the sum of these four values equal to one) then $p_{12}^{(n)}\ldots_6 \to p_{12}p_{34}p_{56}$. If however for $m = 6$ merely $d_{12} = d_{34} = 1$ (realized if, in a notation analogous to (3), v_0, v_5, v_6, v_{56}, v_{12}, v_{34}, v_{125}, v_{126} are the only non-zero values of the l.d.) then $p_{12}^{(n)}\ldots_6 \to p_{12}p_{34}p_5p_6$.

In general, with a proof which consists in a modification of the reasoning (p. 41), of my earlier paper, we may state the following complement to the main limit theorem (9): *If the l.d. is such that $r < m$ disjoint groups G_1, G_2, $\cdots$ G_r of completely linked characters exist, i.e. such that within each group no crossover takes place, each group containing as many of the m numbers as compatible with the definition but not less than two, and all groups together containing $s \leqq m$ of the m elements, then, as $n \to \infty$, $p_{12}^{(n)}\ldots_m$ converges towards the product of those marginal distributions (of the original generation) which correspond to these groups multiplied by the marginal distributions of order one of the remaining free elements which are not contained in any such group. In a formula:*

$$\lim_{n\to\infty} p_{G_1,G_2,\cdots G_r,\gamma_{s+1},\gamma_{s+2},\cdots\gamma_m} = p_{G_1}p_{G_2}\cdots p_{G_r}p_{\gamma_{s+1}}p_{\gamma_{s+2}}\cdots p_{\gamma_m}. \tag{10}$$

We may also characterize these linked groups of maximum size by stating that while within each group no crossover takes place there must be at least one c.p. $\neq 0$ among any two such groups and at least one among any group and any free element. It may however be noted that if there is one c.p. > 0 among two groups of complete linkage (or among a group and a free element) then all c.p.'s among these two groups are different from zero. In fact, it follows by repeated use of the triangular relation (2) that if one c.p. among two disjoint groups of complete linkage is zero, all of them are zero. If, e.g., $(1, 2, 3)$ and $(5, 6, 8)$ are two groups of complete linkage, i.e. $v_{123}(000) = 1$ and $v_{568}(000) = 1$ and if besides $c_{15} = 0$, then $v_{123568}(000000) = 1$ and these six elements form a group of complete linkage.

It may be noticed that the above statement of the generalized limit theorem becomes simpler and more elegant by counting "free elements" as groups. It might then run as follows: *If G_1, G_2, $\cdots$ $G_t (t \leqq m)$ are the maximal groups of completely linked characters, then, under the hypotheses of the earlier paper, the gene distribution in successive generations approaches a limit in which the original (marginal) probabilities within each group G_i are preserved and genes and sets of genes fromd ifferent groups are independently distributed.*

ON THE DEFINITION OF DISTANCE IN THE THEORY OF THE GENE

By Hilda Geiringer
Wheaton College

In several letters to this author Dr. I. M. H. Etherington of the University of Edinburgh has raised questions concerning the author's definition of "distance" proposed in Section 10 of her paper on Mendelian heredity,[1] comparing it with

[1] *Annals of Math. Stat.*, Vol. 15 (1944), pp. 25–57.

the definition implicit in Professor J. B. S. Haldane's earlier treatment.[2] The main content of the author's paper consists of some general limit theorems and the integration of a certain system of difference equations. The distance definition is a by-product subject to discussion.

"Distance" d_{ij} between two genes i and j is defined by the author as the mathematical expectation of the number of crossovers in the interval (i, j) with respect to the "linkage distribution" (l.d.). This basic concept is introduced as follows (page 32): If S is the set of numbers $1, 2, \cdots m$ (m being the number of Mendelian characters), A any subset of A and $A' = S - A$, we denote by $l(A)$ the probability that an individual with "maternal" genes $x_1, \cdots, x_m$ and paternal genes $y_1, \cdots, y_m$ transmit the paternal genes belonging to A and the maternal genes belonging to A'. These 2^m probabilities constitute the l.d. From these definitions the equality (G. (53′))

$$d_{ij} = c_{i,i+1} + c_{i+1,i+2} + \cdots + c_{j-1,j} \qquad (i < j) \tag{1}$$

is derived, where c_{ij} is the probability of a "crossover" (c.p.) in (i, j). This distance has the required additivity: (G. (54))

$$d_{ij} + d_{jk} = d_{ik}, \qquad (i < j < k). \tag{2}$$

Etherington points out that the term "distance" has an established currency in genetics being the basis on which chromosome maps are constructed, and that there is a standard method of calculating it in accordance with which (1) is an "approximation valid only when the adjacent c.p.'s are small." Moreover "the biological uniqueness has been lost for the value of d_{ij} now depends on the particular set of intermediate genes which we happen to be considering. If any of them are omitted from consideration then the inequality (G. (13)).

