A MATHEMATICAL MODEL FOR THE DISTRIBUTION OF LENGTHS OF CHROMOSOMAL DEFICIENCIES INVOLVING A SPECIFIC LOCUS¹

K. H. THOMPSON², B. WALLACE, AND W. T. FEDERER

Department of Plant Breeding, Cornell University, Ithaca, N.Y.

Received December 21, 1964

EXPERIMENTAL distributions of breaks and chromosomal recombinations in *Drosophila melanogaster* have been studied by several authors in attempts to throw light on the mechanism of the origin of chromosomal rearrangement. STADLER (1932), BAUER, DEMEREC, and KAUFMAN (1938), CATCHESIDE (1938), FANO (1943), KAUFMANN (1946) and many others have contributed to this literature. While there is evidence (BAUER *et al.* 1938) that breakage, either spontaneous or X-ray induced, occurs largely at random throughout the length of a chromosome, the accumulated evidence of observed chromosomal alterations indicates that recombination may not be random.

In obtaining alterations to test the hypothesis of random breakage and recombination, inversions and deficiencies are particularly useful because they admit fairly standard cytological or morphological techniques of identification. Inversion loops may be detected and measured with reasonable facility in the salivary chromosomes of *D. melanogaster*, and a series of 49 induced inversions of the X chromosome was obtained by BAUER *et al.* (1938). This distribution was compared by FEDERER, STEEL, and WALLACE (1961) to several mathematical models.

The cytological detection of chromosomal deficiencies, however, is impractical unless the deficiency includes a mutant marker. This technique was used by DEMEREC and FANO (1941) to obtain a series of 37 Notch deficiencies of the X chromosome. These authors observed that the expected frequency of deficiencies involving n bands out of a total number of N bands is equal to $2(N-1-n)/(N-1)(N-2)$ and that the *a priori* probability that any specific band is included in a deficiency of length n is approximately equal to $n/(N-2)$.

These results may be obtained by assuming that any two random and independent breaks may lead with equal probability to a deficiency or to an inversion of the chromosomal segment involved. This assumption probably does not obtain, however, for large chromosomal segments which might be expected to be more viable as inversions than as deficiencies, nor for very short segments which might be lost as deficiencies more often than recombining as inversions. This model also does not take into consideration the shorter deficiencies which may arise as a result of a single break or those deficiencies associated with more complicated

¹ Biometrics Unit Paper No. 104, Paper No. 478 in the Plant Breeding Department. This paper is a modified version of Office of Naval Research Technical Report No. 10, Contract No. Nonr-401(39), Project No. (NR 042-212).

² Present address: Department of Biology, Brookhaven National Laboratories, Upton, Long Island, N.Y.

rearrangements. However, in spite of the fact that we know chromosomal alteration to be a complex mechanism, the two-break model proposed affords a useful hypothesis and perhaps a reasonable approximation.

The calculations of DEMEREC and FANO (1941) are easily extended to obtain the exact distribution of deficiency lengths, given that a specific locus is included. It is convenient to assume that breakage is equally likely between any two bands of a chromosome whose length is divided by N distinct bands. Hence, excluding the two terminal bands, there are $(N-1)(N-2)/2$ different and equally likely deficiencies, of which $(N-1-n)$ are of length n ; $n = 1, 2, \dots, N-2$. It may further be determined that there are $(a-1)(N-a)$ different and equally likely deficiencies which include a specific band a ; $a = 2, 3, \dots, N-1$. Simple counting techniques suffice to show that the frequency of deficiencies of length n which include band a , $a \leq (N+1)/2$, is equal to

$$\begin{aligned} \frac{n}{(a-1)(N-a)} & \quad \text{if} \quad 1 \leq n < (a-1), \\ \frac{1}{(N-a)} & \quad \text{if} \quad (a-1) \leq n \leq (N-a), \text{ and} \\ \frac{N-n-1}{(a-1)(N-a)} & \quad \text{if} \quad (N-a) < n \leq (N-2). \end{aligned}$$

The frequency of deficiencies of length n which include band a is symmetrical in a . For N odd or even, the distribution for a is equal to the distribution for $(N-a+1)$; hence, the frequency of deficiencies of length n which include band a , $a > (N+1)/2$, is equal to

$$\begin{aligned} \frac{n}{(N-a)(a-1)} & \quad \text{if} \quad 1 \leq n < (N-a), \\ \frac{1}{(a-1)} & \quad \text{if} \quad (N-a) \leq n \leq (a-1), \text{ and} \\ \frac{N-n-1}{(N-a)(a-1)} & \quad \text{if} \quad (a-1) < n \leq (N-2). \end{aligned}$$

