MATHEMATICAL MODELS FOR CHROMOSOMAL INVERSIONS[1]

W. T. FEDERER, R. G. D. STEEL[2], AND BRUCE WALLACE

Department of Plant Breeding and Biometry, Cornell University, Ithaca, N.Y. 14850

Received January 9, 1967

SINCE a formal characterization of the nature of inversions has not been made to date, this paper will hypothesize mathematical models for the occurrence of inversions of various lengths, obtain the associated probability distributions of lengths and midpoints of chromosomal inversions, and compare the hypothesized models with two sets of data. A study of the departures from hypothesized models should throw light on the mechanisms of the origin of inversions with respect to their distribution along the chromosome.

The Data

BAUER, DEMEREC, and KAUFMANN [1938] obtained 49 inversions on a Drosophila chromosome by treatment with X rays. Utilizing the 20 approximately equal segments on a cytological map, they classified these inversions according to their length. For example, the class designated as 0–1 (their class 0) contained inversions whose endpoints fell within a segment, class 1–2 (their class 1) contained inversions whose endpoints fell in two adjacent segments, class 2–3 (their class 2) contained inversions whose endpoints fell in two segments separated by one segment, . . . , class 19–20 contained inversions whose endpoints were separated by 18 segments. The frequency distribution of lengths of inversions obtained is presented in the second column of Table 1.

DR. B. P. KAUFMANN (private communication) obtained another series of 98 inversions in the X chromosome of *D. melanogaster*. Again, the chromosome was divided into 114 units; the following numbers of inversions of specified lengths were observed:

[1] Paper No. BU-102 in the Biometrics Unit and No. 473 in the Department of Plant Breeding, Cornell University. Research supported in part under Office of Naval Research Contract NOnr-401(39), Project NR 042-212. Reproduction in whole or in part is permitted for any use by the United States Government.

[2] Now at North Carolina State University, Raleigh, North Carolina 27607.

Length.	Freq.	Length.	Freq.	Length.	Freq.	Length.	Freq.	Length	Freq.	Length	Freq.
1	0	21	1	41	1	61	0	81	1	101	0
2	0	22	3	42	3	62	0	82	0	102	0
3	0	23	2	43	3	63	0	83	1	103	0
4	0	24	0	44	2	64	0	84	2	104	0
5	1	25	2	45	3	65	0	85	0	105	0
6	1	26	4	46	0	66	1	86	0	106	0
7	0	27	2	47	0	67	2	87	0	107	0
8	1	28	0	48	1	68	0	88	0	108	0
9	0	29	0	49	4	69	0	89	1	109	0
10	2	30	1	50	1	70	2	90	0	110	0
11	2	31	2	51	2	71	0	91	0	111	1
12	4	32	1	52	1	72	0	92	0	112	0
13	3	33	2	53	1	73	1	93	1	113	0
14	3	34	3	54	1	74	0	94	0	114	0
15	2	35	1	55	0	75	0	95	0	...	.
16	2	36	0	56	0	76	0	96	0	...	.
17	2	37	1	57	1	77	0	97	0	...	.
18	2	38	1	58	1	78	0	98	0	...	.
19	1	39	0	59	1	79	0	99	1	...	.
20	3	40	1	60	1	80	1	100	0	...	.

TABLE 1

Observed and computed class and cumulative frequencies under Models I to V for the data on 49 inversions

