A MATHEMATICAL MODEL FOR LENGTHS OF INVERSIONS AND MID-POINTS OF INVERSIONS ASSUMING EQUALLY LIKELY BREAKAGES IN ANY PART OF THE CHROMOSOME AND INDEPENDENCE BETWEEN POINTS OF BREAKAGE

W. T. Federer, R. G. D. Steel, and Bruce Wallace

BU-115-M May, 1960

In many ways the karyotype of a species is remarkably constant. The number of chromosomes carried by each individual member is, with few exceptions, the same as that of all others; even the shape of individual chromosomes as determined by the position of the centromere is usually invariable.

The arrangement of gene loci within a chromosome, on the other hand, often varies considerably from individual to individual. In Dipterous flies, whose giant chromosomes offer an excellent opportunity for study, species after species has been found to possess a wealth of contrasting gene arrangements within the same chromosome. The origin of this variation lies almost exclusively in the inversion of chromosomal segments -- a phenomenon which requires that a chromosome be broken in two places, that the segment between the breaks reverse its position within the chromosome, and that the broken ends heal with the segment in its new position.

Unless the wealth of inversions in a species is overwhelmingly large, a phylogeny of gene arrangements can be constructed such that each gene arrangement differs from another by a single two-break inversion. The logic underlying such phylogenies is supported by the frequent discovery of gene arrangements which were previously postulated as hypothetical ones. We must re-emphasize that phylogenies of this sort can be constructed only in the case of certain Dipteran species. It is a well known fact, nevertheless, that inversions are found in a great many plant and animal species; their origin in these species is undoubtedly comparable to that in Diptera.

Why do wild populations of so many species retain a variety of chromosomal rearrangements? Again a definitive answer is available only in the case of some Drosophila species. Dobzhansky and his students have shown repeatedly that individuals carrying two different gene arrangements are superior to individuals homozygous for one or the other of the various inversions in many, if not all, components of fitness. Although experimental evidence is lacking for other

---

Biometrics Unit, Plant Breeding Department, Cornell University

organisms, the simplest explanation for the existence of any polymorphic system is that based on the selective superiority of heterozygous individuals.

It is highly unlikely that the retention of an inversion in any population depends upon the inversion itself; that is, upon the chromosomal breaks, upon the new arrangement, or upon position effects. Indeed, Paget has shown that radiation-induced inversions have an average deleterious effect on fitness. The advantage of inversion heterozygotes which underlies the retention of two or more inversions in a population appear to result from the interaction of blocks of genes; blocks which are held intact by the suppression of normal gene recombination.

We propose to undertake an analysis of the genetic basis of heterosis by utilizing the above facts. Indeed, Sprague and Chao, independently, have used two inversions in analyzing heterosis in corn. We feel, however, that even greater opportunities for analysis are present in Drosophila. The radiation genetics of Drosophila is a well-developed field. These flies breed rapidly; their giant chromosomes can be analyzed with an accuracy unsurpassed in any other group. There already exists an enormous literature (recently reviewed by daCunha) on the inversions which are to be found in a large number of species. Finally, racial and strain hybrids are known to exhibit heterosis.

In brief, our procedure calls for an analysis of the distribution of sizes and position of newly induced inversions retained in populations of hybrid origin but which are, with the exception of the induced inversions, structurally homozygous. This will tell us the location of genes which confer heterosis in these populations as well as the size of the gene blocks needed to confer heterosis. Hybrid populations started by crossing strains from a number of localities can be compared. Material accumulated during these studies can be utilized in attacking many additional problems.

The remainder of this report consists of a mathematical model, in which chromosomes are treated as homogeneous strings subject to breakage at any point, which predicts the theoretical distribution of lengths of inverted chromosomal segments as well as their positions. The distributions of naturally occurring inversions can be compared with those predicted by this model; this comparison will entail library research only. The distribution of newly induced inversions must also be compared with theoretical expectations. This comparison is an essential test of the validity of the present model; it is important that the distribution of newly induced inversions be known so that distortions resulting

from selection within populations can be recognized. Although some information is available in the literature for this study, it is quite possible that the available data will have to be augmented by a new cytological study.

For the purpose of making the theory simple, first assume that there is one and only one inversion per chromosome. That is, only the chromosomes with one inversion will be considered. Secondly, assume that a break is equally likely along any part of the chromosome and that the position of the first break is independent of the position of the second break on the chromosome. This means that the first break (x) and the second break (y) each follow the uniform distribution and that the joint distribution of the two breaks is the product of two uniform distributions.