The calculations of DEMEREC and FANO (1941) may be extended further to include the exact proportion of deficiencies of fixed length n which include a specific band a . This frequency may be seen to depend upon the location of band a . For fixed $n < N/2$, the proportion of deficiencies of length n which include a specific band a is equal to

$$\begin{aligned} \frac{a-1}{N-n-1} & \quad \text{if} \quad 2 \leq a \leq n, \\ \frac{n}{N-n-1} & \quad \text{if} \quad n < a < (N-n+1), \text{ and} \\ \frac{N-a}{N-n-1} & \quad \text{if} \quad (N-n+1) \leq a \leq (N-1). \end{aligned}$$

The number of distinct deficiencies which include band a is symmetrical in n ; that is, the number of deficiencies of fixed length n which include any given band a is equal to the number of deficiencies of fixed length $(N-n-1)$ which include the same band a . Hence, for fixed $n \geq N/2$, the proportion of deficiencies of length n which include a specific band a is equal to

$$\begin{array}{ll} \frac{a-1}{N-n-1} & \text{if } 2 \leq a \leq (N-n-1), \\ 1 & \text{if } (N-n-1) < a < (n+2), \text{ and} \\ \frac{N-a}{N-n-1} & \text{if } (n+2) \leq a \leq (N-1). \end{array}$$

The above frequencies and probabilities are based on a discrete band concept which is commensurate with the techniques of cytological identification and measurement. However, the frequency distribution of deficiency lengths, given that a specific locus a is included, may also be obtained by assuming that breakage may occur at any point of a chromosome of length c . If we assume as before that exactly two independent and random breaks induce a single deficiency, we may let the independent random variables X and Y denote the position of the first and second breaks, respectively. Thus, the joint distribution of X and Y now represents our mathematical model for random two-hit breakage and deficiency and is the bivariate uniform distribution

$$\begin{aligned} f(x,y) &= \frac{1}{c^2} && \text{for } 0 < x < c, \quad 0 < y < c, \\ &= 0 && \text{otherwise.} \end{aligned}$$

By applying the appropriate linear transformations to X and Y , the frequency distributions of deficiency lengths, midpoints, and other characteristics of interest may be obtained. In particular, the distributions of $V=|Z|=|X-Y|$ and $W=(X+Y)/2$ over the region $(0 < x < a < y < c, 0 < y < a < x < c)$ represent the distributions of deficiency lengths and midpoints, given that the deficiency includes a given locus a .

We now proceed to compute the distributions of V and W . Denoting the region of definition by R , by which we mean that (x,y) for fixed $0 < a < c$ satisfies either $0 < x < a < y < c$ or $0 < y < a < x < c$, we begin by observing that

$$\begin{aligned} g(x,y|R) &= \frac{1}{2a(c-a)} && \text{for } (x,y) \in R, \\ &= 0 && \text{otherwise.} \end{aligned}$$

The region R over which this joint density is defined is represented graphically in Figure 1.

It is now convenient to compute the joint distribution of $Z=(X-Y)$ and $W=$

FIGURE 1.—Region for which $g(x,y|R) \neq 0$. FIG. 2.—Region for which $g(w,z|R) \neq 0$.

$(X+Y)/2$ over the same region R . Since the Jacobian of the transformation is equal to 1, $g(w,z|R)$ is uniformly distributed as

$$g(w,z|R) = \frac{1}{2a(c-a)} \begin{array}{ll} \text{for } 0 < 2a < c, & a < 2w < 2a, \quad 2(a-w) < |z| < 2w; \\ \text{for } 0 < 2a < c, & 2a < 2w < c, \quad 2(w-a) < |z| < 2w; \\ \text{for } 0 < 2a < c, & c < 2w < c+a, \quad 2(w-a) < |z| < 2(c-w); \\ \text{for } c < 2a < 2c, & a < 2w < c, \quad 2(a-w) < |z| < 2w; \\ \text{for } c < 2a < 2c, & c < 2w < 2a, \quad 2(a-w) < |z| < 2(c-w); \\ \text{for } c < 2a < 2c, & 2a < 2w < c+a, \quad 2(w-a) < |z| < 2(c-w); \end{array}$$

$$= 0 \quad \text{otherwise.}$$

The region of definition is represented graphically in Figure 2. The joint distribution of $V=|Z|=|X-Y|$ and $W=(X+Y)/2$ is now easily obtained because of the symmetry of $g(w,z|R)$ in z and becomes

