THE ESTIMATION OF LINKAGE IN BACTERIA

NORMAN T. J. BAILEY

Department of Medicine, University of Cambridge

Received 20.iv.50.

CONTENTS

1. INTRODUCTION

THE existence of genetic recombination and linkage in a strain of *Escherichia coli* (Lederberg and Tatum, 1946 *a*, *b*, *c* ; Tatum and Lederberg, 1947 ; Lederberg, 1947) can now be regarded as definitely established. Recombination was first discovered in mixed populations of multiple nutritional mutants by the appearance of prototrophs, which, unlike the parent types, were capable of growing in minimal media. The use of multiple mutants renders the possibility of contamination by back-mutation extremely improbable, especially as prototrophs appear only in mixed populations. The evidence suggests that the vegetative cells are haploid. The existence of linkage between some nutritional factors can be shown by testing colonies grown on a medium supplemented by a single nutritional requirement (Lederberg 1947, pp. 512-13). More conclusive evidence is provided by the introduction of other contrasting characters such as the ability or inability to ferment certain sugars, and resistance or susceptibility to various types of bacteriophage. The frequencies with which the possible combinations arise in prototrophs are most economically interpreted by postulating a linear arrangement of loci, which so far all fall into a single linkage group.

In the usual type of breeding experiment the offspring can all be classified as showing no recombination or as belonging to one of a number of recombination classes. With bacteria, on the other hand, we usually examine only members of the recombination class represented by prototrophs, and even if with an improved technique we could select on the basis of factors other than nutritional factors, we should still have to classify organisms for which *some* recombination had been obligatory. However, it is still possible, with the aid of certain assumptions, not only to test for the existence of linkage but also to estimate recombination fractions and to make tentative linkage-maps.

Recombination, in *E. coli K*-12 at any rate, occurs exceedingly rarely. But when conditions obtain which are favourable to the production of prototrophs, the recombination of a variety of other factors is common. The situation is somewhat analogous to that arising in the detection of linkage in human families, where the difficulty is to find matings of the right type. With bacteria, " matings " of any sort are probably extremely infrequent, but the selection of prototrophs for further examination enables us to sift out just those organisms which will yield information about linkage.

If the pair of contrasting characters, for which both members of a pair are scorable, are distal to a segment over which recombination is obligatory, then the treatment of linkage closely follows the methods normally used in ordinary multiple crosses, at least on the assumption of no interference. But if these pairs lie in the segment over which recombination is obligatory, then special methods are required.

The object of the present paper is to deal with the latter contingency, which has so far proved to be the commonest case. As in classical Mendelian crosses, we should consider the possibility of disturbances due to differential viability or misclassification (partial manifestation). For the simpler four-point crosses we are obliged to make definite assumptions about the degree of interference present since the numbers of degrees of freedom available are not sufficient to permit the estimation of interference as well as linkage, and we cannot make separate estimates of single and triple recombination as in the familiar type of linkage-cross. With crosses involving five points or more the estimation of interference should be possible. In the treatment given here interference will be assumed negligible, although I hope to deal with this topic another time.

2. NOTATION AND DEFINITIONS

It will be convenient to represent pairs of characters, one of which is compulsorily selected, by : A,*a* ; B,*b* ; C,*c* ; . . . where the upper case letters correspond to the character selected. Thus we shall be concerned with crosses of the type : AB*cd* . . . = *ab*CD . . ., yielding prototrophs ABCD . . . . It is common, though not theoretically essential, for such characters to be nutritional. Characters, which can be scored in prototrophs for either member of a contrasting pair, will be represented by : x,x' ; y,y' ; z,z' ; . . . . These will often be characters like phage resistance or sugar fermentation.

As we are confining our attention to loci lying within a segment over which recombination is obligatory, we shall be concerned with crosses of the type A xyz . . . b = a $x'y'z'$ . . . B. Such crosses may in fact contain additional factors C,*c* ; D,*d* ; . . . in the distal portions of the segments to facilitate the selection of prototrophs.

