

THE COMBINATION OF LINKAGE VALUES, AND THE CALCULATION OF DISTANCES BETWEEN THE LOCI OF LINKED FACTORS.

By J. B. S. HALDANE, M.A.,
Fellow of New College, Oxford.

(With One Text-figure.)

ON the theory that the degree of linkage between two factors depends on the distance apart of their loci in a chromosome, Morgan and his fellow-workers have taken the distance between two loci as proportional to the cross-over value¹ of the factors located in them. This theory gives consistent results when the cross-over values are small, but, as recognised by Sturtevant, and by Morgan and Bridges(1), is not accurate for larger values. On the reduplication theory Trow(2) has given a formula for the combination of linkage values which is shown below to be inaccurate when the linkage is not close. In the present paper a more accurate theory of the relations *inter se* of the cross-over values, and of their connexion with the distances apart of the loci of factors in a chromosome, is developed. Some such theory is especially necessary when dealing with a group of factors containing few members not very closely linked.

Suppose A, B, C to be three factors whose loci lie in that order in the same chromosome. Let m be the cross-over value for A and B , n that for A and C . If the chromosomes were perfectly flexible, so that the fact of their having crossed between A and B did not diminish the probability of their crossing again between B and C , we should expect a triply heterozygous organism to produce gametes in the fol-

¹ If zygotes of composition $AB.ab$ and $Ab.aB$ give gametic series

$(1-m)AB : mAb : mA B : (1-m)ab$ and $mAB : (1-m)Ab : (1-m)aB : mab$ respectively, then m is said to be the cross-over value for the factors A and B .

lowing proportions if it were of composition $ABC.abc$, and similarly for other compositions:

No cross-over	$(ABC \text{ and } abc)$.	$(1-m)(1-n)$.
Cross-over between loci of A and B only	$(aBC \text{ and } Abc)$.	$m(1-n)$.
" " " B and C only	$(ABc \text{ and } abC)$.	$(1-m)n$.
" " " A and B and of B and C	$(AbC \text{ and } aBc)$.	mn .

Actually the last class has been shown to be in defect in many cases. This has been thought to be due to the loops formed by the chromosomes during synapsis having a modal length(3). If this were so, we should expect to find an excessive number of double cross-overs when the distance between the loci of A and C was equal to twice the modal distance between points of crossing over. This phenomenon has however not been recorded. The shortage of double cross-overs can equally well be explained by the mere rigidity of the chromosomes, which makes sharp bending difficult. In the sex chromosome of *Drosophila* the ratio¹ of observed to calculated numbers of double cross-overs is .58:1 for eosin (white), vermilion, and sable (4) (where $m+n = .406$), and .21:1 for vermilion, sable, and bar (3) (where $m+n = .239$).

If the calculated number mn of double cross-overs occurred, the cross-over value for A and C would be equal to the total number of single cross-overs, *i.e.* to $m(1-n) + (1-m)n$, or $m+n-2mn$.

If double cross-overs were impossible, but the full numbers of single cross-overs occurred, as would happen if the chromosomes were straight rigid rods, the cross-over value for A and C would obviously be $m+n$ (Morgan and Bridges' formula).

Finally if double cross-overs were impossible, and in every case where one should have occurred according to the calculation above, a single cross-over took its place, the cross-over value for A and C would be $m+n-2mn+mn$, or $m+n-mn$. This case might be approximately realised if the chromosome could not form loops shorter than some definite length.

Hence the cross-over values for A and C should be approximately $m+n$ when m and n are small, $m+n-2mn$ when their sum is large, and $m+n-mn$ for intermediate values.

Table I contains the observed values(5) for all triads of factors in the sex chromosome of *Drosophila* for which each of the cross-over values exceeds .1 (10%). The first column gives the three factors concerned in each case; the second and third columns give the cross-

¹ Called by Muller the "coincidence."

over values for the first and second, and second and third factors respectively, *i.e.* m and n . The fourth, fifth, and sixth columns give the results of the three provisional summation formulae obtained above; the seventh gives the observed cross-over value for the first

TABLE I.