	Class frequency						Cumulative frequency*						
		Computed from model					Observed		Computed from model				
Class (units)	Observed	I	II	III	IV	V	All	Omit 0-1	I	II	III	IV	V
0–1	3	4.8	. .	5.7	. .	6.0	3	.	4.8	. .	5.7	. .	6.0
1–2	5	4.5	4.7	5.3	5.4	5.5	8	5	9.9	4.7	11.0	5.4	11.5
2–3	6	4.3	4.5	4.9	5.1	5.0	14	11	13.6	9.2	15.9	10.5	16.5
3–4	6	4.0	4.2	4.5	4.7	4.6	20	17	17.6	13.4	20.5	15.2	21.0
4–5	4	3.8	4.0	4.1	4.3	4.1	24	21	21.4	17.3	24.6	19.5	25.2
5–6	7	3.6	3.7	3.8	3.9	3.7	31	28	25.0	21.0	28.4	23.4	28.9
6–7	4	3.3	3.4	3.4	3.6	3.3	35	32	28.3	24.5	31.8	27.0	32.2
7–8	1	3.1	3.2	3.0	3.2	3.0	36	33	31.4	27.7	34.8	30.2	35.2
8–9	4	2.8	2.9	2.7	2.9	2.6	40	37	34.2	30.6	37.5	33.1	37.8
9–10	4	2.6	2.7	2.4	2.5	2.3	44	41	36.8	33.3	39.9	35.6	40.1
10–11	0	2.3	2.4	2.0	2.2	2.0	44	41	39.1	35.7	41.9	37.9	42.0
11–12	0	2.1	2.2	1.7	1.9	1.7	44	41	41.2	37.9	43.7	39.8	43.7
12–13	1	1.8	1.9	1.4	1.6	1.4	45	42	43.0	39.8	45.1	41.4	45.1
13–14	0	1.6	1.7	1.2	1.3	1.2	45	42	44.6	41.4	46.3	42.7	46.3
14–15	2	1.3	1.4	0.9	1.1	0.9	47	44	45.9	42.8	47.2	43.8	47.2
15–16	0	1.1	1.1	0.7	0.8	0.7	47	44	47.0	44.0	47.9	44.7	47.9
16–17	1	0.9	0.9	0.5	0.6	0.5	48	45	47.9	44.9	48.4	45.3	48.4
17–18	0	0.6	0.6	0.3	0.4	0.3	48	45	48.5	45.5	48.7	45.7	48.7
18–19	1	0.4	0.4	0.2	0.2	0.2	49	46	48.9	45.9	48.9	45.9	48.9
19–20	0	0.1	0.1	0.1	0.1	0.1	49	46	49.0	46.0	49.0	46.0	49.0

* Additional decimal places were used in computing these values.

TABLE 2

Observed and computed class and cumulative frequencies under Models I to V for the data on 98 inversions

	19 classes								First class omitted from 16 classes					
	Class frequency				Cumulative frequency*				Class frequency			Cumulative frequency*		
		Computed from model				Computed from model				Computed from model			Computed from model	
Class (units)	Observed	I	III	V	Observed	I	III	V	Observed	II	IV	Observed	II	IV
0–1	2	10.0	10.7	10.4	2	10.0	10.7	10.4	...	...	...	...	...	...
1–2	9	9.5	10.0	9.8	11	19.5	20.7	20.2	18	12.3	13.2	18	12.3	13.2
2–3	14	9.0	9.4	9.2	25	28.5	30.1	29.4	14	11.4	12.1	32	23.7	25.3
3–4	10	8.4	8.7	8.6	35	36.9	38.8	37.9	9	10.6	11.1	41	34.3	36.4
4–5	9	7.9	8.1	8.0	44	44.8	46.9	45.9	10	9.7	10.0	51	44.1	46.5
5–6	9	7.3	7.5	7.4	53	52.1	54.4	53.3	11	8.9	9.0	62	53.0	55.5
6–7	7	6.8	6.8	6.8	60	58.9	61.2	60.0	11	8.0	8.0	73	61.0	63.5
7–8	9	6.2	6.2	6.2	69	65.2	67.4	66.2	5	7.2	7.0	78	68.2	70.5
8–9	10	5.7	5.6	5.6	79	70.9	73.0	71.9	2	6.3	6.1	80	74.5	76.6
9–10	4	5.2	5.0	5.1	83	76.0	78.0	76.9	5	5.5	5.1	85	80.0	81.7
10–11	1	4.6	4.4	4.5	84	80.6	82.4	81.4	1	4.6	4.2	86	84.6	86.0
11–12	4	4.1	3.8	3.9	88	84.7	86.2	85.4	5	3.8	3.4	91	88.4	89.3
12–13	1	3.5	3.2	3.4	89	88.2	89.4	88.8	2	2.9	2.5	93	91.3	91.9
13–14	5	3.0	2.7	2.9	94	91.2	92.1	91.6	1	2.1	1.8	94	93.4	93.6
14–15	1	2.4	2.2	2.3	95	93.7	94.3	93.9	0	1.2	1.0	94	94.6	94.7
15–16	1	1.9	1.7	1.8	96	95.6	95.9	95.7	1	0.4	0.3	95	95.0	95.0
16–17	1	1.4	1.2	1.3	97	96.9	97.1	97.0	.	..	..	..	...	...
17–18	0	0.8	0.7	0.8	97	97.7	97.8	97.7	.	..	..	..	...	...
18–19	1	0.3	0.2	0.3	98	98.0	98.0	98.0	.	..	..	..	...	...

* Additional decimal places were used in computing these values.

If the above data are divided into 19 equal classes the resulting frequency distribution is the one in the second column of Table 2.

Models for Distribution of Inversions (Continuous Data)

Model I: We shall restrict ourselves to a treatment of two-break inversions and shall assume that a break (x) is equally likely along any part of the chromosome, and that the position of a second break (y) is independent of the first break (x) and is also equally likely along any part of the chromosome. This means that x and y follow the uniform distribution and that the joint distribution of the two breaks, x and y, is the product of two independent uniform distributions.