The *order* of a cross is denoted by the number of pairs of loci, including end-points, involved in a segment over which recombination is compulsory. Thus A$xyzb$ = $ax'y'z'$B is a five-point cross.

3. TEST OF SIGNIFICANCE FOR LINKAGE

Linkage will be suggested by the failure of (x,x') characters to segregate independently. Disturbed frequencies may result from the physiological properties of the alleles, giving rise to differential viability or partial manifestation. Either of the latter possibilities can be allowed for by employing " reversed " crosses, corresponding to use of matings in both coupling and repulsion in ordinary linkage backcrosses. Table 1 shows how the basic data are classified : *a*, *b*, *c* and *d* are the observed numbers.

TABLE 1

Parents	Prototrophs (ABCD . . .)	
	x	x'
ABcd . . . ; x = abCD . . . ; x'	*a*	*b*
ABcd . . . ; x' = abCD . . . ; x	*c*	*d*

Linkage is indicated if the observed numbers in the fourfold table depart significantly from proportionality.

The first crosses reported by Lederberg and Tatum (1946) gave :

TABLE 1A

Parents	Prototrophs ($B^+M^+P^+T^+$)	
	V_1^r	V_1^s
$B^-M^-P^+T^+$ = $B^+M^+P^-T^-$ . . . . . V_1^r . . . . . V_1^s	8	2
. . . . . V_1^s . . . . . V_1^r	3	7

The exact treatment of the fourfold table gives P = 3·5 per cent. Later work (Lederberg, 1947) confirmed the suggestion of linkage.

Estimation of linkage

We will now consider the problem of estimating recombination fractions, and shall assume that interference is negligible.

4. THREE-POINT CROSSES

Suppose we have a cross of the type : Axb = ax'B, represented diagrammatically in fig. 1.

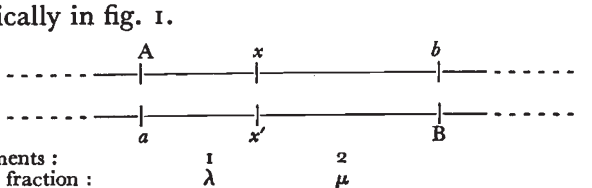

FIG. 1.

The segments (Ax) and (xb) are numbered 1 and 2. The recombination fractions for these segments, are, respectively, λ and μ.

Now recombination over (Ab) is obligatory, by reason of the selection of prototrophs, and we have either recombination over 1 and none over 2, or vice versa. Thus, if the frequencies are undisturbed, the expectations for a sample of n tested colonies is as given in table 2.

TABLE 2

	Prototrophic types		Total
	x' (Rec. over 1)	x (Rec. over 2)	
Expected . . .	$n\lambda(1-\mu)/R$	$n(1-\lambda)\mu/R$	n
Observed . . .	a	b	n

where R is the recombination fraction over the whole segment from A to b, and is given by :

$$R = \lambda(1-\mu)+(1-\lambda)\mu = \lambda+\mu-2\lambda\mu.$$

There is only one degree of freedom at our disposal so we cannot hope to estimate λ and μ separately. Moreover, the likelihood is a function of $\left(\frac{1}{\lambda}-1\right)\Big/\left(\frac{1}{\mu}-1\right)$, so that we cannot even estimate the ratio λ/μ or $p = \lambda/(\lambda+\mu)$ satisfactorily unless both λ and μ are small compared with unity. If the latter condition is fulfilled we have the estimates :

$$\left.\begin{aligned} \hat{p} &= \frac{\lambda}{\lambda+\mu} = \frac{a}{n} \\ \operatorname{var}\hat{p} &= \frac{p(1-p)}{n} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (4.1)$$

5. FOUR-POINT CROSSES

Now let us consider a four-point cross : Axy b = $ax'y'$B, with a notation given in fig. 2.

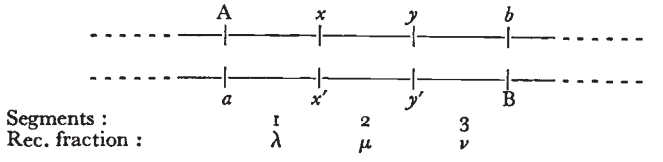

FIG. 2.