1 Factors	2 m	3 n	4 $m+n$	5 $m+n-mn$	6 $m+n-2mn$	7 Observed	8 Class
White sable lethal <i>sc</i> ...	·412	·236	·648	·551	·454	·460	γ
Yellow vermilion rudimentary .	·345	·241	·586	·503	·420	·429	γ
" " bar ...	·345	·239	·584	·502	·419	·479	γ
White depressed bar ...	·203	·380	·583	·506	·429	·436	$\gamma?$
" sable forked ...	·412	·160	·572	·506	·440	·457	γ
Yellow sable rudimentary ...	·429	·143	·572	·511	·449	·429	δ
" bar ...	·429	·138	·567	·508	·449	·479	γ
White vermilion fused ...	·305	·258	·563	·484	·406	·433	γ
Bifid vermilion forked ...	·311	·245	·556	·480	·404	·425	$\gamma?$
White sable rudimentary ...	·412	·143	·555	·496	·437	·424	δ
Bifid vermilion rudimentary ...	·311	·241	·552	·477	·402	·427	γ
White vermilion forked ...	·305	·245	·550	·475	·401	·457	γ
" sable bar ...	·412	·138	·550	·493	·436	·436	$\gamma\delta$
Yellow miniature bar ...	·343	·205	·548	·478	·407	·479	β
White vermilion rudimentary ...	·305	·241	·546	·472	·399	·424	γ
" " bar ...	·305	·239	·544	·471	·398	·436	γ
" miniature bar ...	·332	·205	·537	·469	·401	·436	γ
Yellow miniature rudimentary...	·343	·179	·522	·471	·419	·429	γ
White miniature rudimentary .	·332	·179	·511	·452	·392	·424	γ
" reduplicated bar ...	·289	·206	·495	·435	·375	·436	β
" furrowed forked ...	·303	·191	·494	·436	·378	·457	$\beta?$
Bifid miniature rudimentary ...	·306	·179	·485	·430	·375	·427	$\gamma?$
White furrowed bar ...	·303	·179	·482	·428	·374	·436	$\beta?$
Yellow vermilion sable ...	·345	·101	·446	·411	·376	·429	β
Facet vermilion sable ...	·326	·101	·427	·394	·361	·430	$a?$
Depressed vermilion bar ...	·170	·239	·409	·368	·328	·380	$\beta?$
White vermilion sable ...	·305	·101	·406	·375	·344	·412	a
Shifted vermilion bar ...	·155	·239	·394	·357	·320	·314	$\delta?$
White depressed vermilion ...	·203	·170	·373	·338	·304	·305	$\gamma?$
Yellow club vermilion ...	·177	·188	·365	·332	·298	·345	β
White lethal <i>sb</i> miniature ...	·156	·199	·355	·325	·295	·332	β
White club vermilion ...	·143	·188	·331	·304	·277	·305	β
" lemon vermilion ...	·145	·120	·265	·248	·230	·305	$a?$
Vermilion sable forked ...	·101	·160	·261	·245	·229	·245	$\beta\gamma$
" " rudimentary ...	·101	·143	·244	·230	·215	·241	β
" " bar ...	·101	·138	·239	·225	·211	·239	$a\beta$

and third factors. In the eighth column these observed values are classified as follows:

- Greater than $m+n$ α
- Between $m+n$ and $m+n-mn$ β
- " $m+n-mn$ and $m+n-2mn$ γ
- Less than $m+n-2mn$ δ .

Those exactly equal to $m+n$ are classified as $a\beta$, and so on. The data are placed in the order of the magnitudes of $m+n$. Where any

of the three observed values is based on a count of less than 500 individuals (in which case the probable error of the cross-over value may exceed 1.5 %, as pointed out by the author(6) elsewhere) a query is placed in the last column.

It will be seen that the observed values, when $m+n$ exceeds .5, lie almost wholly between $m+n-mn$ and $m+n-2mn$, as demanded by the theory above. The three discordant values out of 19 are no more than would be expected in view of the probable errors of the observations due both to small numbers and differential mortality. When $m+n$ is less than .5 the results are somewhat more irregular, as the calculated values from the three formulae are not very different, but the majority of observations lie between $m+n$ and $m+n+mn$, as demanded by the theory.

