THE ESTIMATION OF MAP DISTANCES FROM RECOMBINATION VALUES

By D. D. KOSAMBI, *Poona, India*

Suppose three consecutive loci a, b, c of the same linkage group to have the recombination fractions (percentage divided by 100) $(a, b) = y_1$, $(b, c) = y_2$, $(a, c) = y_{12}$. Then it is known that for small values of y_1 and y_2, $y_{12} = y_1 + y_2$ approximately. For slightly larger values, we have a better approximation given by $y_{12} = y_1 + y_2 - y_1 y_2$; for still larger values, the approximation has again to be replaced by $y_{12} = y_1 + y_2 - 2y_1 y_2$. It is desired to obtain one single formula that will cover the entire range 0–$\frac{1}{2}$ of y-values in a reasonably satisfactory manner. This must also correspond to a single-valued, monotonically increasing, continuous function x of y in such a way that the corresponding identity becomes $x_{12} = x_1 + x_2$. The variable x will then be called the map distance corresponding to the given y.

Taking $y = f(x)$, our functional relation, assumed to be independent of the position on and number of the chromosome, must be of the form:

$$f(x + h) = f(x) + f(h) - pf(x) f(h). \tag{1}$$

The evidence that led to the conclusions of the first paragraph indicates that $f(x)/x \to 1$ as $x \to 0$. Also, that the unspecified function p increases from 0 to 2 with increasing x. Transposition and division by h gives

$$\frac{f(x + h) - f(x)}{h} = \frac{f(h)}{h} - pf(x) \frac{f(h)}{h}. \tag{2}$$

Taking limits as $h \to 0$, and assuming $f(x)$ to possess a derivative, we have

$$f'(x) = 1 - pf(x); \quad \text{or} \quad dy/dx = 1 - py. \tag{3}$$

So far, we have followed the arguments and derivation of J. B. S. Haldane (1919), who then fits an empirical curve from observed data for the X-chromosome, to obtain

$$x = 0{\cdot}7y - 0{\cdot}15 \log_e (1 - 2y). \tag{4}$$

This fits the observed data reasonably well, and seems to fit other data also, to a considerable extent. But this amounts to abandoning (3) or taking $p = 0{\cdot}6/(1 - 1{\cdot}4y)$, which does not agree with our hypotheses. At best, (4) would indicate the existence of a general formula of the type desired. It is seen that formula (4) cannot conveniently be inverted, the usual method of use being by means of a table calculated by Haldane at intervals of $0{\cdot}01$ for those ranges of values of y where the deviation from Morgan's first formula $y_{12} = y_1 + y_2$ becomes serious. The method would be, then, to find the values of x for given y (by interpolation if necessary) add, and then change back by using the table again.

It seems, however, possible to take one further step directly from the differential equation (3), by making a very plausible hypothesis about the unknown function p. This depends in some way on x and must increase steadily so far as known. The simplest such function would be one linear in x and y, and the simplest linear function taking the values 0 and 2 at the two ends of the range is, obviously, $4y$ in view of the fact that no recombination value can exceed 50 %. We thus obtain

$$dy/dx = 1 - 4y^2. \tag{5}$$

This integrates at once to the very simple solution:

$$2y = \tanh 2x; \quad x = \tfrac{1}{4}\log\frac{1+2y}{1-2y}. \tag{6}$$

The tables to use are, therefore, those of Fisher & Yates (1938) for the transformation of the correlation coefficient, with $2y = r$, $2x = z$. The chief advantage of formula (6) is that we obtain a direct combination value

$$y_{12} = \frac{y_1 + y_2}{1 + 4y_1 y_2}. \tag{7}$$

The similarity of this with the velocity-addition formula in the special theory of relativity should not be made the basis of more bad philosophy.

Formula (7) eliminates the use of tables and correction curves. In the examples to be found in our text-books, and in such other cases for which I have been able to obtain reasonably good data, the formula works at least as well as Haldane's. The use of tables would be necessary in comparing the lengths of two chromosomes, in accurate determination of the position in terms of x of the spindle-fibre attachment, and so forth. A comprehensive recasting of available data on map distances is not possible at present, because I have no access to the necessary bibliographic material, and also because a good deal of the data seems to have been estimated by statistically unsatisfactory methods.