$$g(w,v|R) = \frac{1}{a(c-a)} \begin{array}{ll} \text{for } 0 < 2a < c, & a < 2w < 2a, \quad 2(a-w) < v < 2w; \\ \text{for } 0 < 2a < c, & 2a < 2w < c, \quad 2(w-a) < v < 2w; \\ \text{for } 0 < 2a < c, & c < 2w < c+a, \quad 2(w-a) < v < 2(c-w); \\ \text{for } c < 2a < 2c, & a < 2w < c, \quad 2(a-w) < v < 2w; \\ \text{for } c < 2a < 2c, & c < 2w < 2a, \quad 2(a-w) < v < 2(c-w); \\ \text{for } c < 2a < 2c, & 2a < 2w < c+a, \quad 2(w-a) < v < 2(c-w); \end{array}$$

$$= 0 \quad \text{otherwise.}$$

The region of definition may be represented as the shaded area of the first quadrant of Figure 2 where the z axis is replaced by v .

Finally, the distribution of the lengths of deficiencies $V=|X-Y|$ which cover the locus a is computed as $g(v|R)=\int g(w,v|R)dw$. Hence, for fixed $0 < 2a < c$,

$$\begin{aligned} g(v|R) &= \frac{1}{a(c-a)} \int_{(2a-v)/2}^{(2a+v)/2} dw = \frac{v}{a(c-a)} \quad \text{for } 0 < v < a, \\ &= \frac{1}{a(c-a)} \int_{v/2}^{(2a+v)/2} dw = \frac{1}{c-a} \quad \text{for } 0 < v < c-a, \\ &= \frac{1}{a(c-a)} \int_{v/2}^{(2c-v)/2} dw = \frac{c-v}{a(c-a)} \quad \text{for } c-a < v < c, \text{ and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

For fixed $c < 2a < 2c$,

$$\begin{aligned} g(v|R) &= \frac{1}{a(c-a)} \int_{(2a-v)/2}^{(2a+v)/2} dw = \frac{v}{a(c-a)} \quad \text{for } 0 < v < c-a, \\ &= \frac{1}{a(c-a)} \int_{(2a-v)/2}^{(2c-v)/2} dw = \frac{1}{a} \quad \text{for } c-a < v < a, \\ &= \frac{1}{a(c-a)} \int_{v/2}^{(2c-v)/2} dw = \frac{c-v}{a(c-a)} \quad \text{for } a < v < c, \text{ and} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

It is now quite clear that without loss of generality we need consider only $0 < 2a < c$; that is, the end of the chromosome from which we measure the locus a is unimportant. The distribution of V for $0 < 2a < c$ is shown graphically in Figure 3 and shows the symmetry in $a=c-a$. The mean and variance of V are $c/2$ and $[(c-a)^2 + a^2]/12$, respectively.

FIGURE 3.—Distribution of $V=|Z|=|X-Y|$ = deficiency length given that the locus $0 < 2a < c$ is included.

The distribution of the midpoints of deficiencies which cover the locus a may be computed similarly as $g(w|R) = \int g(w,v|R) dv$. For fixed $0 < 2a < c$,

$$\begin{aligned} g(w|R) &= \frac{1}{a(c-a)} \int_{2(a-w)}^{2w} dv = \frac{2(2w-a)}{a(c-a)} \quad \text{for } a < 2w < 2a, \\ &= \frac{1}{a(c-a)} \int_{2(w-a)}^{2w} dv = \frac{2}{c-a} \quad \text{for } 2a < 2w < c, \\ &= \frac{1}{a(c-a)} \int_{2(w-a)}^{2(c-w)} dv = \frac{2(c-2w+a)}{a(c-a)} \quad \text{for } c < 2w < a+c, \text{ and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

For fixed $c < 2a < 2c$,

$$\begin{aligned} g(w|R) &= \frac{1}{a(c-a)} \int_{2(a-w)}^{2w} dv = \frac{2(2w-a)}{a(c-a)} \quad \text{for } a < 2w < c, \\ &= \frac{1}{a(c-a)} \int_{2(a-w)}^{2(c-w)} dv = \frac{2}{a} \quad \text{for } c < 2w < 2a, \\ &= \frac{1}{a(c-a)} \int_{2(w-a)}^{2(c-w)} dv = \frac{2(c-2w+a)}{a(c-a)} \quad \text{for } 2a < 2w < a+c, \text{ and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

FIGURE 4.—Distribution of $W=(X+Y)/2$ = deficiency midpoints given that the locus $0 < 2a < c$ is included.