For a chromosome of length c we may write the individual density functions and the joint density function as follows:

$$f(x) = \begin{cases} 1/c \text{ for } 0 < x < c \\ 0 \text{ otherwise} \end{cases}$$

$$f(y) = \begin{cases} 1/c \text{ for } 0 < y < c \\ 0 \text{ otherwise} \end{cases}$$

$$f(x,y) = \begin{cases} 1/c^2 \text{ for } 0 < x < c \text{ and } 0 < y < c \\ 0 \text{ otherwise} \end{cases}$$

The joint probability function is

$$P\{0 < x < x_0, 0 < y < y_0\} = (1/c^2) \int_0^{y_0} \int_0^{x_0} dxdy \ .$$

Let $z = x-y$, $w = (x+y)/2$, and then let $v = |x-y|$. The joint distribution of the variables w = midpoints of inversions and v = length of inversion becomes:

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

Integrating out the variable w, the distribution of lengths of inversions is equal to:

$$h_1(v) \ = \begin{cases} 2(c-v)/c^2 \text{ for } 0 < v < c \\ 0 \text{ otherwise} \end{cases}$$

This function is illustrated graphically in Figure 1.

The mean $= c/3$ and the variance $= c^2/18$ for this distribution. The moment generating function is equal to:

$$\frac{2e^{-ct/3}}{c^2t^2} \{e^{ct} - ct - 1\}$$

In order to compare the observed frequencies for the two sets of data with the theoretical frequencies obtained for the density function $h_1(v)$ we require the areas under the curve for class intervals of one unit in lengths $c/20$ and $c/19$ for the two examples. Thus, cumulative areas are given by:

$$(2/c^2) \int_0^{ic/20} (c-v)dv = 1 - \left(\frac{20-i}{20}\right)^2 ; i = 1, 2, \ldots, 20 \ .$$

The expected number in the first class would be $N = 49$, the sample size, times .0975, the value of the integral for $i = 1$. Likewise, the expected proportion of inversions which are longer than one unit and shorter than two units is:

$$[1 - (20-2/20)^2] - [1 - (20-1/20)^2] = .1900 - .0975 = .0925 \ .$$

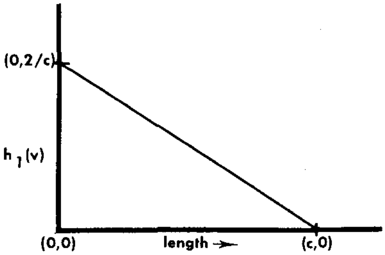

FIGURE 1.—Distribution of length of inversion.

The expected proportions in the remaining classes are obtained similarly. Also, the area between zero and $6i$ units for $c = 114$ is:

$$(2/c^2) \int_0^{ic/19} (c-v)dv = 1 - [(19-i)/19]^2, i = 1, 2, \cdots, 19$$

In some cases it may be desirable to divide the area for $h_1(v)$ into designated fractions, say of equal area. This may be performed by solving the following integral equations for v_0, given the value of α:

$$(2/c^2) \int_0^{v_0} (c-v)dv = \alpha .$$

Upon integrating we find that

$$v^2{}_0 - 2cv_0 + c^2\alpha = 0 .$$

The two roots of the equation are

$$v_0 = c(1 \pm (1-\alpha)^{1/2}).$$

Since $0 < v_0 < c$, the usable root for v_0 is $v_0/c = 1 - (1-\alpha)^{1/2}$. Setting $\alpha = .10$, .20, .30, etc. we obtain the required division points on the v axis as:

$\alpha\%$	10	20	30	40	50	60	70	80	90	100
v/c	0.051	0.106	0.163	0.225	0.293	0.368	0.452	0.553	0.684	1.000

In Tables 1 and 2 the observed and computed frequencies per class and the observed and computed cumulative frequencies under Model I are presented for the two sets of data. In both sets it may be noted that there is a deficiency of inversions observed in the first and perhaps in the second class. A paucity of very short inversions may mean that they arise less frequently than the model predicts (perhaps the chromosome is unable to bend back on itself sharply enough to bring the breakage points near so that an inversion can occur or potential inversions are lost to deficiencies if breaks are close together). The possibility that short inversions are not detected seems unlikely for several reasons.