For example, the data quoted by Haldane gives

yellow-vermilion-rudimentary $= 0{\cdot}345{-}0{\cdot}241{-}0{\cdot}429,$

yellow-vermilion-bar $= 0{\cdot}345{-}0{\cdot}239{-}0{\cdot}479,$

which would indicate that the sum of a given distance to a fixed distance is more when the distance is shorter, contradicting our hypotheses. Similarly, for

yellow-sable-rudimentary $= 0{\cdot}429{-}0{\cdot}143{-}0{\cdot}429,$ yellow-sable-bar $= 0{\cdot}429{-}0{\cdot}138{-}0{\cdot}479.$

These figures are also connected with such questions as the analysis of interference. Bridges & Morgan (1923, p. 6) give the recombination percentage between sepia and Minute-f as 52·4, which is impossible. The same authors give

lethal-iiih-Dichaete-Hairless $= 0{\cdot}177{-}0{\cdot}234{-}0{\cdot}489,$

so that $y_{12} > y_1 + y_2$, which can only be explained by discarding the formulae or by emphasizing the paucity of the data and difficulty of locating lethals. Finally, we are given (Bridges & Morgan, 1923, p. 4) Dichaete-spineless recombination fraction as 0·137 from 3030 primary and as 0·153 from 9143 secondary observations, both sets being supposed (Bridges & Morgan, 1923, p. 21) 'on an equal footing...in calculating recombination percents'. If we restore the original recombination numbers from the given percentages, a rapid calculation gives $\chi^2 = 4{\cdot}62$ (without Yates's correction) which is significant at the 5 % level for a single degree of freedom, making it unlikely that the two sets of figures locate the same point. This is not surprising, as salivary maps by Bridges and others seem to show, if I am not mistaken, that certain loci refer to whole sections of the chromosome. Under these circumstances, the use of any formula is naturally limited.

A few comments may nevertheless be useful. Various modifications of our most useful hypothesis, $p = 4y$, may be made if required by the evidence. All functions $p = a(x) + yb(x)$ lead to Riccatian equations, which may be integrated without much trouble. Another possibility is that of restricting the passage to the limit from equation (2), obtaining a difference instead of a differential equation. But with $p = 4y$, it will be seen that the leading term in the solution will be of the same type as for the differential equation. One possible use of the last modification would be the

derivation of formulae that retain their validity when there is a known minimum length of the chromosome that acts as a cross-over quantum. For the present there seems to be no evidence that would require any definite change of the formulae derived in (6) and (7).

If y, the recombination value, is to be treated as a probability, the methods of Fisher (1937, chap. XI) show the amount of information about the distance x in a sample of n observations to be

$$I_x = \frac{n(1-4y^2)^2}{y(1-y)}. \tag{8}$$

It is this, and not n itself, that should be used as a weight in estimating the same x from parallel observations. *Relative to y*, the maximum information about x is obtained at $y = 0\cdot25$, so that a new locus should be estimated from others which give about a quarter of the total number as recombinations. The point of maximum efficiency relative to x is a little farther to the right, so that slightly greater recombination values would do. The problem of efficient estimation of recombination values has been treated directly by Fisher (1937, pp. 235–52).

Suppose that between a new gene and a known one n observations give m recombinations; for a second gene, n' and m'; between the two reference loci M recombinations occur in N cases. By least squares the 'best' values of the recombination fractions y_1, y_2 between the two markers and the gene to be located would be those minimizing

$$w_1\left(y_1 - \frac{m}{n}\right)^2 + w_2\left(y_2 - \frac{m'}{n'}\right)^2 + w_{12}\left(y_{12} - \frac{M}{N}\right)^2, \tag{9}$$

where y_{12} has the value given by (7). Here y has to be used in place of x because the distribution is much nearer to normal. For the weights most experimenters would choose $w_i = n_i$, the number of observations, though the proper value would be the amounts of information: $w_i = n_i/y_i(1-y_i)$, which make (9) yield the value of χ^2. For the more efficient maximum likelihood estimates, we should equate to zero the first partial derivatives of $S\{m_i \log y_i + (n_i - m_i) \log(1 - y_i)\}$, always taking y_{12} as in (7).