This table also enables us to test the formulae given by Trow(2), based on the reduplication theory. If reduplication takes place so that A and B when coupled give the gametic series

$$qAB:1Ab:1aB:qab \left(\text{cross-over value } m = \frac{1}{q+1} \right),$$

whilst B and C give the series

$$rBC:1Bc:1bC:rbC \left(\text{cross-over value } n = \frac{1}{r+1} \right),$$

then A and C should give the series

$$(qr+1)BC:(q+r)Bc:(q+r)bC:(qr+1)bc \\ \left(\text{cross-over value} = \frac{q+r}{qr+q+r+1} \right).$$

This latter value $\frac{1}{q+1} + \frac{1}{r+1} - \frac{2}{(q+1)(r+1)} = m+n-2mn$. Hence on this hypothesis the observed cross-over values for A and C should cluster round $m+n-2mn$, and approximately equal numbers should be greater or less than it. In other words, as many values should fall in class δ as in classes α , β , and γ together.

The expectation is therefore 18 (δ), 18 (α , β , and γ); the actual numbers are 3.5 (δ), 32.5 (α , β , and γ), reckoning the single value $\gamma\delta$ as half in each class. Hence the above form of Trow's theory is untenable.

On a more complicated form of the same theory, which Sturtevant(7) has shown to be impossible on other grounds, A and C when coupled

alone give a primary series $sAC : lAc : lAc : sac$, and in zygotes of composition $ABC . abc$, a series

$$(qrs + s) AC : (q + r) Ac : (q + r) aC : (qrs + s) ac$$

$$\left(\text{cross-over value} = \frac{q + r}{qrs + q + r + s} \right).$$

As this value is less than that of $m + n - 2mn$, it is still more clearly impossible.

The supporters of the reduplication theory must therefore explain the deficiency of the double cross-over classes of gamete (which from a zygote of composition $ABC . abc$ are AbC and aBc). On the chromosome theory this is due to the rigidity of the chromosomes, and until an equally plausible explanation on the reduplication theory is given, the chromosome theory must be considered the more probable of the two, so far as the class of evidence dealt with in this paper is concerned.

It has been shown above that if A, B , and C are three factors whose loci lie in that order in the same chromosome, and if m and n are the cross-over values for A, B , and B, C respectively, then the value for A and C is $m + n - pmn$, where p is a number between 0 and 2, increasing on the whole with $m + n$, and having the value 1 when $m + n =$ about .5. The distances between loci may now be calculated as follows:

Let x be the distance between the loci of two factors, y their cross-over value, and let the unit of distance be chosen so that when y is sufficiently small x becomes equal to y . This assumption is legitimate if we suppose that crossing over is as likely to occur (other things being equal) at one point in the chromosome as another, i.e. that the chromosome is equally flexible and breakable at all points. The unit of distance is thus 100 times Morgan's unit.

If now we write $y = f(x)$, the form of this function being indeterminate,

$$\therefore f(x + h) = f(x) + f(h) - pf(x)f(h), \text{ where } h \text{ is any increment of } x.$$

$$\therefore \frac{f(x + h) - f(x)}{h} = \frac{f(h) - pf(x)f(h)}{h}$$

Now as h is decreased towards 0, $\frac{f(h)}{h}$ tends to the limit 1.

$$\therefore \frac{dy}{dx} = \text{Lt}_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

$$= \text{Lt}_{h \rightarrow 0} \frac{f(h) - pf(x)f(h)}{h}$$

$$= 1 - pf(x), \text{ where } p \text{ has the value assumed when } m = y, n = 0,$$

$$= 1 - py.$$



Therefore

$$x = \int_0^y \frac{dt}{1-pt}, \text{ since } x \text{ and } y \text{ vanish together, and } py < 1.$$

Hence if p were constant we should have

$$x = \frac{-1}{p} \log_e(1-py), \text{ or } y = \frac{1-e^{-px}}{p} \dots\dots\dots(1).$$

Since however p varies between 0 and 2, the values of x must lie between y and $\frac{-1}{2} \log_e(1-2y)$, those of y between x and $\frac{1-e^{-2x}}{2}$; the equation

$$y = x \dots\dots\dots(2)$$

being nearly accurate for small values of x and y , the equation

$$y = \frac{1-e^{-2x}}{2}, \text{ or } x = \frac{-1}{2} \log_e(1-2y) \dots\dots\dots(3)$$

for large values of x and y , as is obvious, since for large values of x , y approaches the value .5 asymptotically. The equation (2) corresponds to Morgan's summation formula $m+n$, the equation (3) to Trow's formula $m+n-2mn$.