Model II: For Model II we shall assume the same situation as in Model I except that there is a deficiency of inversions in the first class. Essentially this means that a truncated Model I will be fitted to the remaining classes. Since there are 46 inversions in the remaining 19 classes for the set of 49 inversions the factor 46/(area for 19 classes) = 46/(1 − .0975) = 51 = N^* times the area in each class gives the expected frequencies. The sum of the computed frequencies for the last 19 classes equals 46. Extrapolation to the 0–1 class gives the expected value of 51 − 46 = 5 instead of 4.8 as under Model I.

Similarly, if we assume that there is a deficiency of all inversions less than 10 units in length (i.e. the class midpoint is 10 units and the class intervals are 9.5 to 10.5 units), the remaining area is $1 - (2/114^2) \int_0^{9.5} (114-v)dv = 1 - (2/114^2)\ (9.5(114) - 9.5^2/2) = 1 - 2075.75/12996$. Therefore, $N^* = [(98-3)(12996)]/10920.25 = 113$ = effective sample size for the 98 inversion set of data. The expected number of inversions in the first 9.5 units would be $113 - 95 = 18$ whereas only three were observed. A single degree of freedom chi-square contrast of observed with expected in the first class is significant at the 1% level.

Using the computed N^* values, the computed class frequencies and cumulative frequencies may be compared with the observed class frequencies and cumulative frequencies in Tables 1 and 2 under Model II. Since the 98 inversions could be divided into 15 equal classes from 10 to 114, these were the classes used. Whether or not the number of runs of sign change for observed minus expected is significant at the .05 level (see EISENHART and SWED [1943]) is a function of the grouping of observations in the right tail of the distribution. However, for both sets of data it may be noted that there is an excess of observed over computed class frequencies in the shorter classes. Hence it must be concluded that Model II does not adequately fit the data. The assumption of independence of the points of breakage must be violated, at least operationally.

Model III: Utilizing a form for the joint distribution of two variables presented by GUMBEL [1958] and used by PARZEN [1960, page 292] we may write $f(x,y) = f(x)f(y)[1 + \rho(2F(x)-1)(2F(y)-1)]$ where $|\rho| \leq 1$ and where $F(x) = x/c$ and $F(y) = y/c$ are cumulative distributions of x and y, respectively, for x and y uniformly distributed. Therefore, $f(x,y) = (1/c^2)\ [1 + \rho(2x-c)(2y-c)/c^2]$. Again if we let $w = (x+y)/2$ = midpoint and $v = |z| = |x-y|$ = length of inversion, the joint distribution of w and v becomes:

$$h(w,v) = \begin{cases} (2/c^2)\ \{1 + (\rho/c^2)(4w^2 - v^2 - 4cw + c^2)\} & \\ \qquad\qquad \text{for} \begin{cases} 0 < \mathrm{v} < 2w < c \\ 0 < v < 2c - 2w,\ c/2 < w < c \end{cases} \\ 0 \text{ otherwise} \end{cases}$$

Integrating out w the distribution for length of inversions becomes

$$h_3(v) = \begin{cases} (2/c^2)\ \{c - v + (\rho/3c^2)(c^3 - 3c^2 + 2v^3)\} & \text{for } 0 < v < c \\ 0 \text{ otherwise} \end{cases}$$

When ρ = zero, $h_3(v) = h_1(v)$.

In order to observe the type of dependence of y on x, say, we note that the first moment of the conditional distribution gives the regression function, which is:

$$E(y|x) = (1/c)\int_0^c y[1 + \rho(4xy - 2c(x+y) + c^2)/c^2]dy$$

$$= \frac{c}{2} + \frac{\rho}{3}\left(x - \frac{c}{2}\right).$$

This is the linear regression equation since the mean for the uniform distribution is equal to $c/2$. Since $|\rho| \leq 1$ the regression coefficient varies between $\pm 1/3$. A positive value for ρ would mean that the breakage points are closer together than would be expected on the basis of independence and a negative value for ρ would mean that the points are further apart than expected. If radiation damage were to spread for a short distance along the chromosome on either side of the actual "hit", one can imagine that a break would occur anywhere within the damaged segment. For $\rho > 0$ the breaks would tend to be on the innermost ends of the damaged segments; for $\rho < 0$ the breaks would tend to be on the far ends.

To obtain a specific distribution $h_3(v)$ a value for ρ needs to be obtained. Following the procedure used by FEDERER, STEEL, and WALLACE (1961) we note that the numbers of inversions, say m_i, whose lengths fall in the various class intervals, have a multinomial distribution as follows:

$$L = N! \prod_{i=1}^{k} p_i^{m_i}/m_i!$$

where $k =$ number of classes, where $\sum_{k=1}^{k} m_i = N$, and where

$$p_i = \int_{(i-1)c/k}^{ic/k} h_3(v)dv = \frac{2}{k}\left\{1 - \frac{(2i-1)}{2k} + \rho \left[\frac{1}{3} - \frac{2i-1)}{2k} + \frac{4i^3-6i^2+4i-1}{6k^3}\right]\right\}$$
$$= \{t_i + \rho u_i\} / 3k^4$$

where $t_i = 6k^3 - 6ik^2 + 3k^2$ and $u_i = 2k^3 + 3k^2 - 1 - i(6k^2-4) - 6i^2 + 4i^3$.

