} \left(1 - \frac{d}{n}\right)^n = \lim_{x\to\infty} \left(1 - \frac{1}{x}\right)^{xd} = e^{-d}$$

and the last term approaches 1 since the exponent is a constant. The entire function therefore approaches

$$\frac{d^s}{s!} e^{-d}$$

as the probability of exactly s crosses. The probability of an odd number is then

$$p_1 = e^{-d}\left(\frac{d}{1!} + \frac{d^3}{3!} + \frac{d^5}{5!} + \cdots\right)$$

$$= e^{-d}\left(\frac{e^d - e^{-d}}{2}\right) = \frac{1 - e^{-2d}}{2}$$

which is clearly less than 50% for any real d. In case interference is present we cannot multiply probabilities as we did in (24) since the events are no longer independent and it is at least mathematically possible for values of $p_1 > 1/2$. Thus suppose we have a long chromosome which is very likely to cross over at least once, but one cross strongly inhibits any other crosses for a large distance. It is evident that such conditions would allow recombination values greater than 50%.

In the case when the factors are in different chromosomes, equation (20) can be simplified, due to the fact that

$$\lambda^{hi}_{jk} = \lambda^{hk}_{ji}$$

under these conditions. It follows that

$$\lambda^{hi}_{..} = \lambda^{h}_{.i} \; ; \quad \lambda^{jk}_{..} = \lambda^{j}_{.k}$$

so that (20) reduces to

$$\mu^{hi}_{jk} = \left[p_0^{n-1} \lambda^{hi}_{..} + (1-p_0^{n-1}) \lambda^{h}_{..} \lambda^{i}_{..} \right] \cdot \left[p_0^{n-1} \lambda^{jk}_{..} + (1-p_0^{n-1}) \lambda^{j}_{..} \lambda^{k}_{..} \right]$$

For three linked factors the first generation offspring of the population λ^{hij}_{klm} is

$$\left[p_{00} \lambda^{hij}_{...} + p_{01} \lambda^{hi}_{..j} + p_{10} \lambda^{h}_{.ij} + p_{11} \lambda^{h}_{.i.j} \right] \cdot \left[p_{00} \lambda^{klm}_{...} + p_{01} \lambda^{kl}_{..m} + p_{10} \lambda^{k}_{.lm} + p_{11} \lambda^{k}_{.l.m} \right] \quad (23)$$

The second random intermating gives:

$$\begin{aligned} & \left[p_{00} \left(p_{00} \lambda^{hij}_{...} + p_{01} \lambda^{hi}_{..j} + p_{10} \lambda^{h}_{.ij} + p_{11} \lambda^{h}_{.i.j} \right) \right. \\ & + p_{01} \left((p_{00}+p_{01}) \lambda^{hi}_{...} + (p_{10}+p_{11}) \lambda^{h}_{.i.} \right) \lambda^{.j}_{...} \\ & + p_{10} \left((p_{00}+p_{10}) \lambda^{i}_{..j} + (p_{01}+p_{11}) \lambda^{i}_{..j} \right) \lambda^{h}_{...} \quad (25) \\ & \left. + p_{11} \left((p_{00}+p_{11}) \lambda^{h.j}_{...} + (p_{10}+p_{01}) \lambda^{h}_{..j} \right) \lambda^{i}_{...} \right] \\ & \cdot \left[\text{same expression with } h, i, j \text{ replaced by } k, l, m. \right] \end{aligned}$$

The nth generation has also been determined in

this case. It is given by the following proposition.

Theorem XII. Under random intermating of the population λ^{hij}_{klm} the nth offspring generation is given by:

$$\mu^{hij}_{klm} = \Big[p_{00}^{n-1}\left(p_{00}\lambda^{hij}_{\cdots} + p_{01}\lambda^{hi\cdot}_{\cdots j} + p_{10}\lambda^{h\cdot j}_{\cdots i} + p_{11}\lambda^{h\cdot\cdot}_{\cdots ij}\right)$$

$$+ \left((p_{00}+p_{01})^{n-1} - p_{00}^{n-1}\right)\left\{(p_{00}+p_{01})\lambda^{hi\cdot}_{\cdots} + (p_{10}+p_{11})\lambda^{h\cdot\cdot}_{\cdots i}\right\}\lambda^{\cdot\cdot j}_{\cdots}$$

$$+ \left((p_{00}+p_{10})^{n-1} - p_{00}^{n-1}\right)\left\{(p_{00}+p_{10})\lambda^{\cdot i j}_{\cdots} + (p_{10}+p_{11})\lambda^{\cdot i\cdot}_{\cdots j}\right\}\lambda^{h\cdot\cdot}_{\cdots}$$

$$+ \left((p_{00}+p_{11})^{n-1} - p_{00}^{n-1}\right)\left\{(p_{00}+p_{11})\lambda^{h\cdot j}_{\cdots} + (p_{01}+p_{10})\lambda^{h\cdot\cdot}_{\cdots j}\right\}\lambda^{\cdot i\cdot}_{\cdots}$$

$$+ \left(1 + 2p_{00}^{n-1} - (p_{00}+p_{01})^{n-1} - (p_{10}+p_{00})^{n-1} - (p_{00}+p_{11})^{n-1}\right)\lambda^{h\cdot\cdot}_{\cdots}\lambda^{\cdot i\cdot}_{\cdots}\lambda^{\cdot\cdot j}_{\cdots}\Big]$$

$$\cdot\Big[\text{same expression with h, i, j replaced by k, l, m}\Big]$$

This may be proved by mathematical induction exactly as we proved Theorem XI. The expression

approaches $\lambda^{h\cdot\cdot}_{\cdot\cdot\cdot}\,\lambda^{\cdot i\cdot}_{\cdot\cdot\cdot}\,\lambda^{\cdot\cdot j}_{\cdot\cdot\cdot}\,\lambda^{k\cdot\cdot}_{\cdot\cdot\cdot}\,\lambda^{\cdot l\cdot}_{\cdot\cdot\cdot}\,\lambda^{\cdot\cdot m}_{\cdot\cdot\cdot}$ asymptotically as $n \to \infty$. In case the three loci are in different chromosomes it reduces to

$$\frac{1}{16^{n-1}}\left[\lambda^{hij}_{\cdot\cdot\cdot} + (2^{n-1}-1)\left\{\lambda^{hi\cdot}_{\cdot\cdot\cdot}\lambda^{\cdot\cdot j}_{\cdot\cdot\cdot} + \lambda^{h\cdot j}_{\cdot\cdot\cdot}\lambda^{\cdot i\cdot}_{\cdot\cdot\cdot} + \lambda^{\cdot ij}_{\cdot\cdot\cdot}\lambda^{h\cdot\cdot}_{\cdot\cdot\cdot}\right\} + (4^{n-1} - 3\cdot 2^{n-1} + 2)\lambda^{h\cdot\cdot}_{\cdot\cdot\cdot}\lambda^{\cdot i\cdot}_{\cdot\cdot\cdot}\lambda^{\cdot\cdot j}_{\cdot\cdot\cdot}\right] \quad (26)$$

. [same expression with h, i, j replaced by k, l, m]

It is possible to expand a population in a series form which displays the homogeneous components of the population. This series is very similar to the expansion of a Boolean function in Symbolic Logic, and not only throws light on the mathematical nature of the symbols we are using, but is also useful for computational purposes. To develop the expansion we must first define a set of "constants"; homogeneous populations of a certain fixed genetic constitution. These constants will always be represented by the base letter γ, and the indices refer to the particular locus or loci we are considering. All the members of a constant γ population have the same genetic constitution with respect to the factors under consideration.