$$c_{ij} + c_{jk} \geqq c_{ik} \tag{3}$$

shows that in general d_{ij} is diminished while if new genes are taken into consideration d_{ij} may increase." "In order that d_{ij} should not depend on a particular choice of intermediate genes the word 'crossover' in the definition given would have to be interpreted as 'chiasma' instead of 'odd number of chiasmata'; and then d_{ij} cannot be evaluated in terms of the l.d. alone without further assumptions regarding the interference of crossovers."

The point of view adopted in the author's paper was to regard the *l.d.* as the basis from which everything else has to be inferred. The number m of Mendelian characters is considered constant and the distance, being a mathematical expectation with respect to the l.d. necessarily depends on it. In this conception distance is not a geometric property which can be measured for any two genes independently but rather a system of $m(m-1)/2$ consistent numbers associated to the m genes. There is no choice regarding the intermediate genes to be taken into consideration; *all* known genes are to be considered, i.e. one has to use the available relevant information in order to determine the l.d., the c.p.'s and the

[2] Quotation [4a] in the author's paper. References to these papers will be distinguished by the initials *H* and *G*.

distances. If the information is incomplete the results will be provisional and subject to change; if it is satisfactory the same will be true for the distances. Thus it is nothing but natural that d_{ij} is changed if some genes are omitted from consideration, or if new genes are discovered. In this set up "crossover"—defined by means of the marginal distributions of second order of the l.d.—means a transition from the paternal to the maternal set or vice versa. (Expressed in terms of the chiasma-hypothesis this means "odd number of chiasmata between adjacent genes.") Additional assumptions "regarding the interference of crossovers" are neither necessary nor admissible. All this is contained in the l.d.

Haldane's approach as translated by Etherington into the author's notation is as follows. "The genes are considered to be distributed continuously along a chromosome. Thus this approach unlike G.'s is not based on the l.d. of a finite set of genes. We must think of one suffix, i, as referring to a gene at a fixed locus on the chromosome, the others to variable loci, so that the c.p.'s are variable. For any three genes i, j, k a quantity p is defined by the equation

$$c_{ik} = c_{ij} + c_{jk} - pc_{ij}c_{jk}, \qquad (i < j < k) \tag{4}$$

Biological considerations show that p is a number between 0 and 2 (small when c_{ij} and c_{jk} are both small, increasing, on the whole, with $c_{ij} + c_{jk}$). The distance D_{ij} is defined by the statement

$$D_{kj}/c_{kj} \to 1 \quad \text{as } k \text{ approaches } j \qquad (c_{kj} \to 0), \tag{5}$$

together with the additive property, and from this with (4) Haldane's general distance expression is derived:

$$D_{ij} = \int_0^{c_{ij}} \frac{dc_{ij}}{1 - p_0 c_{ij}}. \tag{6}$$

Here $p_0 \equiv p_0(c_{ij})$ denotes the limiting form of p when k approaches j, and represents biologically a property of the chromosome segment (i, j), a measure of interference. Any suitable specification of this function $p_0(c_{ij})$ would constitute a mathematical 'model' of the chromosome. If p were constant we should have $p_0 = p$ and

$$D_{ij} = -\frac{1}{p} \log (1 - pc_{ij}). \tag{7}$$

Both Haldane and Geiringer considered the special cases $p = 2$ (no interference) and $p = 0$ (complete interference) for which respectively

$$D_{ij} = -\tfrac{1}{2} \log (1 - \tfrac{1}{2} c_{ij}) \tag{7'}$$

$$D_{ij} = c_{ij} = d_{ij}. \tag{7''}$$

Since p is always between 0 and 2 Haldane concludes that the true value of D_{ij} is between (7') and (7''), and he gives reasons for saying that (7') is nearly correct for genes 'far apart,' (7'') for genes 'close together.' "