The equation (3) may be deduced more directly as follows for a perfectly flexible chromosome:

Let a length x of the chromosome be considered as divided into a very large number N of small equal portions. Then the chance of a cross-over in each of these is approximately $\frac{x}{N}$. Hence the chance of a cross-over in t of these segments and no more is

$$\frac{N!}{t!(N-t)!} \left(\frac{x}{N}\right)^t \left(1-\frac{x}{N}\right)^{N-t}$$

When N becomes infinite the limiting value of this expression, *i.e.* the probability of exactly t and no more cross-overs in a length x , is

$$c_t = \frac{x^t e^{-x}}{t!} \dots\dots\dots(4).$$

Hence the value of y for a given value of x is the sum of the probabilities of all odd numbers of cross-overs.

$$\begin{aligned} \therefore y &= c_1 + c_3 + c_5 + c_7 + \dots\dots \\ &= e^{-x} \left(\frac{x}{1!} + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots\dots \right) \\ &= e^{-x} \sinh x \\ &= \frac{1-e^{-2x}}{2} \dots\dots\dots(5). \end{aligned}$$

In practice, however, owing to the rigidity of the chromosome, the value of c_1 thus calculated is too small, and those of $c_0, c_2,$ etc. too large. They are however more accurate for great lengths, where the rigidity of the chromosomes affects the results to a less extent.

It is suggested that the unit of distance in a chromosome as defined above be termed a "morgan," on the analogy of the ohm, volt, etc. Morgan's unit of distance is therefore a centimorgan.

To obtain a more accurate relation between x and y we may plot the curves representing equations (2) and (3), and then obtain empirically a curve lying between the two which fits the observed results as closely as possible. This has been done in the figure, where line (a) represents equation (2), curve (b) equation (3), and curve (c)

$$x = .7y - \frac{3}{2} \log_e (1 - 2y) \dots\dots\dots(5).$$

Equation (5) is merely chosen to give as good a fit as possible and has probably no theoretical significance. The points representing observations are plotted as follows:

The values of y in columns 2 and 3 of Table I are taken, and the corresponding distances in morgans (values of x) read off from curve (c) or Table II, which is calculated from equation (5). These latter are added together, and a point plotted with their sum as abscissa and the observed cross-over value from column 7 of Table I as ordinate. For example the first row of Table I gives the following result:

The cross-over values .412 and .236 correspond, according to the curve (c), or better, by interpolation from Table II, to distances of .549 and .261 morgans respectively. The sum of these distances is .810, and the observed cross-over value from column 7 is .460. The point farthest to the right is accordingly plotted with abscissa .810 and ordinate .460. Curve (c) gives the value .479 for y , and the error of y is accordingly .019, or 1.9%.

It will be seen that 18 of the observations lie above the curve (c), 18 below, and that in only 4 cases, 3 of which are among the results queried in Table I, does the error of y exceed .04 or 4%. The probable error of the cross-over values, as calculated from the curve, is 1.8%, or, omitting the queried results, 1.6%. This result is not large considering the probable errors of the values of y for the points plotted, which range from 3.1% downwards.

The curve gives satisfactory results for smaller cross-over values, but these are not plotted, as they do not allow of much discrimination be-



tween the three equations. If the points had been plotted from either line (*a*) or curve (*b*), $3\frac{1}{2}$ would have lain on one side, $32\frac{1}{2}$ on the other, as may be seen from Table I.

Hence the curve (*c*) may be taken as a fairly accurate guide to the combination of linkage values, and this remains equally true whether the chromosome theory is adopted or not. For this reason a series of values of $100x$ and $100y$ (*i.e.* distances in centimorgans and cross-over values as percentages) calculated from equation (5) are given in Table II. As more results accumulate it should be possible to correct these values, which are rather uncertain for large values of x and y .