For a single locus our definitions are as follows:

$$\gamma^{a}_{a\,i\,j} = \begin{cases} 1 & \text{if } i = j = a \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

If $a \neq b$

$$\gamma^{a}_{b\,i\,j} = \begin{cases} \frac{1}{2} & \text{if } i = a \text{ and } j = b \text{ or if } i = b \text{ and } j = a \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

Here a and b are dead indices; they represent certain fixed numbers while the live indices i and j represent any of several values. The dead indices merely serve to distinguish one γ from another and there is, in general, no index balance on them. It will be seen from our definition that ${}^{a}_{b}\gamma^{i}_{j}$ represents a population whose members are all of the genetic type $A_a A_b$. Thus ${}^{1}_{2}\gamma^{i}_{j}$ is a homogeneous population with the formula $A_1 A_2$.

If we are considering only a single dimorphic factor the series expansion of a population λ^{i}_{j} is as follows:

$$\lambda^{i}_{j} = \lambda'_1 \, {}^{1}\gamma^{i}_{j} + 2\lambda'_1\lambda'_2 \, {}^{1}_{2}\gamma^{i}_{j} + \lambda'^{2}_{2} \, {}^{2}\gamma^{i}_{j} \qquad (29)$$

To prove this it is merely necessary to note that it reduces to an identity for all values of i and j. Thus with i = 1, j = 1 all the terms on the right are zero, except the first which reduces to λ'_1. For i = 1, j = 2 only the second term is effective giving $\lambda'_2 = 2\lambda'_2\frac{1}{2} = \lambda'_2$, etc.

The expansion (29) displays λ^{i}_{j} as a population made up of three homogeneous parts ${}^{1}\gamma^{i}_{j}$, ${}^{1}_{2}\gamma^{i}_{j}$, ${}^{2}\gamma^{i}_{j}$. The fraction of each type is the coefficient of the corresponding γ in the expansion. For more than two alleles the expansion takes the form:

$$\lambda^{i}_{j} = \lambda^{1}_{1}\,{}_{1}\gamma^{i}_{j} + \lambda^{2}_{2}\,{}_{2}\gamma^{i}_{j} + \ldots + \lambda^{n}_{n}\,{}_{n}\gamma^{i}_{j}$$
$$+ 2\lambda^{1}_{2}\,{}_{2}\gamma^{i}_{j} + 2\lambda^{1}_{3}\,{}_{3}\gamma^{i}_{j} + \ldots + 2\lambda^{1}_{n}\,{}_{n}\gamma^{i}_{j} \qquad (30)$$
$$+ 2\lambda^{2}_{3}\,{}_{3}\gamma^{i}_{j} + \ldots + 2\lambda^{n-1}_{n}\,{}_{n}\gamma^{i}_{j}$$

With more than one factor under consideration we define the γ populations according to the following scheme:

$$ {}_{ab}\gamma^{hi}_{jk} = \begin{cases} 1 \text{ if } \begin{cases} h = j = a \\ i = k = b \end{cases} \\ 0 \text{ otherwise} \end{cases} \qquad (31)$$

If $a \neq c$ or $b \neq d$ or both

$$ {}_{cd}\gamma^{hi}_{jk} = \begin{cases} 1/2 \text{ if } \begin{cases} h = a & i = b \\ j = c & k = d \end{cases} \text{ or } \begin{cases} h = c & i = d \\ j = a & k = b \end{cases} \\ 0 \text{ otherwise} \end{cases} \qquad (32)$$

Thus ${}_{32}\gamma^{hi}_{jk}$ represents a homogeneous population whose members all have the formula $A_1\ A_3\ B_2\ B_2$. Constants for more than two factors are defined in a completely analogous manner. Thus ${}_{def}\gamma^{hij}_{kpm}$ is a population whose members are of the type $A_a\ A_d\ B_b\ B_c\ C_c\ C_f$.

The series expansion for more than one locus has the form (taking three factors to be specific):

$$\lambda^{h\,i\,j}_{l\,m} = \lambda^{1\,1\,1}_{1\,1\,1}\gamma^{h\,i\,j}_{k\,l\,m} + 2\lambda^{1\,1\,2}_{1\,1\,1}\gamma^{h\,i\,j}_{k\,l\,m} + \ldots + \lambda^{r_1\,r_2\,r_3}_{r_1\,r_2\,r_3}\gamma^{h\,i\,j}_{r_1\,r_2\,r_3\;k\,l\,m} \qquad (33)$$

where r_1, r_2, and r_3 are the number of alleles of the three factors. Each term corresponding to a part of the population homozygous in all factors has the coefficient one; if the corresponding part of the population is heterozygous in one or more factors the coefficient should be two.

The cross product of any two γ population may be written as a linear combination of γ's. Thus

$$\gamma^{a\,h}_{b\,i} \times \gamma^{c\,h}_{d\,i} = \tfrac{1}{4}\gamma^{a\,h}_{c\,i} + \tfrac{1}{4}\gamma^{a\,h}_{d\,i} + \tfrac{1}{4}\gamma^{b\,h}_{c\,i} + \tfrac{1}{4}\gamma^{b\,h}_{d\,i} \qquad (34)$$

In case some of the numbers a, b, c, d are equal this expression is still true but may be simplified. Thus if a = b

$$\gamma^{a\,h}_{a\,i} \times \gamma^{c\,h}_{d\,i} = \tfrac{1}{2}\gamma^{a\,h}_{c\,i} + \tfrac{1}{2}\gamma^{a\,h}_{d\,i}$$