The observations fall into four classes according to whether there is recombination over only one segment, or over all three. There are now three degrees of freedom available and we should be able to

estimate the actual values of λ, μ and ν. The observed and expected values are as in table 3.

TABLE 3

	Prototrophic types				Total
	$x'y'$ (Rec. over 1)	xy' (Rec. over 2)	xy (Rec. over 3)	$x'y$ (Rec. over 1, 2 and 3)	
Expected	$n\lambda(1-\mu)(1-\nu)/R$	$n(1-\lambda)\mu(1-\nu)/R$	$n(1-\lambda)(1-\mu)\nu/R$	$n\lambda\mu\nu/R$	
Observed	a	b	c	d	n

where R, the recombination fraction over the whole interval 1+2+3, is given by

$$\left.\begin{aligned} R &= \lambda+\mu+\nu-2(\lambda\mu+\mu\nu+\nu\lambda)+4\lambda\mu\nu \\ &= \tfrac{1}{2}[(1-(1-2\lambda)(1-2\mu)(1-2\nu)] \end{aligned}\right\} \quad . \quad . \quad (5.1)$$

The logarithm of the likelihood is :

$$\begin{aligned} L = {} & (a+d)\log\lambda+(b+d)\log\mu+(c+d)\log\nu+(b+c)\log(1-\lambda) \\ & +(a+c)\log(1-\mu)+(a+b)\log(1-\nu)-n\log R \end{aligned} \quad (5.2)$$

Therefore, the maximum likelihood equations are :

$$\left.\begin{aligned} \frac{\partial L}{\partial\lambda} &= \frac{a+d}{\lambda}-\frac{b+c}{1-\lambda}-\frac{n}{R}(1-2\mu)(1-2\nu) = 0 \\ \frac{\partial L}{\partial\mu} &= \frac{b+d}{\mu}-\frac{a+c}{1-\mu}-\frac{n}{R}(1-2\nu)(1-2\lambda) = 0 \\ \frac{\partial L}{\partial\nu} &= \frac{c+d}{\nu}-\frac{a+b}{1-\nu}-\frac{n}{R}(1-2\lambda)(1-2\mu) = 0 \end{aligned}\right\} \quad . \quad . \quad (5.3)$$

Since we have three degrees of freedom and three unknowns we should expect the solution of (5.3) to be the same as the solution of the equations given by equating the expectations to the observations, *i.e.* :

$$\left.\begin{aligned} n\lambda(1-\mu)(1-\nu)/R &= a \\ n(1-\lambda)\mu(1-\nu)/R &= b \\ n(1-\lambda)(1-\mu)\nu/R &= c \\ n\lambda\mu\nu/R &= d \end{aligned}\right\} \quad . \quad . \quad . \quad (5.4)$$

The solution of (5.4) is clearly given by :

$$\frac{\lambda}{1-\lambda}=\left(\frac{ad}{bc}\right)^{\frac{1}{2}}, \quad \frac{\mu}{1-\mu}=\left(\frac{bd}{ac}\right)^{\frac{1}{2}}, \quad \frac{\nu}{1-\nu}=\left(\frac{cd}{ab}\right)^{\frac{1}{2}} \quad . \quad . \quad (5.5)$$

and it is easy to verify that (5.5) is in fact the solution of (5.3).

The formulæ (5.5) are analogous to those previously used in

linkage backcrosses (Fisher, 1935-49 ; Bailey, 1949) and application of the usual formula

$$\text{var T} = \Sigma m\left(\frac{\partial T}{\partial a}\right)^2 - n\left(\frac{\partial T}{\partial n}\right)^2$$

gives the following large sample variances :

$$\left.\begin{aligned} \text{var}\,\hat{\lambda} &= \lambda^2(1-\lambda)^2/h \\ \text{var}\,\hat{\mu} &= \mu^2(1-\mu)^2/h \\ \text{var}\,\hat{\nu} &= \nu^2(1-\nu)^2/h \end{aligned}\right\} \quad . \quad . \quad . \quad (5.6)$$

where $\frac{4}{h} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$, and λ, μ, ν are calculated from (5.5).