In mathematical terms, for a chromosome of length c,

$$\left.\begin{aligned} f(x) &= \frac{1}{c} \\ &= 0 \end{aligned}\right\} \quad \begin{aligned} &0<x<c \\ &\text{otherwise} \end{aligned}$$

$$\left.\begin{aligned} f(y) &= \frac{1}{c} \\ &= 0 \end{aligned}\right\} \quad \begin{aligned} &0<y<c \\ &\text{otherwise} \end{aligned}$$

(The probability that x (or y) falls in any given interval is $\frac{1}{c}$ times the length of the given interval.)

The joint density function of x and y is:

$$f(x,y) = f(x)f(y) = \frac{1}{c^2}, \quad 0 \leq x,y \leq c$$

and the joint probability function is

$$P\left\{0<x<x_0, \quad 0<y<y_0\right\}$$
$$= \int_0^{y_0}\int_0^{x_0} f(x,y)dxdy = \frac{1}{c^2}\int_0^{y_0}\int_0^{x_0} dxdy$$

The region over which the joint density function is defined may be represented graphically as Figure 1.

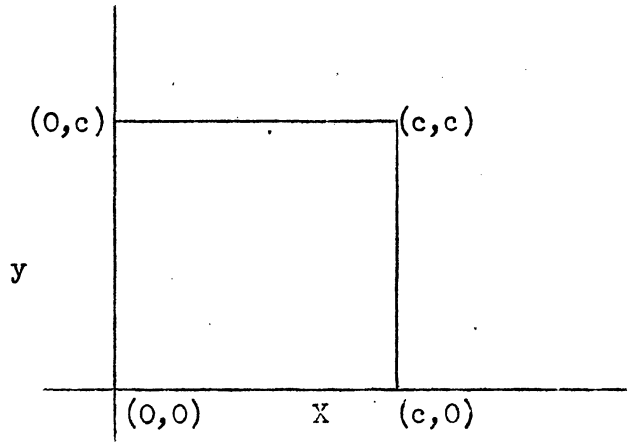

Figure 1. Region for which $f(x,y) \neq 0$.

The problems are to find the distributions of (1) the lengths of the inverted part of the chromosome, (2) the mid-points of the inversions regardless of length, and (3) the mid-points of inversions of fixed length. That is, it is desired to obtain the distributions of $|x-y|$, of $\frac{x+y}{2}$ for all lengths and of $(x+y)/2$ for any fixed value of $|x-y|$.

First, set $z=x-y$ and $w=(x+y)/2$. Then the region of definition of the function $g(w,z)$ is given in Figure 2.

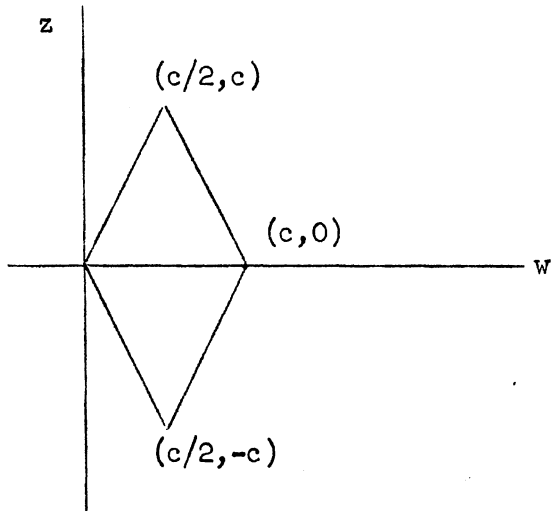

Figure 2. Region for which $g(w,z) \neq 0$.

To find the joint distribution of w and z, we first find

$$J = \pm \begin{vmatrix} \frac{\delta w}{\delta x} & \frac{\delta z}{\delta x} \\ \frac{\delta w}{\delta y} & \frac{\delta z}{\delta y} \end{vmatrix} = - \begin{vmatrix} 1/2 & 1 \\ 1/2 & -1 \end{vmatrix} = 1$$

Now

$$f(x,y) = \frac{1}{c^2} J = \frac{1}{c^2}$$

$$= g(w,z), \quad \begin{cases} -2w < z < 2w < c \\ -2c+2w < z < 2c-2w, \ c/2 < w < c \end{cases}$$

Let $v=|z|=|x-y|$. Then because of the symmetry and uniformity of $g(w,z)$, we have

$$h(w,v) = \frac{2}{c^2}, \quad \begin{cases} 0 < v < 2w < c \\ 0 < v < 2c-2w, \ c/2 < w < c \end{cases}$$

This is the joint distribution of the two variables of interest and is defined over the region shown in Figure 3.