For two loci the law of multiplication of the γ's is

$$\gamma^{a\,b\,h\,i}_{c\,d\,j\,k} \times \gamma^{e\,f\,h\,i}_{g\,h\,j\,k} = \frac{p_0^2}{4}\left[\gamma^{a\,b\,h\,i}_{e\,f\,j\,k} + \gamma^{a\,b\,h\,i}_{g\,h\,j\,k} + \gamma^{c\,d\,h\,i}_{e\,f\,j\,k} + \gamma^{c\,d\,h\,i}_{g\,h\,j\,k}\right]$$
$$+ \frac{p_0 p_1}{4}\left[\gamma^{a\,b\,h\,i}_{e\,h\,j\,k} + \gamma^{a\,b\,h\,i}_{g\,f\,j\,k} + \gamma^{c\,d\,h\,i}_{e\,h\,j\,k} + \gamma^{c\,d\,h\,i}_{g\,f\,j\,k} + \gamma^{a\,d\,h\,i}_{e\,f\,j\,k} + \gamma^{a\,d\,h\,i}_{g\,h\,j\,k}\right.$$
$$\left. + \gamma^{c\,b\,h\,i}_{e\,f\,j\,k} + \gamma^{c\,b\,h\,i}_{g\,h\,j\,k}\right] + \frac{p_1^2}{4}\left[\gamma^{a\,d\,h\,i}_{e\,h\,j\,k} + \gamma^{a\,d\,h\,i}_{g\,h\,j\,k} + \gamma^{c\,b\,h\,i}_{e\,h\,j\,k} + \gamma^{c\,b\,h\,i}_{g\,f\,j\,k}\right]$$

The series expansion and law of multiplication of the γ's display a population as a hypercomplex number, i.e. as a symbol of the form $(a_1 e_1 + a_2 e_2 + \ldots + a_n e_n)$ where the coefficients $a_1, a_2 \ldots a_n$ are numbers and the symbols $e_1, e_2 \ldots e_n$ are "unit vectors" with some given law of multiplication such that the product of any two of the e's may be written as a linear combination of e's. It is well known that except for trivial cases and the case of ordinary complex numbers, no law of multiplication preserves all the commutative, associative and distributive laws of ordinary numbers. In our case the associative law of multiplication is sacrificed. Thus the product $({}_1\gamma_i^h \times {}_1\gamma_i^h) \times {}_2\gamma_i^h$ is, by (34):

$$({}_1\gamma_i^h \times {}_1\gamma_i^h) \times {}_2\gamma_i^h = {}_1\gamma_i^h \times {}_2\gamma_i^h = \tfrac{1}{2}\,{}_1\gamma_i^h + \tfrac{1}{2}\,{}_2\gamma_i^h$$

while on the other hand

$${}_1\gamma_i^h \times ({}_1\gamma_i^h \times {}_2\gamma_i^h) = {}_1\gamma_i^h \times (\tfrac{1}{2}\,{}_1\gamma_i^h + \tfrac{1}{2}\,{}_2\gamma_i^h)$$

$$= \tfrac{1}{2}\,{}_1\gamma_i^h + \tfrac{1}{2}(\tfrac{1}{2}\,{}_1\gamma_i^h + \tfrac{1}{2}\,{}_2\gamma_i^h)$$

$$= \tfrac{3}{4}\,{}_1\gamma_i^h + \tfrac{1}{4}\,{}_2\gamma_i^h$$

This is a simple example of a multiplication in which the associative law does not hold, and shows that in a cross product of several factors it is essential that parentheses be retained to indicate the order in which

multiplication is performed.

The series expansion (33) of a population shows how an arbitrary population may be written as the sum of a set of particular homogeneous populations, the γ's. The choice of this particular set of populations as components was a matter of convenience, not of necessity. We now show that any set of populations satisfying a certain simple condition would do as well.

Consider a set of n populations $\lambda, \mu, \ldots, \sigma$. We omit writing the indices, but there may be any number of loci. We will say that these populations are linearly independent if there is no set of numbers $a_1, a_2, \ldots, a_n$, not all zero, and that

$$a_1 \lambda + a_2 \mu + \ldots + a_n \sigma = 0 \qquad (35)$$

for all values of the live indices.

Theorem XIII: Any population φ may be expressed uniquely as a linear combination of n linearly independent populations where n is the number of different possible genetic formulae for the factors considered.

To prove this, note that a necessary and sufficient condition that (35) have no solution for the a's (not all zero) is that the determinant

$$\begin{vmatrix} \lambda^{11\ldots1}_{11\ldots1} & \mu^{11\ldots1}_{11\ldots1} & \ldots & \sigma^{11\ldots1}_{11\ldots1} \\ \lambda^{11\ldots1}_{11\ldots2} & \mu^{11\ldots1}_{11\ldots2} & \ldots & \sigma^{11\ldots1}_{11\ldots2} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \lambda^{\alpha_1\alpha_2\ldots\alpha_s}_{\alpha_1\alpha_2\ldots\alpha_s} & \mu^{\alpha_1\alpha_2\ldots\alpha_s}_{\alpha_1\alpha_2\ldots\alpha_s} & \ldots & \sigma^{\alpha_1\alpha_2\ldots\alpha_s}_{\alpha_1\alpha_2\ldots\alpha_s} \end{vmatrix} \qquad (36)$$

be different from zero. In this determinant each population takes on the values of all components in a column; i.e. the values obtained by giving the indices all possible values.

Now the non-vanishing of (36) is also a necessary and sufficient condition for the existence of a unique solution for the b's in the equations:

$$\varphi = b_1\lambda + b_2\mu + \ldots + b_n\sigma$$

and this proves the theorem.

In passing we note that if we have a linked factors with $r_1, r_2, \ldots, r_s$ alleles respectively, then n, the number of different components is given by

$$n = \frac{r_1 r_2 \ldots r_s (r_1 r_2 \ldots r_s + 1)}{2} \qquad (37)$$

We may think of 2s positions in whice genes may be placed. There are r_1, possibilities for the first and second positions, r_2 for the third and fourth, etc., and therefore a total of $r_1^2 r_2^2 r_3^2 \ldots r_s^2$. However, as an interchange of

the two chromosomes does not affect the genetic constitution we should divide this by two, except for the ones which are homozygous in all factors and were not counted twice. There are $r_1 r_2 \ldots r_s$ types of fully homozygous individuals and we may correct our formula, then, by adding this and then dividing by two:

$$n = \frac{r_1^2 r_2^2 \ldots r_s^2 + r_1 r_2 \ldots r_s}{2} = 1/2 \left[r_1 r_2 \ldots r_s (r_1 r_2 \ldots r_s + 1) \right]$$

In case the loci are not all in the same chromosome but spread out in a number of different ones, we may evaluate the expression (37) for each chromosome involved and multiply these results.

IV. The Solution of Equations Involving Unknown Populations

It is easy to write down equations of various types involving unknown populations. Many of these may be solved for the unknowns in terms of the known populations by means of the theorems we have developed. In general an equation represents some breeding experiment involving a population of unknown genetic constitution resulting in a genetically known population. In the following we shall use the letters $\varphi, \psi, \chi \ldots$ for base letters in unknown populations and $\lambda, \mu, \nu \ldots$ for known populations.