TABLE II.

100 <i>y</i> (Cross-over value as percentage)	0.0	5.0	8.0	10.0	11.0	12.0	13.0
100 <i>x</i> (Distance in centimorgans)	0.0	5.1	8.2	10.3	11.4	12.5	13.6
100 <i>y</i> ... 14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0
100 <i>x</i> ... 14.7	15.9	17.0	18.1	19.3	20.5	21.7	22.9
100 <i>y</i> ... 25.0	26.0	27.0	28.0	29.0	30.0	31.0	32.0
100 <i>x</i> ... 27.9	29.2	30.5	31.9	33.3	34.7	36.2	37.7
100 <i>y</i> ... 36.0	37.0	38.0	39.0	40.0	41.0	42.0	43.0
100 <i>x</i> ... 44.3	46.1	48.0	50.0	52.2	54.4	56.9	59.6
100 <i>y</i> ... 47.0	48.0	49.0	49.5	49.7	49.8	49.9	50.0
100 <i>x</i> ... 75.1	81.9	93.0	99.2	109.4	117.7	128.1	∞

As an example of the use of this table the following problem may be taken:

"The factors *A* and *B* give a cross-over value of 38.5%, the factors *B* and *C* a value of 22.7%. What is the value for *A* and *C*?"

From the table we find by interpolation that the distance *AB* is 49.0 centimorgans, the distance *BC* 24.9. Hence the distance

$$AC = AB \pm BC = 73.9 \text{ or } 24.1 \text{ centimorgans.}$$

The cross-over value is therefore 46.8% or 22.0%, according as *C* lies outside *AB* or between *A* and *B*. Morgan's formula would have given 61.2% (an impossible value), or 15.8%; Trow's formula 43.7%, or 28.9% (by solving the equation $m + .227 - 2m \times .227 = .385$). On the reduplication theory the result from *AB* + *BC* corresponds to the view that the reduplication between *A* and *C* is "secondary" to those between *A*, *B* and *B*, *C*; the result from *AB* - *BC* to the view that the reduplication between *A* and *B* is secondary to those between *A*, *C* and *C*, *B*.

It should be remarked that the existence of a quantity x which has the property demonstrated above is not a conclusive proof of the chromosome theory, and indeed such a quantity may occur in certain forms (*e.g.* Trow's) of the reduplication theory. However the fact that the

values of x correspond to those demanded for the distance on the hypothesis that the factors are located in a semi-rigid chromosome is a strong point in favour of that hypothesis.

We have now the data for a fairly accurate estimate of the total length of the known portion of a chromosome, *e.g.* the sex chromosome in *Drosophila*. Taking some of the best authenticated measurements we have:

Factors	$100 y$ (Cross-over value in per cent.)	100 x (from Table II)
Yellow-White	1.1	1.1
White-Vermilion	30.5	35.4
Vermilion-Bar	23.9	26.5
Bar-Lethal <i>sc</i>	8.3	8.5
Totals	63.8	71.5

This gives a total length of 71.5 centimorgans against Morgan and Bridges' estimate(8) of 66.2. The discrepancy is due to the fact that in some comparatively long segments of the chromosome (*e.g.* between the loci of Sable and Rudimentary, a distance of about 15 centimorgans) no factors have been located, and such distances tend to be underestimated. It may also be due in part to the large probable error involved in using a large number of small distances.

From equation (4) we may calculate the proportion of chromosomes giving t cross-overs in the known region. These values are incorrect, owing to the rigidity of the chromosome, c_1 being too low, the remainder too high. The theoretical values are:

No cross-over in $c_0 = e^{-.715}$, or 49.1 % of the chromosomes

One " in $c_1 = .715e^{-.715}$, or 34.4 % " "

Two cross-overs in $c_2 = \frac{.715^2 e^{-.715}}{2}$, or 12.6 % " "

Three , in $c_3 = \frac{.715^3 e^{-.715}}{6}$, or 3.0 % " "

Four " in $c_4 = \frac{.715^4 e^{-.715}}{24}$, or .36 % " "

and so on.