The product formulæ (5.5) are presumably those employed by Lederberg (1947) in calculating recombination fractions, although he gives neither the formulæ nor the standard errors. If λ, μ and ν are all small the triple-recombination class will be very small and $1/h \sim 1/4d$ will be large. Thus the standard errors will all be relatively large, as stated by Lederberg. If λ, μ and ν are not small the theoretical precisions will be much higher, although it may no longer be easy to recognise the triple recombination class and then the order of the loci may be uncertain.

A test for differential viability *

The above treatment of a four-point cross ignores the possibility of differential viability, and to take account of this we are obliged to employ different mating types. The easiest test for the existence of differential viability is obtained when we have data for crosses which are " reversed " with respect to both x,x' and y,y'.

Let us assume that the viability of x' relative to x is u ; and of y' relative to y is v. Then the expectations and observations are as follows :

TABLE 4

Rec. over	1. $Axyb = ax'y'B$			2. $Ax'y'b = axyB$		
	Prototrophic type	Expected	Observed	Prototrophic type	Expected	Observed
1	$x'y'$	$uv.\ \lambda(1-\mu)(1-\nu)$	a	xy	$1.\ \lambda(1-\mu)(1-\nu)$	e
2	xy'	$v.\ (1-\lambda)\mu(1-\nu)$	b	$x'y$	$u.\ (1-\lambda)\mu(1-\nu)$	f
3	xy	$1.\ (1-\lambda)(1-\mu)\nu$	c	$x'y'$	$uv.\ (1-\lambda)(1-\mu)\nu$	g
123	$x'y$	$u.\ \lambda\mu\nu$	d	xy'	$v.\ \lambda\mu\nu$	h
		$\times n_1/R_1$	n_1		$\times n_2/R_2$	n_2

$$\left.\begin{aligned} \text{where } R_1 &= uv.\lambda(1-\mu)(1-\nu) + \ldots + \ldots + \ldots \\ R_2 &= 1.\lambda(1-\mu)(1-\nu) + \ldots + \ldots + \ldots \end{aligned}\right\} \quad . \quad (5.7)$$

* We can of course examine the data for general disturbances by testing the 2×4 contingency table. The present test is relevant to a more specific hypothesis and permits the two viabilities to be separated.

And let us write

$$\left.\begin{aligned} a+e &= \mathrm{A} \\ b+f &= \mathrm{B} \\ c+g &= \mathrm{C} \\ d+h &= \mathrm{D} \\ n_1+n_2 &= \mathrm{N} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (5.8)$$

$$\begin{aligned} \therefore \mathrm{L} = {} & (\mathrm{A}+\mathrm{D})\log\lambda+(\mathrm{B}+\mathrm{D})\log\mu+(\mathrm{C}+\mathrm{D})\log\nu+(\mathrm{B}+\mathrm{C})\log(1-\lambda) \\ & +(\mathrm{A}+\mathrm{C})\log(1-\mu)+(\mathrm{A}+\mathrm{B})\log(1-\nu) \\ & +(a+d+f+g)\log u+(a+b+g+h)\log v \\ & -n_1\log \mathrm{R}_1-n_2\log \mathrm{R}_2 \quad . \quad . \quad . \quad . \quad . \quad (5.9) \end{aligned}$$

To derive a significance test we calculate scores in the usual way by differentiating L and then put $u = v = 1$.

The equation $\frac{\partial \mathrm{L}}{\partial\lambda} = \frac{\partial \mathrm{L}}{\partial\mu} = \frac{\partial \mathrm{L}}{\partial\nu} = 0$ are similar to (5.3) when we put $u = v = 1$, but with A, B, C, D written for a, b, c, d. The solutions are therefore :

$$\frac{\lambda}{1-\lambda} = \left\{\frac{\mathrm{AD}}{\mathrm{BC}}\right\}^{\frac{1}{2}}, \quad \frac{\mu}{1-\mu} = \left\{\frac{\mathrm{BC}}{\mathrm{AC}}\right\}^{\frac{1}{2}}, \quad \frac{\nu}{1-\nu} = \left\{\frac{\mathrm{CD}}{\mathrm{AB}}\right\}^{\frac{1}{2}} \quad . \quad (5.10)$$

by comparison with (5.5).