FIGURE 5.—Distribution of $W=(X+Y)/2$ = deficiency midpoints given that the locus $c < 2a < 2c$ is included.

This distribution is shown graphically in Figures 4 and 5 to show the nature of the symmetry in $a=c-a$. The mean and variance of W are $(c+2a)/4$ and $[(c-a)^2+a^2]/48$, respectively.

The conditional distributions may now be computed as $g(w|v,R)=g(w,v|R)/g(v|R)$ and $g(v|w,R)=g(w,v|R)/g(w|R)$. The conditional distribution of W , the midpoint of the deficiencies, for any fixed value of V , the deficiency length, given that the deficiency includes the locus a , is given as

$$\begin{aligned} g(w|v,R) &= \frac{1}{v} \text{ for } 0 < 2a < c, \quad 0 < v < a, \quad 2a-v < 2w < v+2a; \\ &= \frac{1}{a} \text{ for } 0 < 2a < c, \quad a < v < c-a, \quad v < 2w < v+2a; \\ &= \frac{1}{c-v} \text{ for } 0 < 2a < c, \quad c-a < v < c, \quad v < 2w < 2c-v; \\ &= \frac{1}{v} \text{ for } c < 2a < 2c, \quad 0 < v < c-a, \quad 2a-v < 2w < v+2a; \\ &= \frac{1}{c-a} \text{ for } c < 2a < 2c, \quad c-a < v < a, \quad 2a-v < 2w < 2c-v; \\ &= \frac{1}{c-v} \text{ for } c < 2a < 2c, \quad a < v < c, \quad v < 2w < 2c-v; \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The conditional distribution of V , the length of the deficiency, for any fixed value of W , the deficiency midpoint, given that the deficiency includes the locus a , is given as

$$\begin{aligned} g(v|w,R) &= \frac{1}{2(2w-a)} \text{ for } 0 < 2a < c, \quad a < 2w < 2a, \quad 2(a-w) < v < 2w; \\ &= \frac{1}{2a} \text{ for } 0 < 2a < c, \quad 2a < 2w < c, \quad 2(w-a) < v < 2w; \\ &= \frac{1}{2(c-2w+a)} \text{ for } 0 < 2a < c, \quad c < 2w < c+a, \quad 2(w-a) < v < 2(c-w); \\ &= \frac{1}{2(2w-a)} \text{ for } c < 2a < 2c, \quad a < 2w < c, \quad 2(a-w) < v < 2w; \\ &= \frac{1}{2(c-a)} \text{ for } c < 2a < 2c, \quad c < 2w < 2a, \quad 2(a-w) < v < 2(c-w); \\ &= \frac{1}{2(c-2w+a)} \text{ for } c < 2a < 2c, \quad 2a < 2w < c+a, \quad 2(w-a) < v < 2(c-w); \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The mathematical model discussed to this point has assumed that the loss of a chromosomal segment is based solely upon the location of two chance breaks in the chromosome and not upon the distance between the two breaks. In fact, chromosomal deficiencies of large lengths involving a significant portion of a chromosome are rarely observed. One reason may be that viability is in some

way proportional to the length of the deleted portion of the chromosome. It is also likely that there exist several loci distributed along the length of every chromosome which are essential to viability. Deficiencies including these loci would therefore never be observed. However, without loss of generality, the mathematical model that we have considered is applicable to this situation where the locus a is bounded at an unknown distance on each side by an essential locus; say a_s and a_t where $0 < a_s < a_t < c$. a_s and a_t now become the effective ends of the chromosome insofar as deficiencies which include the locus a are concerned, although their locations must be estimated.

If we let $S=W-(V/2)$ and $T=W+(V/2)$ represent respectively the proximal and distal breaks of a deficiency which includes the locus a ; that is, such that $0 < a_s < s < a < t < a_t < c$, it is easy to show from previous considerations of the model that

$$\begin{aligned} g(s|R) &= \frac{1}{a-a_s} \quad \text{for } 0 < a_s < s < a, \\ &= 0 \quad \text{otherwise, and} \\ g(t|R) &= \frac{1}{a_t-a} \quad \text{for } a < t < a_t < c, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