The general method of attack on these problems may be outlined as follows:

1. By summing on various indices we are able to evaluate gene frequencies, gene pair frequencies, etc. for the unknown populations. This ordinarily involves no more than the solution of one or more linear algebraic equations.

2. Knowing these we can evaluate cross products in which the unknowns appear, since a cross product depends only on the values of the population symbol with half the indices dotted.

3. With only linear terms remaining it is usually easy to solve the equations by or-

dinary methods for algebraic equations.

To illustrate how this is done we shall consider several examples. Suppose first, for simplicity, that only one locus is involved, and that we have the equation:

$$R_1 \varphi_i^h + R_2 \varphi_i^h \times \varphi_i^h + R_3 \varphi_i^h \times (\varphi_i^h \times \varphi_i^h) + \ldots = \lambda_i^h \qquad \Sigma R_i = 1 \tag{38}$$

with φ_i^h unknown and λ_i^h known.

By Theorem IX this reduces immediately to the form

$$R \varphi_i^h + S \varphi_i^h \times \varphi_i^h = \lambda_i^h \quad ; \quad R + S = 1 \tag{39}$$

Summing on i we have by Theorem VII

$$R \varphi_\cdot^h + \tfrac{1}{2} S \varphi_\cdot^h + \tfrac{1}{2} S \varphi_\cdot^h = \lambda_\cdot^h$$

or $$\varphi_\cdot^h = \lambda_\cdot^h$$

Hence by Theorem VI we may replace φ_i^h by λ_i^h in any product. Returning then to equation (39) we have

$$R \varphi_i^h + S \lambda_i^h \times \lambda_i^h = \lambda_i^h$$

$$\varphi_i^h = \tfrac{1}{R} \lambda_i^h - \tfrac{S}{R} \lambda_i^h \times \lambda_i^h$$

and this must be the unique solution of the equation, if a solution exists. To prove that is a solution we merely try it in the equation and find that it is satisfied.

A more general equation in one unknown is the following:

$$R\,\varphi_i^h(\mu_i^h \times \varphi_i^h) + S\,\varphi_i^h \times \varphi_i^h + T\,\varphi_i^h = \lambda_i^h \qquad (40)$$

Summing on i:

$$R\left(\tfrac{1}{2}\varphi_\cdot^h + \tfrac{1}{4}\mu_\cdot^h + \tfrac{1}{4}\varphi_\cdot^h\right) + S\varphi_\cdot^h + T\varphi_\cdot^h = \lambda_\cdot^h$$

$$\varphi_\cdot^h = \frac{4}{3R+4S+4T}\lambda_\cdot^h - \frac{R}{3R+4S+4T}\mu_\cdot^h$$

$$= \frac{4}{4-R}\lambda_\cdot^h - \frac{R}{4-R}\mu_\cdot^h$$

Replacing φ_i^h in each product of (39) by the expression $\frac{4}{4-R}\lambda_\cdot^h - \frac{R}{4-R}\mu_\cdot^h$ gives:

$$\varphi_i^h = \frac{1}{T}\lambda_i^h - \left(\frac{4}{4-R}\lambda_i^h - \frac{R}{4-R}\mu_i^h\right) \cdot$$

$$\times\left[\frac{4}{4-R}\lambda_i^h \times \mu_i^h - \frac{R^2}{4-R}\mu_i^h \times \mu_i^h + \frac{4S}{4-R}\lambda_i^h - \frac{RS}{4-R}\mu_i^h\right]$$

as the solution of (39).

It may be easily shown that the above method is applicable to any single equation in one unknown φ_i^h providing the coefficient of φ_i^h does not equal zero. Such an equation always has a unique solution, although this solution may not always represent a realizable population.

A system of linear simultaneous equations may be solved by the ordinary methods for algebraic equations, since by fixing the indices we actually have such a sys-

tem of linear algebraic equations. Thus suppose

$$\sum_{i=1}^{w} R_{ij}\ {}_{i}\varphi_{\ell}^{\kappa} = {}_{j}\lambda_{\ell}^{\kappa} \qquad j = 1, 2, \ldots, w \tag{41}$$

represents the system of equation with ${}_{1}\varphi_{\ell}^{\kappa}, {}_{2}\varphi_{\ell}^{\kappa}, \ldots, {}_{w}\varphi_{\ell}^{\kappa}$ unknowns. The indices i and j here serve to distinguish between the different φ and λ populations. The solution of this system is

$${}_{i}\varphi_{\ell}^{\kappa} = \sum_{j=1}^{w} \frac{M_{ji}}{|R_{ji}|}\ {}_{j}\lambda_{\ell}^{\kappa} \tag{42}$$

where $|R_{ij}|$, the determinant of the coefficients $\neq 0$ $\neq 0$, and M_{ji} is the cofactor of R_{ij}.

The reader may verify that simultaneous systems of equations with cross products involving the unknown populations may also be solved by these methods. Turning now to problems involving more than one gene locus the situation becomes a bit more complicated. Suppose again we have the equation in one unknown

$$R\,\varphi_{j\kappa}^{\ell i} \times \varphi_{j\kappa}^{\ell i} + S\,\varphi_{j\kappa}^{\ell i} = \lambda_{j\kappa}^{\ell i} \tag{43}$$

Referring to the definition of a cross product for two factors we see that knowing the two quantities $\varphi_{j\cdot}^{\ell i}$ and $\varphi_{\cdot\kappa}^{\ell i}$ the cross product in (43) is completely determined. Summing, then, on j and k we get

$$R\,\varphi_{\cdot\cdot}^{\ell i} \times \varphi_{\cdot\cdot}^{\ell i} + S\,\varphi_{\cdot\kappa}^{\ell i} = \lambda_{\cdot\kappa}^{\ell i} \tag{44}$$

and we se that to determine $\varphi_{\cdot\kappa}^{\ell i}$ we should find $\varphi_{\cdot\cdot}^{\ell i}$

Summing (45) on i gives us

$$\varphi^{h}_{\cdot\cdot} = \lambda^{h}_{\cdot\cdot}$$

similarly

$$\varphi^{\cdot i}_{\cdot\cdot} = \lambda^{\cdot i}_{\cdot\cdot}$$

hence

$$R\,\lambda^{h}_{\cdot\cdot}\,\lambda^{\cdot i}_{\cdot\cdot} + S\,\varphi^{h\,i}_{\cdot\,\cdot} = \lambda^{h\,i}_{\cdot\,\cdot}$$

$$\varphi^{h\,i}_{\cdot\,\cdot} = \frac{1}{S}\,\lambda^{h\,i}_{\cdot\,\cdot} - \frac{R}{S}\,\lambda^{h}_{\cdot\cdot}\,\lambda^{\cdot i}_{\cdot\cdot} \qquad (46)$$