Substituting these values in the scores for u and v, we find :

$$\left.\begin{aligned} \mathrm{S}_{u=1} &= \frac{2}{\mathrm{N}}\left\{n_2(a+d)-n_1(e+h)\right\} \\ \mathrm{S}_{v=1} &= \frac{2}{\mathrm{N}}\left\{n_2(a+b)-n_1(e+f)\right\} \end{aligned}\right\} \quad . \quad . \quad . \quad (5.11)$$

The variances and covariances of these expressions are easily obtained by the usual multinomial distribution theory and are given by the matrix :

$$V = \begin{bmatrix} \alpha(\mathrm{A}+\mathrm{D})(\mathrm{B}+\mathrm{C}) & \alpha(\mathrm{AC}-\mathrm{BD}) \\ \alpha(\mathrm{AC}-\mathrm{BD}) & \alpha(\mathrm{A}+\mathrm{B})(\mathrm{C}+\mathrm{D}) \end{bmatrix} \quad . \quad . \quad (5.12)$$

where $\alpha = 4n_1n_2/\mathrm{N}$.[3]

We shall require the inverse of V, *i.e.*

$$V^{-1} = \begin{bmatrix} \beta(\mathrm{A}+\mathrm{B})(\mathrm{C}+\mathrm{D}) & \beta(\mathrm{BD}-\mathrm{AC}) \\ \beta(\mathrm{BD}-\mathrm{AC}) & \beta(\mathrm{A}+\mathrm{D})(\mathrm{B}+\mathrm{C}) \end{bmatrix} \quad . \quad . \quad (5.13)$$

where $\beta = \mathrm{N}^2/4\mathrm{K}n_1n_2\mathrm{ABCD}$

and $\mathrm{K} = \frac{1}{\mathrm{A}} + \frac{1}{\mathrm{B}} + \frac{1}{\mathrm{C}} + \frac{1}{\mathrm{D}}$

We can then use the joint test for viability given by a χ^2 with 2 degrees of freedom :

$$\chi_2^2 = S'V^{-1}S \quad . \quad . \quad . \quad . \quad (5.14)$$

where $S = \begin{bmatrix} \mathrm{S}_{u=1} \\ \mathrm{S}_{v=1} \end{bmatrix}$

This test would be much more convenient if we could find a relatively simple algebraic expression for (5.14).

If viability is found significant then we must return to (5.9) and proceed in the usual way by the use of scores and amounts of partial information.

The use of the significance test is illustrated by its application to some of Lederberg's data.

Lederberg (1947, page 514) gives the results of various crosses. Let us consider two types of cross only, viz.

1. $(B^+M^+)Lac^-V_1^r(T^-L^-) = (B^-M^-)Lac^+V_1^s(T^+L^+)$
2. $(B^+M^+)Lac^+V_1^s(T^-L^-) = (B^-M^-)Lac^-V_1^r(T^+L^+)$

Although it is doubtful whether the B_1^+ and the B_1^{-**} results are really homogeneous, we shall group these together to give the following data :

TABLE 4A

Rec. over	1	2	Totals
1	$Lac^+V_1^s$ a : 141	$Lac^-V_1^r$ e : 130	A : 271
2	$Lac^-V_1^s$ b : 296	$Lac^+V_1^r$ f : 247	B : 543
3	$Lac^-V_1^r$ c : 241	$Lac^+V_1^s$ g : 128	C : 369
123	$Lac^+V_1^r$ d : 18	$Lac^-V_1^s$ h : 13	D : 31
Total	n_1 : 696	n_2 : 518	N : 1214