The maximum likelihood estimate is that value of ρ, say $\hat{\rho}$, which is the solution of the following equation:

$$\sum_{i=1}^{k} \frac{m_i u_i}{t_i+\hat{\rho}u_i} = 0 .$$

Then, the variance of $\hat{\rho}$ from a sample of size N is obtained as:

$$V(\hat{\rho}) = 1/N \sum_{i=1}^{k} p_i u_i^2/(t_i+\rho u_i)^2 = 3k^4/N \sum_{i=1}^{k} u_i^2/(t_i+\rho u_i) .$$

The estimated variance is obtained by substituting $\hat{\rho}$ for ρ in the above expression. For the two sets of data the maximum likelihood estimates of ρ as found from the equation above are $\hat{\rho}_{49,\text{III}} = 0.61614$ and $\hat{\rho}_{98,\text{III}} = 0.20128$.

With the above values of $\hat{\rho}$ substituted in the equation for the p_i, the computed values for the class frequencies are calculated as Np_i. The class frequencies along with the cumulative frequencies for the two sets of data are presented in Table 1 for the 49 inversions and in Table 2 for the 98 inversions.

Model IV: As noted under Model II, there appears to be a deficiency of inversions in the first class. Denoting the total number of inversions as N, the observed number of inversions in the first class as m_1, and the proportion of the area in the first class as p_1, the multinomial distribution after deleting the first class becomes:

$$(N-m_1)! \prod_{i=2}^{k} \left(\frac{p_i}{1-p_1}\right)^{m_i} \Big/ \; m_i! = L \; ;$$
$$\frac{\partial \log L}{\partial \hat{\rho}} = \sum_{i=2}^{\text{k}} \frac{m_i u_i}{t_i+\hat{\rho}u_i} + \frac{(N-m_1)u_1}{3k^4-t_1-\hat{\rho}u_1} = 0 .$$

The maximum likelihood estimator $\hat{\rho}$ is that value of ρ satisfying the above equation. Now, the estimated variance of $\hat{\rho}$ is

$$V(\hat{\rho}) = 3k^4/N \left(\sum_{i=2}^{k^2} u_i^2/(t_i+\hat{\rho}u_i) - u_1^2/(3k^4-t_1-\hat{\rho}u_1)\right) .$$

For these data $\hat{\rho}_{46,\text{IV}} = 0.47898$ and $\hat{\rho}_{95,\text{IV}} = 0.24907$. With the value $\hat{\rho}$ substituted for ρ in the p_i we compute the quantities $(N-m_1)p_i/(1-p_1)$ for $i = 2,3,\ldots,$

k to obtain the calculated values for each of the classes excluding the zero class. The class and cumulative frequencies in Tables 1 and 2 for Model IV were computed using this procedure. If desired, the effective sample size may be calculated as $N^*_{46} = N/(1-p_1) = 46/.88774 = 51.8$ and $N^*_{95} = 95/.87950 = 109.3$ for the 49 and 98 inversion sets of data, respectively, when the 0–1 class data are not utilized.

Model V: Another family of functions for the distribution of lengths of inversions (see FEDERER, STEEL, and WALLACE [1961]) is:

$$h_5(v) = \begin{cases} 1/c + \sum_{i=1}^{b} \beta_i(v^i - c^i/(i+1)), & 0 < v < c \\ 0 \text{ otherwise} \end{cases}$$

subject to the constraint that $h_5(v=c) = 1/c + \sum_{i=1}^{b} \beta_i c^i i/(i+1) = 0$. For $b = 2$, $h_5(v)$ becomes

$$h_5(v) = \frac{1}{c} + \beta_1\left(v - \frac{c}{2}\right) + \beta_2\left(v^2 - \frac{c^2}{3}\right)$$

$$= \frac{2(c-v)}{c^2}\left\{1 - (c-3v)\left[\frac{1}{4c} + \frac{c\beta_1}{8}\right]\right\}$$

since

$$\beta_2 = -\frac{3}{2c^3} - \frac{3\beta_1}{4c} .$$

When $\beta = \beta_1 = -2/c^2$, and therefore $\beta_2 = 0$, $h_5(v)$ becomes $h_1(v)$. There appears to be no direct relation between $h_5(v)$ for $b = 2$ and $h_3(v)$ for Model III except when $\rho = 0$ in $h_3(v)$ and when $\beta = -2/c^2$ in $h_5(v)$. Both h_3(v) and $h_5(v)$ may be written in terms of $h_1(v)$ as follows:

$$h_3(v) = h_1(v)\left\{1 + \frac{\rho}{3c^2}(c^2 - 2cv - 2v^2)\right\}$$

and

$$h_5(v) = h_1(v)\left\{1 + \left(\frac{2+c^2\beta}{8c}\right)(3v-c)\right\} .$$

For $b = 2$, the mean and variance associated with $h_5(v)$ are equal to

$$\frac{c}{12}\,[6 + \beta_1 c^2 + \beta_2 c^3] \text{ and } \frac{c^2}{12}\left\{1 + \frac{\beta_2 c^3}{15} + \frac{(\beta_1 c^2 + \beta_2 c^3)^2}{12}\right\}$$

respectively. The distribution function for $h_5(v)$ is shown in Figure 2 for β negative.

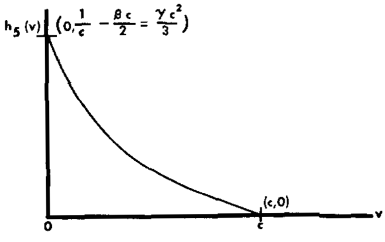

FIGURE 2.—Distribution of length of inversion.

The distribution function $h_5(v)$ may not appeal to the geneticist because the basis for chromosome breakage is not postulated. This procedure falls in the area of curve-fitting at present. In order for this model to be palatable to the biologist it would be necessary to determine the type of relationship involved between x and y. The model is presented as a statistical alternative to those in the previous sections and because it is equal to $h_1(v)$ when $\beta = -2/c^2$.

Dividing the length c into k equal parts the area under the curve between the $(i-1)$st unit and the ith unit, $(i = 1, 2, \ldots, k)$ is

$$p_i = \int_{(i-1)c/k}^{ic/k} h_5(v)dv = \{f_i + c^2\beta g_i\}/4k^3$$

where $f_i = 6k^2 - 2 + 6i - 6i^2$ and $g_i = 3i - 3i^2 + k(4i-2) - k^2 - 1$.

To obtain the maximum likelihood estimator of β we note that the m_i have a multinomial distribution as follows:

$$L = N! \prod_{i=1}^{k} p_i^{m_i}/m_i! \ .$$

The value of β, for c known (usually $c=1$), which maximizes $\log L$ is obtained as a solution to the following equation:

$$\sum_{i=1}^{k} m_i g_i/(f_i + c^2\hat{\beta} g_i) = 0 \ .$$

The variance of $\hat{\beta}$ is obtained as:

$$V(\hat{\beta}) = 16k^6/Nc^4 \sum_{i=1}^{k} g_i^2/p_i \ .$$

The $V(\hat{\beta})$ is approximated by substituing $\hat{\beta}$ for β in p_i and setting $c=1$; denote p_i as $\hat{p}_i$ in this case.

From the data for 49 inversions $\hat{\beta}_{49} = -4.1663$ for $c=1$ and $\hat{\beta}_{98} = -2.3135$. Substituting the $\hat{\beta}$ values in the formula for the p_i, the various estimated $\hat{p}_i$ are obtained. Then, $N\hat{p}_i$ is the estimated or computed value for the ith class. These values are given in Tables 1 and 2 for the two sets of data.

Model VI: We may delete the first class, or any other class, from the data and proceed as described under Model IV. This would result in obtaining an estimate of the parameter β from all classes except the deleted one or ones. The computations were not made for this case because of the lack of a biological formulation leading to $h_5(v)$.

Models for Distribution of Inversions (Grouped Data)

Model VII: Suppose that the chromosome is divided into k equal (These need not be equal, but it simplifies the argument to make them equal.) segments and the recorded length is as described under *The Data*. Observe that the 0-1 class of inversions contains inversions occurring within the premarked segment and which vary in length from zero to one class interval, that the 1-2 class of inversions contains the number of inversions which had endpoints in two adjacent segments or classes and which vary in length from zero to two segments, that those in the 2-3 class vary from one to three segments in length, etc. Under this scheme and for a bivariate uniform distribution as under Model I we note that the two end-points (x,y) of an inversion may be characterized as:

$$P\{x,y \text{ fall in the same segment}\} = 1/k$$
$$P\{x,y \text{ fall in adjacent segments}\} = 2(k-1)/k^2$$
$$P\{x,y \text{ fall in segments separated by one segment}\} = 2(k-2)/k^2$$
$$\vdots$$
$$P\{x,y \text{ fall in segments separated by } k-2 \text{ segments}\} = 2/k^2$$