Substituting in (44)

$$\varphi^{h\,i}_{\cdot\,\cdot} = \frac{1}{S+Rp_{\cdot}}\left(\lambda^{h\,i}_{\cdot\,\cdot} - \frac{Rp_{\cdot}}{S}\,\lambda^{h\,i}_{\cdot\,\cdot} + R\,\lambda^{h}_{\cdot\cdot}\,\lambda^{\cdot i}_{\cdot\cdot}\right) \qquad (47)$$

Now from our original equation (43):

$$\varphi^{h\,i}_{j\,k} = \frac{1}{S}\,\lambda^{h\,i}_{j\,k} - \frac{R}{S}\,\varphi^{h\,i}_{j\,k} \times \varphi^{h\,i}_{j\,k}$$

and the cross product on the right may be calculated from the equations (46) and (47). This, then, gives the unique solution to the problem.

Generalizing these methods to systems of simultaneous equations in more than one unknown and with any number of gene loci is not difficult. Of course, the equations become larger and more cumbersome, but no new theoretical difficulties appear.

V. Lethal Factors and Selection

The results of a selective action may be calculated by the methods we have developed. Suppose the chances of survival of the types A_h A_i are R^h_i where h and i take values over all alleles. Then starting with a population λ^h_i the population reaching maturity will be

$$\frac{1}{D_1} R^h_i \lambda^h_i$$

where D_1 is a normalizing factor given by

$$D_1 = \sum_h \sum_i R^h_i \lambda^h_i \qquad (48)$$

Sums on more than one index are of frequent occurence in selection work and we adopt the convention that summation on two or more indices simultaneously will be indicated by placing a bar over these indices. Thus (48) would be written

$$D_1 = R^{\bar{h}}_{\bar{i}} \lambda^{\bar{h}}_{\bar{i}}$$

The first generation offspring of λ^h_i would be

$$\mu^h_i = \frac{1}{D_1^2} (R^h_i \lambda^h_i) \times (R^h_i \lambda^h_i)$$
$$= \frac{1}{D_1^2} (R^h_s \lambda^h_s R^t_i \lambda^t_i)$$

and if

$$D_2 = D_1^2 R^{\bar{h}}_{\bar{i}} \mu^{\bar{h}}_{\bar{i}}$$

then

$$\frac{1}{D_2} R^h_i R^h_s \lambda^h_s R^{\bar{t}}_i \lambda^{\bar{t}}_i$$

will reach maturity.

The next generation is given by

$$\frac{1}{D_{\bar{\imath}}} R^{\ell}_{\bar{\jmath}} R^{\ell}_{\bar{s}} \lambda^{\ell}_{\bar{s}} R^{\bar{t}}_{\bar{\jmath}} \lambda^{\bar{t}}_{\bar{\jmath}} R^{\bar{\imath}}_{i} R^{\bar{\kappa}}_{\bar{s}} \lambda^{\bar{\kappa}}_{s} R^{\bar{t}}_{i} \lambda^{\bar{t}}_{i}$$

etc.

A population will be in equilibrium if and only if the equation

$$\lambda^{\ell}_{i} = \frac{1}{D} R^{\ell}_{i} \lambda^{\ell}_{\cdot} \lambda^{i}_{\cdot} \tag{49}$$

is satisfied. This requires that

$$\lambda^{\ell}_{\cdot} = \frac{1}{D} R^{\ell}_{\bar{\imath}} \lambda^{\ell}_{\cdot} \lambda^{\bar{\imath}}_{\cdot} \tag{50}$$

or

$$\lambda^{\ell}_{\cdot} \left(\frac{1}{D} R^{\ell}_{\bar{\imath}} \lambda^{\bar{\imath}}_{\cdot} - 1 \right) = 0 \tag{51}$$

This may be satisfied in two different ways. If none of the $\lambda^{\ell}_{\cdot} = 0$, we must have

$$R^{\ell}_{\bar{\imath}} \lambda^{\bar{\imath}}_{\cdot} = D \tag{52}$$

a system of linear algebraic equations with the unique solution (providing $|R^{\ell}_{i}| \neq 0$)

$$\lambda^{\ell}_{\cdot} = \kappa \begin{vmatrix} R^{1}_{1} & R^{1}_{2} & \cdots & R^{1}_{\ell-1} & R^{1}_{\ell+1} & \cdots & R^{1}_{n} \\ R^{2}_{1} & R^{2}_{2} & \cdots & R^{2}_{\ell-1} & R^{2}_{\ell+1} & \cdots & R^{2}_{n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ R^{n}_{1} & R^{n}_{2} & \cdots & R^{n}_{\ell-1} & R^{n}_{\ell+1} & \cdots & R^{n}_{n} \end{vmatrix} \tag{53}$$

where k is a constant determined by the condition $\lambda_i = 1$ Equation (51) may also be satisfied if some of the $\lambda_i^h = 0$. We may not divide though by these components, but the remainder gives us a set of, say, $m < w$ equations in the m nonvanishing components which will, in general, have a unique solution.

A population satisfying (49) above would be in equilibrium in the sense that the first generation expected offspring would be of the same genetic constitution. However, this equilibrium may not be stable, for the actual offspring will in general deviate somewhat from the expected offspring and the population will be stable if and only if this deviation tends to cause the next generation to return to the equilibrium position. The situation may be likened to a ball which may be either balanced on the top of a hill or placed at the lowest point in a valley. In either case the ball is in equilibrium, but only in the valley is it stable, for if given a slight displacement on the hill the ball will tend to run down, while in the valley it tends to return to the lowest point.

Although a complete set of necessary and sufficient conditions for stability of a population have not been found, we have the following proposition:

Theorem XIV. A necessary condition for stability of a realizable equilibrium population λ_i^q (no gene

frequency equals zero) under the selective action R_{i}^{h} is that

$$R_{h}^{h} < R_{s}^{h} \qquad s \neq h \tag{54}$$

Proof: Let the $R_{\cdot}^{h}$ coefficients be multiplied by such a constant that the equilibrium population satisfies the normalized equation:

$$\lambda_{\cdot}^{h} = R_{\cdot}^{h} \lambda_{\cdot}^{h} \lambda_{\cdot}^{\cdot} \tag{55}$$

or, since no $\lambda_{\cdot}^{h} = 0$

$$R_{\cdot}^{h} \lambda_{\cdot}^{\cdot} = 1 \tag{56}$$

Let this population take on a small increment $\Delta\lambda_{\cdot}^{h}$ with $\Delta\lambda_{\cdot}^{\cdot} = 0$. The result of one generation random intermating of this displaced population $\lambda_{\cdot}^{h} + \Delta\lambda_{\cdot}^{h}$ is

$$\mu_{\cdot}^{h} = \frac{1}{D} R_{\cdot}^{h} (\lambda_{\cdot}^{h} + \Delta\lambda_{\cdot}^{h})(\lambda_{\cdot}^{\cdot} + \Delta\lambda_{\cdot}^{\cdot})$$