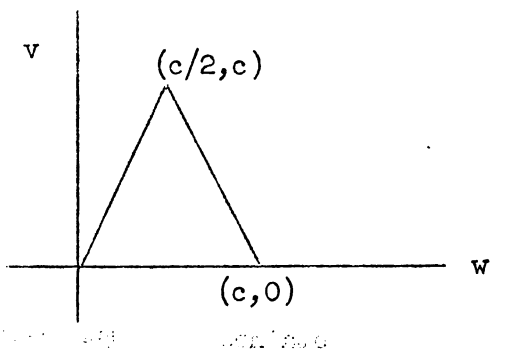

Figure 3. Region for which $h(w,v) \neq 0$

From the distribution of $h(w,v)$, we now find the three distributions of interest, together with the means, variances and other moments of the variables.

First, we find the distribution of lengths of inversions, namely, $h_1(v)$.

$$h_1(v) = \int h(w,v)dw$$

$$= \begin{cases} \frac{2}{c^2} \int_{v/2}^{c/2} dw , & 0<v<c \\ \frac{2}{c^2} \int_{c/2}^{c-v/2} dw , & 0<v<c \end{cases}$$

$$= \frac{2(c-v)}{c^2} , \quad 0<v<c$$

This function is shown graphically in Figure 4.

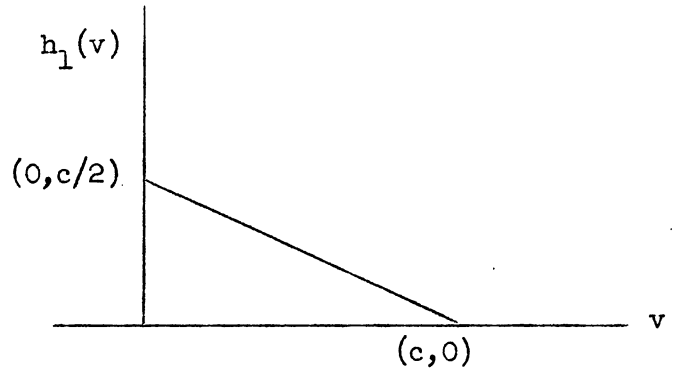

Figure 4. Distribution of $v=|z|=|x-y|$ = length of break.

The mean of v is given by

$$\mu = E(v) = \frac{2}{c^2} \int_0^c (c-v)v \, dv$$

$$= c/3$$

The moments about the mean may be obtained from the following moment generating function.

$$m(t) = E(e^{t(v-c/3)})$$

$$= \frac{2}{c^2}\int_0^c (c-v)e^{t(v-c/3)}dv$$

$$= \frac{2}{c^2}\left[\frac{2}{3}\frac{ce^{t(v-c/3)}}{t}\Big|_0^c - \frac{e^{t(v-c/3)}}{t^2}\left(t\left\{v-c/3\right\}-1\right)\Big|_0^c\right]$$

$$= \frac{2\ e^{\frac{2}{3}ct}}{c^2\ t^2} - \frac{2e^{-\frac{ct}{3}}}{t} - \frac{2e^{-\frac{ct}{3}}}{c^2\ t^2}$$

$$= 1 + 2c\,\frac{(2^3-3(3)+1)}{3^3(3)(2)}\cdot\frac{t}{1} + 2c^2\frac{(2^4+3(4)-1)}{3^4(4)(3)}\cdot\frac{t^2}{2!}$$

$$+ \cdots + 2c^n\,\frac{\{2^{n+2}+(-1)^n 3(n+2)+(-1)^{n+1}\}}{3^{n+2}(n+2)(n+1)}\cdot\frac{t^n}{n!}$$

$$+ \cdots$$

The coefficient of $t^2/2!$ is the variance,

$$\sigma^2 = \frac{2c^2(2^4+3(4)-1)}{3^4(4)(3)}$$

$$= c^2/18$$

Secondly, we required the distributions of midpoints of inversions. This is simply $h_2(w)$.

$$h_2(w) = \int h(w,v)dv$$

$$= \begin{cases} \frac{2}{c^2}\int_0^{2w} dv\ , & 0<w<c/2 \\ \frac{2}{c^2}\int_0^{2c-2w} dv\ , & c/2<w<c \end{cases}$$

$$= \begin{cases} \frac{4w}{c^2}\ , & 0<w<c/2 \\ \frac{4(c-w)}{c^2}\ , & c/2<w<c \end{cases}$$

This function is shown in Figure 5.

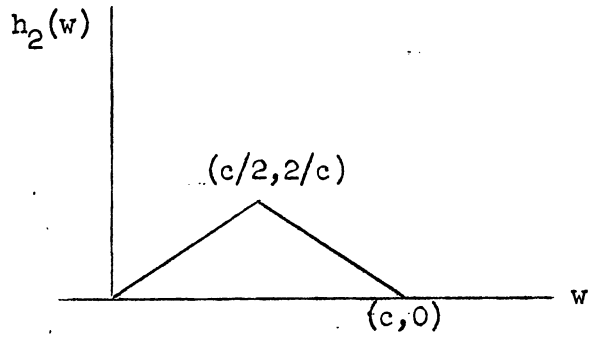

Figure 5. Distribution of w=(x+y)/2 = midpoint.