The value of c_1 is too low, the others too high. The real value of $c_1 + c_3 + c_5 + \dots$ is the cross-over value of 46.3 %, and Morgan(8) gives $c_2 + c_4 + c_6 + \dots$, the number of double cross-overs (including quadruples, etc.), as about 10 %, so that c_0 should be about 43 %. When the relation between x and y is accurately known it will be possible to calculate the values of c_t with accuracy by integration.

It is believed that the above method of estimating distances will prove of considerable value when applied to comparatively long chromosomes in which factors are sparsely located, such as the second and third in *Drosophila*, since there is no reason to suppose that the relation arrived at between distance and cross-over value is peculiar to the sex chromosome in *Drosophila*. The results of investigations on these chromosomes should go far to confirm or refute the theory.

Outside *Drosophila* the best series of results on which to test it are those of Altenburg(9) with the three factors *M*, *S*, and *G* in *Primula sinensis*, quoted by Punnett(10) in a recent paper. Here the cross-over value for *M* and *S* is 11.6%, for *M* and *G* 34.0%, for *S* and *G* 40.6%, each result being based on 3684 individuals. By Table II the distance *SM* is 12.1 centimorgans, *MG* 40.9, and hence *SG* is 55.0 centimorgans (assuming the loci to lie in the order *SMG*). Hence the cross-over value for *S* and *G* should be 40.4%, the observed value being 40.6%, a very nearly perfect fit. The addition formula gives 45.6%, Trow's formula 37.7%. The probable error of the calculated result is .64%, of the observed .55%. Hence the probable value of their difference is .84%, and though the close agreement is accidental, both the alternative formulae are impossible.

In the case of Punnett's(10) results for sweet peas the agreement is also good, but owing to the closeness of the linkage, the three formulae give nearly equal values. There is, however, no reason to suppose that Table II does not represent with fair accuracy the relation between distance and cross-over value in all organisms, though the absolute value in $\mu\mu$ of the unit of distance, or morgan, is presumably different in different cases.

SUMMARY.

By a consideration of the observed gametic ratios of the sex-linked factors in *Drosophila*, the following results, among others, are arrived at:

1. If *A*, *B*, and *C* are three factors lying in a chromosome in that order, and if *m* is the cross-over value for *A* and *B*, *n* that for *B* and *C*, then the value for *A* and *C* lies between *m* + *n* and *m* + *n* - 2*mn*, being nearer to the former when *m* + *n* is small, to the latter when it is large.

2. A relation is arrived at, on the hypothesis that the chromosomes are partially rigid, between cross-over value and distance, which permits of the calculation of one of the cross-over values for three factors from the other two, with a probable error of less than 2%.

3. This relation may also be used to calculate the total length of a chromosome, and the number of double and triple cross-overs to be expected in a large distance.

4. The results from *Drosophila* are incompatible with Trow's form of the reduplication theory, but perhaps not with other possible forms of it.

5. The theory developed above fits all the observed data in plants.

REFERENCES.

1. MORGAN and BRIDGES. "Sex-linked inheritance in *Drosophila*." *Carnegie Institution of Washington*, 1916, p. 21.
2. TROW. "Primary and Secondary Reduplication." *Journal of Genetics*, Vol. II. p. 22, 1913.
3. MORGAN and BRIDGES. *Loc. cit.* p. 43.
4. ——— ——— *Loc. cit.* p. 37.
5. ——— ——— *Loc. cit.* p. 84.
6. HALDANE. "The Probable Errors of Observed Linkage Values." *Journal of Genetics*, Vol. VIII. p. 291, 1919.
7. STURTEVANT. "The Reduplication hypothesis applied to *Drosophila*." *American Naturalist*, Vol. XLVIII. p. 535, 1914.
8. MORGAN and BRIDGES. *Loc. cit.* p. 8.
9. ALTENBURG. *Genetics*, Vol. I. p. 354, 1916.
10. PUNNETT. "Reduplication series in Sweet Peas." *Journal of Genetics*, Vol. VI. No. 3, 1917.