Whence

$$\mu_{\cdot}^{h} = \frac{1}{D} R_{\cdot}^{h} (\lambda_{\cdot}^{h} + \Delta\lambda_{\cdot}^{h})(\lambda_{\cdot}^{\cdot} + \Delta\lambda_{\cdot}^{\cdot})$$

$$= \frac{1}{D} \left[R_{\cdot}^{h} \lambda_{\cdot}^{h} \lambda_{\cdot}^{\cdot} + R_{\cdot}^{h} \lambda_{\cdot}^{h} \Delta\lambda_{\cdot}^{\cdot} + R_{\cdot}^{h} \lambda_{\cdot}^{\cdot} \Delta\lambda_{\cdot}^{h} + R_{\cdot}^{h} \Delta\lambda_{\cdot}^{h} \Delta\lambda_{\cdot}^{\cdot} \right]$$

Now the first term $R_{\cdot}^{h} \lambda_{\cdot}^{h} \lambda_{\cdot}^{\cdot} = \lambda_{\cdot}^{h}$,

and the third $R_{\cdot}^{h} \lambda_{\cdot}^{\cdot} \Delta\lambda_{\cdot}^{h} = \Delta\lambda_{\cdot}^{h}$ from (56). For small increments the last term is of the second order

and may be neglected so that

$$\mu^{h}_{\cdot} = \frac{1}{D}\left[\lambda^{h}_{\cdot} + \Delta\lambda^{h}_{\cdot} + R^{h}_{\bar{\imath}}\,\lambda^{h}_{\cdot}\,\Delta\lambda^{\bar{\imath}}_{\cdot}\right]$$

To evaluate the constant D, we first sum on h.

$$\mu^{\cdot}_{\cdot} = \frac{1}{D}\left[\lambda^{\cdot}_{\cdot} + \Delta\lambda^{\cdot}_{\cdot} + R^{\bar{h}}_{\bar{\imath}}\,\lambda^{\bar{h}}_{\cdot}\,\Delta\lambda^{\bar{\imath}}_{\cdot}\right]$$

$$1 = \frac{1}{D}\left[1 + 0 + \Delta\lambda^{\bar{\imath}}_{\cdot}\right] = \frac{1}{D}$$

thus

$$\mu^{h}_{\cdot} = \lambda^{h}_{\cdot} + \Delta\lambda^{h}_{\cdot} + R^{h}_{\bar{\imath}}\,\lambda^{h}_{\cdot}\,\Delta\lambda^{\bar{\imath}}_{\cdot}$$

We see that the offspring of the displaced population is equal to this population plus an additional increment $R^{h}_{\bar{\imath}}\,\lambda^{h}_{\cdot}\,\Delta\lambda^{\bar{\imath}}_{\cdot}$. Clearly a necessary condition for stability is that this be opposite in sign to the original increment $\Delta\lambda^{h}_{\cdot}$. Now the original increment was completely arbitrary. Let us fix h and s as two constant indices and suppose the components of the increment were $\Delta\lambda^{h}_{\cdot} = +\epsilon$, $\Delta\lambda^{s}_{\cdot} = -\epsilon$ and all other components zero. Taking ϵ positive we have the condition

$$R^{h}_{\bar{\imath}}\,\lambda^{h}_{\cdot}\,\Delta\lambda^{\bar{\imath}}_{\cdot} < 0$$

or

$$\left(R^{h}_{1}\,\Delta\lambda^{1}_{\cdot} + R^{h}_{2}\,\Delta\lambda^{2}_{\cdot} + \ldots + R^{h}_{n}\,\Delta\lambda^{n}_{\cdot}\right)\lambda^{h}_{\cdot} < 0$$

For a realizable population $\lambda^{h}_{\cdot}$ is positive and all the terms in the parentheses are zero except $\Delta\lambda^{h}_{\cdot}$ and $\Delta\lambda^{s}_{\cdot}$ giving

$$R^{h}_{h}\,\epsilon - R^{h}_{s}\,\epsilon < 0$$

or $$R^{h}_{h} < R^{h}_{s}$$

This proves the theorem.

For a dimorphic factor this condition is also easily shown to be sufficient for stability; but examples show that it is not always sufficient for more than two alleles. Sufficient conditions (not necessary) for any number of alleles are that

$$R^{h}_{h} = K_1 \quad \text{, a constant independent of h} \tag{57}$$

$$R^{h}_{s} = K_2 \quad \text{, a constant independent of h and s}$$

$$K_1 < K_2$$

For then the correction term is

$$\lambda^{h}_{\cdot} \left[R^{h}_{s} \sum_{\substack{i=1 \\ i \neq h}}^{n} \Delta \lambda^{i}_{\cdot} + R^{h}_{h} \Delta \lambda^{h}_{\cdot} \right]$$

$$= \lambda^{h}_{\cdot} \left[- R^{h}_{s} \Delta \lambda^{h}_{\cdot} + R^{h}_{h} \Delta \lambda^{h}_{\cdot} \right]$$

$$= \lambda^{h}_{\cdot} \Delta \lambda^{h}_{\cdot} (K_1 - K_2)$$

which is clearly opposite in sign and less in absolute value than $\Delta \lambda^{h}_{\cdot}$

VI. A Calculus of Populations

Up to the present, all our population symbols have been constants, i.e. each represented a certain particular population. The manipulation of these discrete sets of numbers constitutes an algebra. Sometimes, however, it is convenient to consider continuous time variations of a population. Such a study leads to a calculus of populations. We have already used, in the preceding section, the idea of an incremental population.

In this section we will define the "derivative" of a population and develop some of the fundamentals of the calculus.

First let us generalize our idea of population to include variable populations, i.e. populations that are functions of time. We indicate this functional dependence by the usual notation e.g.

$$\lambda_{j\,\kappa}^{h\,i}(t) \tag{58}$$

represents the genetic constitution of the population $\lambda_{j\,\kappa}^{h\,i}$ at the time t. In case no ambiguity is introduced we will sometimes omit the argument t, it being understood that

$$\lambda_{j\,\kappa}^{h\,i} = \lambda_{j\,\kappa}^{h\,i}(t)$$

We define the derivative of a population as the indexed symbol whose components are the time derivatives of the components of the population in question. Thus:

$$\frac{d}{dt}\lambda_{jk}^{hi}(t) = \lim_{\Delta t \to 0} \frac{\lambda_{jk}^{hi}(t+\Delta t) - \lambda_{jk}^{hi}(t)}{\Delta t} \tag{59}$$

We assume the population large enough and the variation of λ smooth enough for the limit to exist in a practical sense.