Applying the foregoing results we obtain :

$$S_{u=1} = -28{\cdot}2800,\ S_{v=1} = -59{\cdot}3509 \quad . \quad . \quad (5.15)$$

$$V = \begin{bmatrix} 221{\cdot}9954 & 67{\cdot}0329 \\ 67{\cdot}0329 & 262{\cdot}4379 \end{bmatrix}, \quad . \quad . \quad . \quad . \quad (5.16)$$

and

$$V^{-1} = \begin{bmatrix} 4{\cdot}88105 \times 10^{-3} & -1{\cdot}24674 \times 10^{-3} \\ -1{\cdot}24674 \times 10^{-3} & 4{\cdot}12886 \times 10^{-3} \end{bmatrix} \quad . \quad . \quad (5.17)$$

Using (5.14) this gives $\chi_2^2 = 14{\cdot}26$, which is highly significant.*

It should be mentioned that the slightly more obvious procedure of distributing the six available degrees of freedom between the three recombination fractions and *three* differential viabilities (for four phenotypes) in fact fails to be of service, as the resulting maximum likelihood equations are not uniquely soluble.

6. FIVE-POINT CROSSES

A five-point cross can be represented by

$$A\ xyzb = ax'y'z'B$$

* Separate examination of S_u and S_v shows that it is only the viability of V^s_1 relative to V^r_1 which is clearly significant.

Applying the method of maximum likelihood we obtain the following scores and amounts of partial information :

$$\left.\begin{aligned} S_\kappa &= \frac{a+e+f+g}{\kappa} - \frac{b+c+d+h}{1-\kappa} - \frac{n\left(\frac{1}{R}-2\right)}{1-2\kappa} \\ S_\lambda &= \frac{b+e+f+h}{\lambda} - \frac{a+c+d+g}{1-\lambda} - \frac{n\left(\frac{1}{R}-2\right)}{1-2\lambda} \\ S_\mu &= \frac{c+e+g+h}{\mu} - \frac{a+b+d+f}{1-\mu} - \frac{n\left(\frac{1}{R}-2\right)}{1-2\mu} \\ S_\nu &= \frac{d+f+g+h}{\nu} - \frac{a+b+c+e}{1-\nu} - \frac{n\left(\frac{1}{R}-2\right)}{1-2\nu} \end{aligned}\right\} \quad . \quad (6.2)$$

$$\left.\begin{aligned} I_{\kappa\kappa} &= \frac{n[R(1-R)-\kappa(1-\kappa)]}{R^2\kappa(1-\kappa)(1-2\kappa)^2} \\ &\text{and three similar expressions.} \\ I_{\kappa\lambda} &= \frac{-n(1-2\mu)(1-2\nu)}{R^2} \\ &\text{with five similar expressions.} \end{aligned}\right\} \quad . \quad . \quad . \quad (6.3)$$

Now if θ is a column vector representing four approximate values for the recombination fractions ; if S is the corresponding set of scores ; and if I is the matrix of the amounts of partial information ; then an improved set of approximations to the maximum likelihood estimates is given by θ_1, where

$$\theta_1 = \theta + I^{-1}S \quad . \quad . \quad . \quad . \quad (6.4)$$

If necessary, the process can be repeated using the set of approximations θ_1 as starting point, to give new estimates θ_2. More than two cycles of iteration are rarely needed.

From a consideration of table 5 it is easy to see that we could obtain initial estimates of the recombination fractions by using product formulæ of the type

$$\left(\frac{\kappa}{1-\kappa}\right)^4 = \frac{m_1m_5m_6m_7}{m_2m_3m_4m_8} = \frac{aefg}{bcdh} \text{ etc.}$$

but these would be very sensitive to sampling variations in the small numbers of triple-recombinants. We can avoid this difficulty by grouping together the four triple-recombination classes, and then solving the equations formed by setting the expected values equal to the observed values—this gives four independent equations in four

Suppose that the arrangement is as in the following diagram, where the recombination fractions for the various segments are also shown :

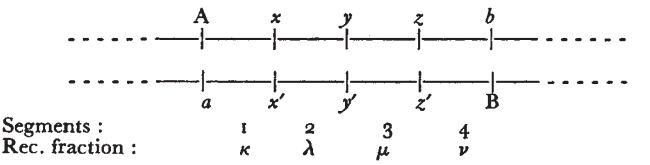

FIG. 3.