If the author is right, this seems to be the standard definition accepted in genetics as mentioned above by Etherington. A few, not exhaustive, comments may be added. Writing in (6) t for the variable of integration and $p_0 = p_0(t)$ it is seen that the expression

$$D_{ij} = \int_0^{c_{ij}} \frac{dt}{1 - tp_0(t)} \tag{6}$$

contains the unknown function $p_0(t)$, which is unspecified except for the statement that it is bounded between 0 and 2. It is immediately seen that with an arbitrary $p_0(t)$ and without a restriction taking the place of (4) this distance (6) will not be additive in the sense of (2). By imposing, after a choice of $p_0(t)$, appropriate restrictions on the c_{ij} additivity may be achieved. For instance in the particular case $p_0(t) = p = \text{const}$, (2) holds by virtue of (4). For such a set of restrictions it has then to be proved that the corresponding "model" is "consistent," i.e. that the so restricted c.p.'s form a compatible set of marginal distributions of second order of an m-variate distribution, the l.d.

These different points will be exemplified presently by studying the particular case $p_0(t) = p$, where p is a suitably chosen constant; the parameter p is to be fitted to the observations under consideration. It may be impossible to reproduce a set of observations satisfactorily if one parameter only is available. In fact, Haldane's paper suggests that it is not only the particular case $p = \text{const}$ he has in mind. It seems however that if D_{ij} is given by (6) with a non constant $p_0(t)$, complicated and perhaps (biologically) not very meaningful conditions may have to be introduced in order to assure additivity of the distances and consistency of the respective model. This author was unable to work out examples of more general and at the same time appropriate and fairly simple assumptions for the unknown function $p_0(t)$.

If $p = \text{const}$, then (7) under the restriction (4) furnishes an additive distance definition because:

$$\begin{aligned} -p[D_{ij} + D_{jk}] &= \log(1 - pc_{ij}) + \log(1 - pc_{jk}) \\ &= \log(1 - pc_{ij} - pc_{jk} + p^2 c_{ij}c_{jk}) = \log(1 - pc_{ik}) = -pD_{ik}, \end{aligned}$$

because of (4). Let us now investigate whether there is a consistent system of c.p.'s satisfying (4). Put, as in G.(48), $c_{i,i+1} = p_i$, combine (4) with G.(50) and write $p = 2\epsilon$. It follows that (4) is satisfied with $0 \leqq \epsilon \leqq 1$, if:

$$p_{ij} = \epsilon p_i p_j, \qquad p_{ijk} = \epsilon^2 p_i p_j p_k, \cdots. \tag{8}$$

Here p_{ij} is the probability of the simultaneous occurrence of the "events" numbered i and j, etc. For $\epsilon = 0$ we get "disjoint events" (see G.i) for the discussion of consistency). Assume now $\epsilon > 0$. By some considerations, analogous to those p. 54 G, the following necessary and sufficient condition of consistency follows:

$$\prod_{i=1}^{m-l} (1 - \epsilon p_i) \geqq 1 - \epsilon \qquad (\epsilon > 0). \tag{9}$$

This restriction (not considered by Haldane or Etherington) is, of course, relevant. If e.g. $m = 3$, $p_1 = p_2 = 4/5$, then ϵ must be $\geqq 15/16$; or if $m = 4$, $p_1 = p_2 = p_3 = \frac{1}{2}$, $\epsilon \geqq 3 - \sqrt{5}$ results. The restriction required by the "linear theory" is

$$p_i \leqq \frac{1}{2\epsilon}, \qquad (i = 1, 2, \cdots, m-1). \tag{10}$$

Hence this model is consistent under certain restrictions. It is, in contrast to Etherington's contention, different from iii) G. p. 54. The corresponding distance definition (7) is different from the author's. The D_{ij} thus defined are additive, and D_{ij} depends on c_{ij} only and not on the intermediate genes. The author's definition of distances, d_{ij}, is general, additive and seems to the author to be well adapted to the biological situation; since the definition of d_{ij} is not related to any particular model it is compatible with any model, which may contain any desired—consistent—assumptions about "interference," etc. For example in G. iv) p. 55, an n-parametric model has been suggested which seems fairly flexible.

It may however seem more acceptable to the biologist not to use a general distance definition but to define "distance" merely in relation to some sufficiently general "model" (such that the distance definition would vary with the model), instead of accepting an all-over definition as ventured in the author's paper. The particular model (8) in connection with its related distance definition (7) might give an example of such an approach.[3, 4]

---

[3] As Etherington remarks, eq. (14′) in the author's original paper is not correct. One can only state that (47) holds. The mistake is however without consequence since no conclusions are drawn from (14′). The same mistake was pointed out by Professor Kai Lai Chung.