In fact, S and T are independent, and their distributions are useful in estimating a_s and a_t . If we have n independent deficiencies which include the locus a , then

$$\xi = \left(\frac{n+1}{n} \right) [\max(T_1, \dots, T_n)] - \frac{a}{n}$$

is an unbiased estimator of a_t , and

$$\zeta = \left(\frac{n+1}{n} \right) [\min(S_1, \dots, S_n)] - \frac{a}{n}$$

is an unbiased estimator of a_s . Since $\max(T_1, \dots, T_n)$ and $\min(S_1, \dots, S_n)$ are complete and sufficient for a_t and a_s respectively, ξ and ζ have minimum variance among all unbiased estimators of a_t and a_s ; namely,

$$\begin{aligned} \text{Var}(\xi) &= \frac{(a_t-a)^2}{n(n+2)}, \text{ and} \\ \text{Var}(\zeta) &= \frac{(a-a_s)^2}{n(n+2)}. \end{aligned}$$

In particular, such unbiased estimators of a_t and a_s as $(2\bar{w}-a+\bar{v})$ and $(2\bar{w}-a-\bar{v})$ respectively, where $\bar{w}=\Sigma W_i/n$ and $\bar{v}=\Sigma V_i/n$, are inadmissible under this model.

Experimental data: Experimental data with which the calculations presented here may be tested are scanty indeed. By far the largest number of deficiencies recorded are the Notch-deficiencies summarized by BRIDGES and BREHME (1944). Fifty of these were subjected to analysis; the mid-point of these deficiencies fell

within one band of that (3C7) within which Notch is known to be located. Thus, it appears that no essential regions exist in the immediate neighborhood of Notch. Stated somewhat differently, the lengths of the Notch-deficiencies do not seem to be governed by discrete points, located close by, which act as dominant lethals. The remote possibility that Notch is located by chance exactly mid-way between two such essential regions can be ruled out by determining the mid-point of deficiencies that include Notch and a second locus. The mid-points of those Notch-deficiencies that include white, roughest, split, facet, diminutive, and echinus (taken singly) correspond to the points mid-way between the known locations of Notch and these other mutant loci. Now, if Notch were exactly mid-way between two essential regions, it is impossible that a point mid-way between Notch and echinus, for example, to be exactly mid-way between these two essential regions as well.

More pertinent to these calculations is the case of the white-deficiencies listed in BRIDGES and BREHME (1944). Six white-deficiencies are listed. In none of these does the rightmost break fall more than 2 or 3 bands from the known locus of white. The leftmost breaks, in contrast, in two instances fall a dozen or more bands to the left of white. Unfortunately, these six deficiencies are not sufficient to test the theoretical model. It is quite clear that the "essential region" in the case of the white-deficiencies is the near-by Notch locus; if a deficiency includes this locus, it is called a Notch-deficiency rather than a white-deficiency. By accident of semantics, Notch behaves as a dominant lethal in respect to deficiencies called "white-deficiencies".

SUMMARY

A mathematical model of random two-hit breakage and deficiency is developed for the condition that the chromosomal segment lost includes an arbitrary but specific locus. Both a discrete band model and a continuous chromosome model are considered in presenting theoretical distributions for deficiency lengths, deficiency mid-points, and other related characteristics. The problem of estimating the location of essential loci is discussed for the continuous model.

LITERATURE CITED

- BAUER, H., M. DEMEREC, and B. P. KAUFMANN, 1938 X-ray induced chromosomal alterations in *Drosophila melanogaster*. *Genetics* **23**: 610-630.
- BRIDGES, C. B., and K. S. BREHME, 1944 The mutants of *Drosophila melanogaster*. Carnegie Inst. Wash. Publ. **552**.
- CATCHESIDE, D. G., 1938 The effect of X-ray dosage upon the frequency of induced structural changes in the chromosomes of *Drosophila melanogaster*. *J. Genet.* **36**: 307-320.
- DEMEREC, M., and U. FANO, 1941 Mechanism of the origin of X-ray induced notch deficiencies in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.* **27**: 24-31.
- FANO, U., 1953 Mechanism of induction of gross chromosomal rearrangements in *Drosophila* sperms. *Proc. Natl. Acad. Sci. U.S.* **29**: 12-18.

- FEDERER, W. T., R. G. D. STEEL, and B. WALLACE, 1961 A mathematical model for lengths and mid-points of inversions in chromosomes. ONR Technical Report No. 3, Contract No. Nonr-401(39), Project No. (NR 042-212), Cornell University.
- KAUFMANN, B. P., 1946 Organization of the chromosome. I. Break distribution and chromosome recombination in *Drosophila melanogaster*. J. Exptl. Zool. **102**: 293-320.
- STADLER, L. J., 1932 On the genetic nature of induced mutations in plants. Proc. 6th Intern. Congr. Genet. **1**: 274-294.