To illustrate, we simplify still further by taking the recombination value between markers as precisely known (M, N very large). Then $y_{12} = a$, $y_1 = y$, $y_2 = (a-y)/(1-4ay)$, and we have

$$\chi^2 = \frac{(ny-m)^2}{ny(1-y)} + \frac{\{n'(a-y) - m'(1-4ay)\}^2}{n'(a-y)\{(1-a) + y(1-4a)\}}, \tag{10}$$

and

$$\log L = \text{const.} + m \log y + (n-m) \log(1-y) - n' \log(1-4ay)$$
$$+ m' \log(a-y) + (n'-m') \log\{(1-a) + y(1-4a)\}. \tag{11}$$

The minimum χ^2 gives an equation of the sixth degree, whereas equating L' to zero gives a quartic:

$$ny^4 - b_1 y^3 + b_2 y^2 - b_3 y + b_4 = 0, \tag{12}$$

where

$$b_1 = n(a+r+s) + m + n'(r-s) + m'(s-a),$$
$$b_2 = n(ar+as+rs) + m(a+r+s) + n'(r-s)(1+a) + m'(s-a)(1+r),$$
$$b_3 = n(ars) + m(ar+as+rs) + n'a(r-s) + m'r(s-a),$$
$$b_4 = m(ars); \quad \text{with} \quad r = \tfrac{1}{4}a \text{ and } s = (1-a)/(4a-1).$$

D. de Winton & J. B. S. Haldane (1935, p. 75) give for *Primula*–II ♀ the y-values PF-FCh-PCh corrected as 15·10–10·35–23·92, whereas our formula (7) should give 23·95 for the last, a very good fit; the uncorrected (backcross cross-overs) values are 14·52–10·83–23·10, where the last should have been 23·85 for consistence by (7). If we took 23·92 as the fixed value and worked only

with the backcross data, we should have $a = 0.2392$, $n = 1253$, $m = 182$, $n' = 2613$, $m' = 283$, which gives

$$y^4 - 18.77314y^3 + 15.2263y^2 + 2.5593y - 0.63951 = 0,$$

the root between 0 and 0.5 being 0.146 to the nearest three figures, which is hardly an improvement worth the trouble, the present case being merely an illustration. The standard error is immediately calculated, as usual, by taking the reciprocal of $-L''$ as the variance when the value of y from (12) is substituted. For the more general formulae with several y-values determined simultaneously, the best method is to substitute the observed values in the maximum-likelihood equations and proceed by successive approximations.

De Winton & Haldane (1935, pp. 96–7) extend our postulates by considering the nature of the coincidence. This amounts to the extra condition that $p(x,y)/2y \to$ const. as $y \to 0$. Taking $p = 8y/(1+2y)$, Haldane integrates the differential equation to get

$$12x = \log(1+4y) - 4\log(1-2y).$$

But $p = 2y/(1-y)$ also satisfies all conditions to give $6x = 4\log(1+y) - \log(1-2y)$, which gives somewhat better consistency in the values of x, here the sole criterion, as there is no intrinsic unit of map distance. That formula is the most suitable where the distances are additive to within the limits of significance. The data from *Primula*–I may be used for the purposes of comparison (de Winton & Haldane, 1935, p. 98):

	SB	SG	SL	BG	BL	GL
♀ y	6.25	34.40	37.53	32.14	36.73	3.61
$\bar{x}_0$	6.26	42.20	48.72	38.15	46.93	3.61
$\bar{x}_1$	6.33	46.06	54.03	41.21	51.77	3.61
$\bar{x}_2$	6.27	39.12	44.39	35.72	42.97	3.61
♂ y	11.55	41.03	41.45	35.01	38.86	1.82
$\bar{x}_0$	11.76	57.94	59.24	43.38	47.22	1.82
$\bar{x}_1$	11.92	65.37	67.07	47.45	52.11	1.82
$\bar{x}_2$	11.66	51.56	52.55	40.09	43.19	1.82

where $\bar{x}_0$ is from formula (6) of this note, $\bar{x}_1$ is Haldane's revised formula above, and $\bar{x}_2$ ours. Others could be devised very easily, as for example by taking the simple value $p(x,y) = 2y + 4y^2$, which satisfies all the conditions to give the very clumsy result

$$10x = 4\tan^{-1}(1+2y) + \log(2y^2 + 2y + 1) - 2\log(1-2y) - \pi;$$

this gives more trouble in the calculation with actually less consistence in the fit for map distances. Besides being less trouble to calculate, the inverse hyperbolic tangent formula has the tremendous advantage of the handy composition rule (7), which also allows use in actually fitting cross-over values and map distances from the observational data.

REFERENCES

C. B. BRIDGES & T. H. MORGAN (1923). *The Third Chromosome Group of Mutant Characters of* Drosophila melanogaster. Carnegie Institute of Washington, Publication no. 327.

R. A. FISHER (1937). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

R. A. FISHER & F. YATES (1938). *Statistical Tables*, Table VII. Oliver and Boyd, Edinburgh.

J. B. S. HALDANE (1919). *J. Genet.* 8, 299–309.

D. DE WINTON & J. B. S. HALDANE (1935). *J. Genet.* 31, 67–100.