The observations are distributed in eight groups : four single-recombination groups and four triple-recombination groups. If the recombination fractions are not too large then it will be easy to distinguish the single from the triple-recombination classes, thus establishing the order of the loci. Let us suppose that this can be done.

Now the straightforward approach is to use the method of maximum likelihood, but since we have seven degrees of freedom to estimate four parameters we shall not be able to replace the likelihood equations by equations in which expectations are set equal to the observations. We can, however, employ the usual scoring procedure. It is convenient to have approximate estimates of the recombination fractions to start with and an easy method of obtaining such approximations is given below.

Suppose that the observed and expected values are, on the assumption of no interference, as set out in the following table :

TABLE 5

	Prototrophic type	Recombination over	Expected	Observed
Singles	AB, $x'y'z'$	1	$m_1 = n\kappa(1-\lambda)(1-\mu)(1-\nu)/R$	a
	$xy'z'$	2	$m_2 = n(1-\kappa)\lambda(1-\mu)(1-\nu)/R$	b
	xyz'	3	$m_3 = n(1-\kappa)(1-\lambda)\mu(1-\nu)/R$	c
	xyz	4	$m_4 = n(1-\kappa)(1-\lambda)(1-\mu)\nu/R$	d
Triples	AB, $x'yz'$	123	$m_5 = n\kappa\lambda\mu(1-\nu)/R$	e
	$x'yz$	124	$m_6 = n\kappa\lambda(1-\mu)\nu/R$	f
	$x'y'z$	134	$m_7 = n\kappa(1-\lambda)\mu\nu/R$	g
	$xy'z$	234	$m_8 = n(1-\kappa)\lambda\mu\nu/R$	h
		Total	n	n

where R, the recombination fraction over the whole interval from A to B, is given by :

$$\left.\begin{aligned} R &= \Sigma\kappa - 2\Sigma\kappa\lambda + 4\Sigma\kappa\lambda\mu - 8\kappa\lambda\mu\nu \\ &= \tfrac{1}{2}[1-(1-2\kappa)(1-2\lambda)(1-2\mu)(1-2\nu)] \\ \text{and } j &= e+f+g+h. \end{aligned}\right\} \quad . \qquad (6.1)$$

unknowns, and is equivalent to a maximum likelihood solution with the triple-recombinants grouped. The estimates are given by :

$$\left.\begin{aligned} \kappa = \frac{a}{a+F},\ \lambda = \frac{b}{b+F},\ \text{etc.} \\ \text{where } F^2 = \frac{abcd}{j}\left\{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right\} \\ \text{and } j = e+f+g+h \end{aligned}\right\} \quad . \quad . \quad (6.5)$$

As an example, let us consider the following data for *Escherichia coli*, for the use of which I am indebted to Dr L. L. Cavalli.

The crosses made were :

$$M^+Gal^-Lac^-V_1{}^r(T^-L^-) = M^-Gal^+Lac^+V_1{}^s(T^+L^+) \quad . \quad . \quad (6.6)$$

where in these experiments (TL) acted as a single unit. The strains crossed were K12 : W583 and 58-161.

The prototrophs recovered were all $M^+(T^+L^+)$ and were scored for *Gal*, *Lac*, and V_1. It appeared from the data that the order of the loci was as shown in (6.6) and on this assumption we may exhibit the data as in table 5*a* :

TABLE 5A

Prototrophic type	Recombination over	Observed number	Expected number
$M^+(T^+L^+)$,			
Gal^+ Lac^+ $V_1{}^s$	1	a : 44	48·05
Gal^- Lac^+ $V_1{}^s$	2	b : 82	85·39
Gal^- Lac^- $V_1{}^s$	3	c : 191	187·63
Gal^- Lac^- $V_1{}^r$	4	d : 99	95·18
$M^+(T^+L^+)$,			
Gal^+ Lac^- $V_1{}^s$	123	e : 11	6·68
Gal^+ Lac^- $V_1{}^r$	124	f : 7	3·39
Gal^+ Lac^+ $V_1{}^r$	134	g : 4	7·45
Gal^- Lac^+ $V_1{}^r$	234	h : 9	13·24
		31 : j	30·76
	Total	n : 447	447·01