Note that the derivative of a population is <u>not</u> a population. A population has the property $\lambda_{..}^{..} = 1$ while its derivative has the following property:

<u>Theorem XV.</u>

$$\frac{d}{dt}\lambda_{..}^{..} = 0 \tag{60}$$

This is true if we first sum on all indices and then take the derivative or vice versa. Both follow immediately on taking the derivative of (6).

As in ordinary calculus we have simple rules for taking derivatives of sums, products, etc. These are all exactly the same as those of ordinary calculus. We have:

Theorem XVI.

1. If $\lambda_{j\kappa}^{h\,i}(t)$ is a constant

$$\frac{d}{dt}\,\lambda_{j\kappa}^{h\,i} = 0 \tag{61}$$

2. If

$$\lambda_{j\kappa}^{h\,i} = R_1 \mu_{j\kappa}^{h\,i} + R_2 \nu_{j\kappa}^{h\,i}$$

then (62)

$$\frac{d}{dt}\,\lambda_{j\kappa}^{h\,i} = R_1 \frac{d}{dt}\,\mu_{j\kappa}^{h\,i} + R_2 \frac{d}{dt}\,\nu_{j\kappa}^{h\,i}$$

3. If

$$\lambda_{j\kappa}^{h\,i} = \mu_{j\kappa}^{h\,i} \times \nu_{j\kappa}^{h\,i} \tag{63}$$

then

$$\frac{d}{dt}\,\lambda_{j\kappa}^{h\,i} = \mu_{j\kappa}^{h\,i} \times \frac{d}{dt}\,\nu_{j\kappa}^{h\,i} + \nu_{j\kappa}^{h\,i} \times \frac{d}{dt}\,\mu_{j\kappa}^{h\,i}$$

The first two of these rules for differentiation are obvious, since by fixing the indices they merely state the ordinary rules for differentiating constants and sums. The third, which is the analogue of Leibnitz rule for differentiating a product, requires proof. Starting with the definition of a derivative we have:

$$\frac{d}{dt}\,\lambda_{j\kappa}^{h\,i} = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\left[\lambda_{j\kappa}^{h\,i}(t+\Delta t) - \lambda_{j\kappa}^{h\,i}(t)\right]$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\left[\left\{\mu_{j\kappa}^{h\,i}(t) + \Delta\mu_{j\kappa}^{h\,i}(t)\right\} \times \left\{\nu_{j\kappa}^{h\,i}(t) + \Delta\nu_{j\kappa}^{h\,i}(t)\right\} - \mu_{j\kappa}^{h\,i}(t) \times \nu_{j\kappa}^{h\,i}(t)\right]$$

where $\Delta \mu_{j\kappa}^{h\,i}(t) = \mu_{j\kappa}^{h\,i}(t+\Delta t) - \mu_{j\kappa}^{h\,i}(t)$ and similarly for $\Delta \nu_{j\kappa}^{h\,i}(t)$. Now the first cross product may be multiplied out by our distributive law (Theorem VIII) giving

$$\lim_{\Delta t \to 0} \left[\mu_{j\kappa}^{h\,i}(t) \times \frac{\Delta \nu_{j\kappa}^{h\,i}(t)}{\Delta t} + \nu_{j\kappa}^{h\,i}(t) \frac{\Delta \mu_{j\kappa}^{h\,i}(t)}{\Delta t} + \Delta \mu_{j\kappa}^{h\,i}(t) \times \frac{\Delta \nu_{j\kappa}^{h\,i}(t)}{\Delta t} \right]$$

The third term in general tends to zero with Δt so that our limit is:

$$\mu_{j\kappa}^{h\,i} \frac{d}{dt} \nu_{j\kappa}^{h\,i} + \nu_{j\kappa}^{h\,i} \frac{d}{dt} \mu_{j\kappa}^{h\,i}$$

the desired result.

For a population $\lambda_{j\kappa}^{h\,i}$ intermating at random this reduces to

$$\frac{d}{dt}\left(\lambda_{j\kappa}^{h\,i} \times \lambda_{j\kappa}^{h\,i}\right) = 2\, \lambda_{j\kappa}^{h\,i} \times \frac{d}{dt} \lambda_{j\kappa}^{h\,i}$$

A population whose components are analytic functions of time may be expanded in a Taylor series

$$\lambda_{j\kappa}^{h\,i}(t) = \lambda_{j\kappa}^{h\,i}(0) + \frac{d}{dt} \lambda_{j\kappa}^{h\,i}\Big|_{t=0} \cdot t + \frac{d^2}{dt^2} \lambda_{j\kappa}^{h\,i}\Big|_{t=0} \frac{t^2}{2} + \ldots$$

for by fixing the indices we are again merely stating the standard Taylor Theorem.

Bibliography

Inasmuch as no work has been done previously along the specific algebraic lines indicated in this thesis, our references must be of a fairly general nature. For a good introductory treatment of genetics, we recommend:

Sinnott, E.W. and Dunn, L.C. Principles of Genetics. McGraw Hill, 1932.

For mathematical treatment of certain phases of hereditary phenomena, the following may be consulted:

Fisher, R.A. The genetical theory of natural selection. Oxford Clarendon Press, 1930, Pp. xiv 272.

Wright, Sewall. Genetics, 1921, V.6, Systems of Mating. II. The effects of inbreeding on the genetic composition of a population.

Ibid. Pp.162-166. IV. The effects of selection.

Haldane, J.B.S. The cause of evolution. London, Longmans Green, 1932.

Robbins, R.B. Some applications of mathematics to breeding problems. I. Genetics, 1917, 2, 489-504; II. Ibid, 1918, V.3, 73-92; III. Ibid, 1918, V.3, 375-389.

Hogben, L. A matrix notation for Mendelian populations. Proc. Royal Soc. Edinburgh, 1933, 53, 7-25.

Biography of the Author

Claude E. Shannon was born in Petoskey, Michigan, April 30, 1916. The first sixteen years of his life were spent in Gaylord, Michigan, and there he attended the Public School, graduating from the Gaylord High School in 1932. He entered the University of Michigan in the fall of that year and took an engineering course, specializing in Electrical Engineering and Mathematics. While a senior, he was elected a member of Phi Kappa Phi and an associate member of Sigma Xi. In 1936, he obtained the degrees of Bachelor of Science in Electrical Engineering and Bachelor of Science in Mathematics from the University of Michigan and in the same year started working as a Research Assistant in Electrical Engineering at the Massachusetts Institute of Technology.