Since the set of 49 inversions would be most affected by the method (for the 98 inversions six classes were grouped to form one class in Table 2.) of measuring lengths of inversions, the following probabilities (P) and expected values (E = 49P) have been computed for these data (O):

Class	P	E	O	Class	P	E	O	Class	P	E	O
0–1	.050	2.45	3	7–8	.065	3.18	1	14–15	.030	1.47	2
1–2	.095	4.66	5	8–9	.060	2.94	4	15–16	.025	1.22	0
2–3	.090	4.41	6	9–10	.055	2.70	4	16–17	.020	0.98	1
3–4	.085	4.16	6	10–11	.050	2.45	0	17–18	.015	0.74	0
4–5	.080	3.92	4	11–12	.045	2.20	0	18–19	.010	0.49	1
5–6	.075	3.68	7	12–13	.040	1.96	1	19–20	.005	0.24	0
6–7	.070	3.43	4	13–14	.035	1.72	0				

Model VIII: If we truncate the frequency distribution eliminating the first class to obtain the counterpart for Model II, this involves no difficulties as we simply divide each of the last k-1 probabilities above by $1 - P\{x,y$ fall in the same segment$\} = (k-1)/k$. Then for the 49 inversions, $N^* = 46/19/20 = 51.1$ and the various P, E, and O values are:

Class	P	E	O	Class	P	E	O	Class	P	E	O
0–1	. . .	. . .	3	7–8	.068	3.15	1	14–15	.032	1.45	2
1–2	.100	4.60	5	8–9	.063	2.90	4	15–16	.026	1.21	0
2–3	.095	4.36	6	9–10	.058	2.66	4	16–17	.021	0.97	1
3–4	.090	4.12	6	10–11	.053	2.42	0	17–18	.016	0.73	0
4–5	.084	3.87	4	11–12	.047	2.18	0	18–19	.010	0.48	1
5–6	.079	3.63	7	12–13	.042	1.94	1	19–20	.005	0.24	0
6–7	.074	3.39	4	13–14	.037	1.70	0				

It should be noted that the first seven expected values (E) for all the data above are lower than the observed values (O) and even for the truncated set, the first six values of expected (E) are less than the corresponding observed (O) values. Hence, the method of this section gave a run of plus signs for O—E in much the same manner as for Model II. However, it should be noted that relatively low values of $x^2 = \sum_{i=1}^{k} (0-E)^2/E$ would be obtained for all methods. The method of this section fits the data for the 49 inversions better in the first class than do Models I, III, and V. For the 98 inversions, the methods of this section give about the same results as obtained for Models I to V. Hence the method of measurement had little effect on the fit for the five models for the data involving 98 inversions.

A study of the computed values under Models I to V in comparison with the observed values in Tables 1 and 2 reveals that Models III and V fit the data somewhat better than Model I. The deficiency of small inversions as observed in the first class is evidence that a truncated procedure should be utilized. Using truncated data and procedures after omitting the data in the first class indicates that Model IV yields a better fit than Model II. For the two sets of data it would appear that independence of points of breakage as indicated by the length and location of the inversion may be suspect, and that inversions may not be recovered at random with respect to such breaks. In order to differentiate between the truncated models much larger sets of data would be required.

Distribution of Midpoints and Proximal and Distal Points of Inversions Under Models I and III

The distribution of midpoints of inversions under Model I is given by FEDERER, STEEL and WALLACE [1961] as:

$$h_1(w) = \begin{cases} 4w/c^2, \ 0 < w < c/2 \\ 4(c-w)/c^2, c/2 < w < c \\ 0, \text{ otherwise} \end{cases}$$

The mean is $c/2$, the variance is $c^2/24$, and the generating function for moments about the mean is $m(t) = E(e^{t(w-c/2)}) = (8/c^2t^2)(-1 + \cosh ct/2)$. The distribution of w = midpoint of inversions is given in Figure 3.

Setting $s = w - v/2$ = distance to proximal point and setting $r = w + v/2$ = distance to distal point of inversion, the respective distributions under Model I for r and s were found to be:

$$k_1(s) = \begin{cases} (2/c^2)(c-s), \ 0 < s < c \\ 0, \text{ otherwise} \end{cases}$$

$$k_1(r) = \begin{cases} (2/c^2)(c-r), \ 0 < r < c \\ 0, \text{ otherwise} \end{cases}$$

$$k_1(r,s) = \begin{cases} (2/c^2, \ 0 < r, s < c \\ 0, \text{ otherwise} \end{cases}$$

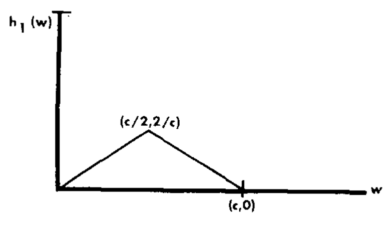

FIGURE 3.—Distribution of $w = (x+y)/2$ = midpoint.