Clearly this is a symmetric function with mean,

$$\mu = c/2$$

Moments about the mean may be obtained from the following moment generating function.

$$m(t) = E(e^{t(w-c/2)})$$

$$= \int_0^{c/2} e^{t(w-c/2)} \frac{4w}{c^2}\, dw + \int_{c/2}^{c} e^{t(w-c/2)} \frac{4(c-w)}{c^2}\, dw$$

$$= \frac{4}{c^2}\left[\left\{ \frac{e^{t(w-c/2)}}{t^2}\,[t(w-c/2)-1] + \frac{ce^{t(w-c/2)}}{2t} \right\} \Bigg|_0^{c/2} \right.$$

$$\left. + \left\{ \frac{ce^{t(w-c/2)}}{2t} - \frac{e^{t(w-c/2)}}{t^2}\,[\,t(w-c/2)-1\,] \right\} \Bigg|_{c/2}^{c} \right]$$

$$= \frac{8}{c^2t^2}\left[-1 + \cosh\frac{ct}{2}\right]$$

$$= \frac{8}{c^2t^2}\left[-1+\left(1-\frac{i^2c^2t^2}{2^2(2!)}+\frac{i^4c^4t^4}{2^4(4!)}-\frac{i^6c^6t^6}{2^6(6!)}+\cdots\right)\right]$$

$$= 1+\frac{4c^2}{2^3(4)(3)}\frac{t^2}{2!}+\frac{4c^4}{2^5(6)(5)}\left(\frac{t^4}{4!}\right)+\cdots$$

$$+\frac{4c^{2n}}{2^{2n+1}(2n+2)(2n+1)}\frac{t^{2n}}{(2n)!}+\cdots$$

This clearly shows that all odd moments about the mean are zero, as they must be for a symmetric distribution. The variance is the coefficient of $t^2/2!$.

$$\sigma^2 = \frac{4c^2}{2^3(4)(3)}$$

$$= \frac{c^2}{24}$$

Finally, for the distribution of the midpoints of inversions of fixed length, we require the conditional distribution of w given v. This is denoted by h(w|v).

$$h(w|v) = \frac{h(w,v)}{h_1(v)}$$

$$= \left(\frac{2}{c^2}\right)\left(\frac{c^2}{2(c-v)}\right)$$

$$= \frac{1}{c-v} \quad , \qquad \frac{v}{2} < w < c-\frac{v}{2}$$

This distribution is shown graphically in Figure 6.

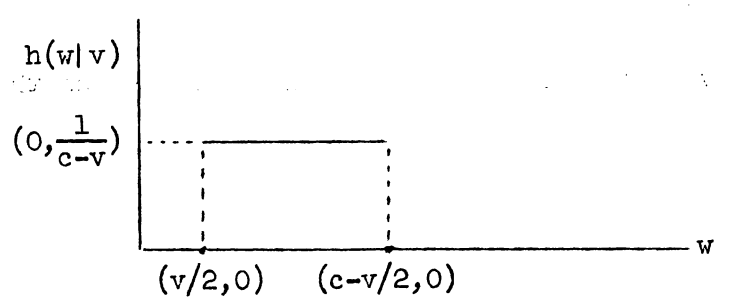

Figure 6. Distribution of w=(x+y)/2 for fixed v=|x-y| .

The mean of the variable in this conditional distribution is

$$E(w|v) = \int h(w|v)dw$$

$$= \int_{v/2}^{c-v/2} \frac{1}{c-v}\, dw$$

$$= c/2$$

For moments about the mean, we compute the appropriate moment generating function. This is

$$E(e^{t(w-c/2)}|v) = \int e^{t(w-c/2)} h(w|v)\, dw$$

$$= \frac{1}{c-v} \int_{v/2}^{c-v/2} e^{t(w-c/2)}\, dw$$

$$= \frac{1}{c-v}\left[\frac{2 \sinh \frac{t(c-v)}{2}}{t}\right]$$

$$= 1 + \frac{(c-v)^2(t^2)}{2^2(3)(2!)} + \frac{(c-v)^4(t^4)}{2^4(5)(4!)} + \cdots$$

$$+ \frac{(c-v)^{2n}\, t^{2n}}{2^{2n}(2n+1)\,(2n)!} + \cdots$$

Again, all odd moments about the mean are zero as they should be for a symmetric, in this case the uniform, distribution. It is also clear that

$$\sigma^2 = \frac{(c-v)^2}{12} \quad .$$

This shows that the variance is zero for v=c and increases as the length of the fixed interval decreases.