From September, 1936 to June, 1938, he was in charge of the Differential Analyzer at the Massachusetts Institute of Technology, a machine used for obtaining numerical solutions to differential equations. During the summer of 1938, he did research work on the design of the Bush Rapid Selector, chiefly connected with the vacuum tube circuits employed in this device. In September, he left the Electrical Engineering Department to become an Assistant in the Department of Mathematics. In the spring of 1939, he was elected to full membership of Sigma Xi. The summer of 1939

was spent at the biological station of the Carnegie Institution of Washington at Cold Spring Harbor, New York, in the study of the application of algebraic methods to hereditary phenomena. In the autumn, he returned to the Massachusetts Institute of Technology as Bolles Fellow and continued work toward a degree of Doctor of Philosophy in Mathematics. In January, 1940, he was married. In the same month, he was awarded the Alfred Noble Prize, of the combined engineering societies of the United States, an award given each year to a person not over thirty for a paper published in one of the journals of the participating societies. This was given in recognition of the paper "A Symbolic Analysis of Relay and Switching Circuits". That spring, he was elected to membership in the American Mathematical Society as an institutional nominee of the Massachusetts Institute of Technology. He is also an associate member of the American Institute of Electrical Engineers and of the Institute of Radio Engineers. He was recently awarded a National Research Fellowship in Mathematics for the year 1940-41.

Published Papers:

"A Symbolic Analysis of Relay and Switching Circuits", Transactions of the American Institute of Electrical Engineers, 1938. Pp. 713-23

ABSTRACT

of

"AN ALGEBRA FOR THEORETICAL GENETICS"

by

Claude E. Shannon

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

from the

Massachusetts Institute of Technology

1940

In this thesis, an algebra is constructed for studying the dynamics of Mendelian populations. The symbols of the algebra represent groups of individuals or populations. The indexed symbol λ^{hi}_{jk}, for example, represents a population in which two gene loci are under consideration (the number of loci corresponds to the number of pairs of indices). The number of allelomorphs in each locus is completely arbitrary, as is the recombination value for the loci. The different components of a population symbol, represented by fixing the indices at specific values, are numbers whose values correspond to the fractions of the population with certain genetic formulae. It is convenient in some cases to consider as populations symbols whose components are negative or even complex. Such symbols cannot, of course, represent an actual group of individuals and are called <u>unrealizable</u> populations, but their use sometimes facilitates the solution of problems.

Addition of two population symbols, $R\lambda^{hi}_{jk} + S\mu^{hi}_{jk}$, results in a third population symbol which is defined in such a way as to represent the population obtained by merely combining the original populations in fractional proportions corresponding to the scalar coefficients of R and S. Cross multiplication of population symbols $\lambda^{hi}_{jk} \times \mu^{hi}_{jk}$ gives a population symbol which

is defined in such a way as to represent the expected offspring population when the two original populations are crossmated at random. When two gene loci are considered, this is realized by the mathematical definition

$$\lambda^{hi}_{jk} \times \mu^{hi}_{jk} = \tfrac{1}{2}\left[p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h}_{\cdot i}\right]\left[p_0\mu^{jk}_{\cdot\cdot} + p_1\mu^{j}_{\cdot k}\right]$$
$$+ \tfrac{1}{2}\left[p_0\lambda^{jk}_{\cdot\cdot} + p_1\lambda^{j}_{\cdot k}\right]\left[p_0\mu^{hi}_{\cdot\cdot} + p_1\mu^{h}_{\cdot i}\right]$$

in which $p_1 = 1-p_0$ is the recombination value for the two loci, and replacing an index by a dot indicates summation of the population symbol on that index. Cross multiplication is defined analogously for n loci. It is shown that this algebra is commutative on addition and multiplication, distributive, and associative on addition but not on multiplication. These laws together with two fundamental manipulation theorems: one, that summation of a population on all indices gives unity and two, that inverting the upper and lower rows of indices of a population leaves it unchanged, form the basic algorisms of the algebra.

A number of the well known theorems of theoretical genetics are easily proved by means of this algebra. In addition, a number of new results are found. Completely general formulae for the nth generation offspring under random intermating of an arbitrary initial population are developed both for the cases of two and of three linked

factors. For two linked factors, the formula for the nth generation is

$$\mu^{hi}_{jk} = \left[p_0^{n-1}(p_0\lambda^{hi}_{\cdot\cdot} + p_1\lambda^{h\cdot}_{\cdot i}) + (1-p_0^{n-1})\lambda^{h\cdot}_{\cdot\cdot}\lambda^{\cdot i}_{\cdot\cdot}\right]$$
$$\left[p_0^{n-1}(p_0\lambda^{j\cdot}_{\cdot\cdot} + p_1\lambda^{j\cdot}_{\cdot k}) + (1-p_0^{n-1})\lambda^{j\cdot}_{\cdot\cdot}\lambda^{\cdot k}_{\cdot\cdot}\right]$$

in which λ^{hi}_{jk} is the initial population and p_1 the recombination value. Incidental to this, it is shown that a recombination value $> \frac{1}{2}$ is impossible when there is no interference. Conditions are found for the stability under random intermating of a population when one or more loci are considered. For the case of one locus, three sets of equivalent necessary and sufficient conditions are established.

By means of certain homogeneous γ populations an arbitrary population may be expanded in a finite series displaying its various components. This expansion, together with the multiplication law for the γ populations, displays the elements of this algebra as hypercomplex numbers. It is shown that an arbitrary population may be expanded uniquely as a sum of any n linearly independent populations where n is the number of different possible genetic formulae for the factors considered.

It is possible to write down various types of equations involving known and unknown populations using the

operations of addition and cross multiplication. In general, such an equation can be interpreted as a breeding experiment involving one or more unknown populations and resulting in a genetically known population. Methods are developed whereby most such equations can be solved in case a solution exists. Briefly this method of solution may be summarized as follows. By summing on one or more indices of the unknown populations, enough data about them is obtained to uniquely determine any cross products in which they appear. The cross product terms in the original equations then become known and the equations may be sloved in exactly the same way as ordinary linear algebraic equations.

In case a selective action exists favoring individuals of a certain genetic constitution, the previous formulae for stability no longer hold. Although this more difficult problem has not been completely solved, a necessary condition for the possible existence of a stable population under an arbitrary given selective action is established, and a formula for this population is developed. This has only been done for the case of a single locus.

A start has been made toward the development of a calculus of populations, i.e. the study of populations which may vary continuously with time. The time derivative of a population is defined. The derivative of a population, although an indexed symbol, is not itself a popu-

lation. All the ordinary rules of derivation including the Leibnitz rule for the derivative of a cross product of populations are shown to hold true. Also, a population may be expanded in a Taylor series in powers of time, of the same form as the ordinary Taylor series.