Also, the distribution of midpoints of inversions w of fixed length v is:

$$h_1(w|v) = \begin{cases} 1/(c-v), & v/2 < w < c - v/2 \\ 0, & \text{otherwise} \end{cases}$$

Under Model III the distribution of midpoints of inversions is given by:

$$h_3(w) = \begin{cases} (4w/c^2)[1+(\rho/c^2)((2w-c)^2 - 4w^2/3)], & 0 < w < c/2 \\ (4(c-w)/c^2)[1+(\rho/c^2)((2w-c)^2 - 4(c-w)^2/3)], & c/2 < w < c \\ 0, & \text{otherwise} \end{cases}$$

Proceeding as for Model I, the additional distributions for midpoints for a fixed length, for distal points and for proximal points are readily obtainable. These distribution functions under Model V are not obtainable because $h_5(w,v)$, the joint distribution of midpoints and lengths of inversions, is not given.

Thompson, Wallace, and Federer [1965] have presented theoretical distributions of deficiency lengths, deficiency midpoints, and other related characteristics both for a discrete band model and a continuous chromosome model, as given by Model I.

DISCUSSION

In order to form a point of reference on a chromosome for determining the point of breakage it may be necessary to use two known reference points on a chromosome and to identify the point of breakage only as falling between these two points. This would involve grouping the data. Under Model I assumptions, Model VII was constructed for grouped data. This model resulted in good agreement of observed and expected values in the shortest inversion class for one set of data, but not for the second set. Model VIII is simply the truncated counterpart of Model VII. In some respects the models for grouped data fit the observed values better than the models for continuous data but in other respects one could not differentiate between the various models.

The authors are greatly indebted to Dr. B. P. Kaufmann for sending us the data involving 98 inversions.

SUMMARY

Six mathematical models, Models I to VI, for continuous data and two, Models VII and VIII, for grouped data were postulated to describe the distribution of lengths of inversions for a given chromosome. Model I postulated that the points of breakage (and reunion) were independent, and that any point of breakage was equally likely throughout the chromosome. Model III postulated that the points of breakage and reunion were related in that the endpoints of inversions tended to be closer together, or alternatively further apart, than would be expected on the basis of independence; the expected value of one endpoint given the value of the second endpoint is of the same form as linear regression. In Model V, a distribution of lengths of inversions was postulated, but the form of the dependence was not determined so that the problem was circumvented; this model may not appeal

to the biologist until the form of the dependence is determined.—Since observed data on lengths of inversions appeared to indicate a deficiency in the first class, this class was omitted and mathematical Models II, IV, and VI were constructed for the remaining classes. These procedures are the truncated counterparts of Models I, III, and V, respectively.—Maximum likelihood estimators and their variances for the dependence parameter were developed for Models III, IV, and V, and the procedure for doing this for Model VI was indicated. Utilizing these results where appropriate, Models I to V were fitted to the two available sets of data on lengths, totalling 147 inversions. All models fitted the data well, even though the models involving dependence of points of breakage resulted in a better fit. This, however, was to be expected, since another parameter was estimated. Biological events which decree that recoverable inversions may not be a random sample of chromosomes carrying two truly independent breaks may be responsible for the deficiency of short inversions.—The distribution of midpoints, distal points, and proximal points of inversions was obtained for Model I. The distribution of midpoints of inversions under Model III was also given.

LITERATURE CITED

Bauer, H., M. Demerec, and B. P. Kaufmann, 1938 X-ray induced chromosomal alterations in *Drosophila melanogaster*. Genetics **23**: 610–630.

Eisenhart, C., and F. S. Swed, 1943 Tables for testing randomness of grouping in a sequence of alternatives. Annals Math. Stat. **14**: 66–87.

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 3. Office of Naval Research. Contract Nonr-401(39), Project NR 042-212, January.

Gumbel, E. J., 1958 Distributions á plusieurs variables dont les marges sont données. C. R. Acad. Sci., Paris **246**: 2717–2720.

Parzen, E., 1960 *Modern Probability Theory and Its Application.* Wiley, New York.

Thompson, K. H., B. Wallace, and W. T. Federer, 1965 A mathematical model for the distribution of lengths of chromosomal deficiencies involving a specific locus. Genetics **51**: 887–896.