[4] Etherington writes: "I have been kindly allowed to read Professor Geiringer's MS. and feel that some comments are necessary.

The standard procedure for calculating the distance between two linked genes is as follows. A selection of intermediate genes is taken and the adjacent crossover values calculated, giving a provisional estimate of the distance as in Geiringer's formula (1). When further intermediate genes are added to the selection, it is found that the provisional distance increases, but there is apparently a maximum value beyond which it cannot be increased. This unknown maximum value is the distance, and the geneticist accepts (1) as the distance when he is sure that he has observed a sufficient number of intermediate genes to give a good enough approximation to the true distance. Thus Geiringer's formula (1) gives the geneticist's true distance only on the understanding that it includes all genes intermediate between i and j; but generally speaking the great majority of these genes may be unobservable in the sense that they have no observably distinct alleles by means of which the c.p.'s could be calculated, though from time to time fresh genes may become observable by mutation.

In some cases the above procedure fails because not enough intermediate genes can be observed; then Haldane's analysis is useful. It should be emphasized that his distance is additive by definition. (For a geometrical analogy, think of the genes as points closely distributed along a curve, chords representing c.p.'s. Haldane's definition of the distance is analogous to defining arc length of the curve as a limiting sum of chords.) In my tran-

scription of his treatment, I should perhaps have made it clearer that the derived formula (6) gives only the distance D_{ij} measured from the initially chosen and fixed gene i to an arbitrary gene j. Other distances D_{jk}, $(i < j < k)$, are deduced from it by the postulate of additivity $(D_{jk} = D_{ik} - D_{ij})$. If the origin i is changed, there will be a similar formula (6), but it should not be assumed that the function p_0 is the same. In referring to certain conditions necessary 'to assure additivity,' Geiringer evidently means conditions that the function p_0 may be the same for all origins i. These conditions would be interpreted biologically as asserting uniformity of interference along the chromosome. I agree that there are further points to be cleared up in this connection.

If I might sum up the discussion, I would say that the geneticist's conception of the distance between genes is an actual property of the corresponding chromosome segment. Geiringer's definition represents the best possible general approach to this from the limited data of the l.d. alone. Haldane's definition fits the geneticist's conception, and his investigation is an attempt to get the best estimate of the distance by making approximate assumptions as to what happens between the observed genes. It is based on the unobservable crossover-distribution of a supposed infinite set of genes, but can be applied to particular models of this infinite c.d. so as to derive results which involve only a finite and observable c.d. Finally it should be mentioned that in the paper quoted, Haldane gave also an alternative method for the case $p = 2$, leading to the same formula (7'), which is really equivalent to defining the distance as the mathematical expectation of the number of chiasmata (not crossovers in G.'s sense) in the interval (i, j)."

A CRITERION OF CONVERGENCE FOR THE CLASSICAL ITERATIVE METHOD OF SOLVING LINEAR SIMULTANEOUS EQUATIONS

By Clifford E. Berry

Consolidated Engineering Corporation, Pasadena, Calif.

The recent development of two devices[1, 2] for solving linear simultaneous equations by means of the classical iterative method[3] has stimulated the writer to investigate convergence criteria for the method. There are in the literature[4] necessary and sufficient criteria for convergence of symmetric systems, and sufficiency criteria for general systems. So far as the writer knows, however, this is the first development of a necessary and sufficient criterion for convergence in the general case. The results obtained are applicable to any arbitrary square non-singular matrix in which $a_{ii} \neq 0$.

Let the set of equations be represented by

$$AX = G, \tag{1}$$

[1] Morgan, T. D., Crawford, F. W., "Time-saving computing instruments designed for spectroscopic analysis", *The Oil and Gas Journal*, August 26 (1944), pp. 100–105.

[2] Berry, C. E., Wilcox, D. E., Rock, S. M., Washburn, H. W., "A computer for solving linear simultaneous equations", to be published.

[3] Hotelling, Harold, "Some new methods in matrix calculation", *The Annals of Mathematical Statistics*, Vol. XIV (1943), pp. 1–34.

[4] Mises, R. von and Pollaczek-Geiringer, Hilda, "Zusammenfassende Berichte. Praktische Verfahren der Gleichungsauflösung". *Zeitschrift für angewandte Math. und Mechanik*, Vol. 9 (1929), pp. 58–77, and 152–164.