Using the formulæ of (6.5) we obtain the preliminary estimates :

$$\theta = \begin{bmatrix} \kappa \\ \lambda \\ \mu \\ \nu \end{bmatrix} = \begin{bmatrix} 0{\cdot}12 \\ 0{\cdot}20 \\ 0{\cdot}36 \\ 0{\cdot}23 \end{bmatrix} \quad . \quad . \quad . \quad (6.7)$$

Substituting these values in (6.2) gives the scores

$$S = \begin{bmatrix} S_\kappa \\ S_\lambda \\ S_\mu \\ S_\nu \end{bmatrix} = \begin{bmatrix} 24{\cdot}0895 \\ 4{\cdot}7558 \\ -17{\cdot}5867 \\ -39{\cdot}4096 \end{bmatrix} \quad . \quad . \quad (6.8)$$

And substituting in (6.3) gives the information matrix :

$$I = \begin{Vmatrix} 5167{\cdot}9 & -332{\cdot}8 & -713{\cdot}1 & -369{\cdot}8 \\ -332{\cdot}8 & 3393{\cdot}7 & -903{\cdot}3 & -468{\cdot}4 \\ -713{\cdot}1 & -903{\cdot}3 & 2243{\cdot}5 & -1003{\cdot}7 \\ -369{\cdot}8 & -468{\cdot}4 & -1003{\cdot}7 & 3056{\cdot}4 \end{Vmatrix} . \quad . \quad (6.9)$$

Inverting (6.9) gives :

$$I^{-1} = 10^{-4} \times \begin{Vmatrix} 2{\cdot}22463 & 0{\cdot}69097 & 1{\cdot}35168 & 0{\cdot}81892 \\ 0{\cdot}69097 & 3{\cdot}86689 & 2{\cdot}43714 & 1{\cdot}47657 \\ 1{\cdot}35168 & 2{\cdot}43714 & 7{\cdot}16048 & 2{\cdot}88851 \\ 0{\cdot}81892 & 1{\cdot}47657 & 2{\cdot}88851 & 4{\cdot}54578 \end{Vmatrix} . \quad (6.10)$$

Therefore, using (6.4), the corrections to apply to the estimates in (6.7) are :

$$I^{-1}S = \begin{Vmatrix} 0{\cdot}000 \\ -0{\cdot}007 \\ -0{\cdot}020 \\ -0{\cdot}020 \end{Vmatrix} . \quad . \quad . \quad (6.11)$$

Therefore

$$\theta_1 = \begin{Vmatrix} 0{\cdot}120 \\ 0{\cdot}193 \\ 0{\cdot}340 \\ 0{\cdot}210 \end{Vmatrix} . \quad . \quad . \quad . \quad (6.12)$$

We can use the improved estimates of (6.12) to recalculate the scores and hence derive a better set of approximations, θ_2. It is not worthwhile to recalculate the information matrix as this changes relatively slowly. The standard errors of the final estimates are given by the square roots of the terms in the leading diagonal of (6.10). Thus we have :

$$\left.\begin{aligned} \hat{\kappa} &= 12{\cdot}4 \pm 1{\cdot}5 \text{ per cent.} \\ \hat{\lambda} &= 20{\cdot}1 \pm 2{\cdot}0 \quad \text{,,} \\ \hat{\mu} &= 35{\cdot}6 \pm 2{\cdot}7 \quad \text{,,} \\ \hat{\nu} &= 21{\cdot}9 \pm 2{\cdot}1 \quad \text{,,} \end{aligned}\right\} . \quad . \quad (6.13)$$

The five loci can now be mapped as in fig. 4. Recombination fractions are given for the separate segments, and also a map distance calculated from Haldane's formula : $x = -\frac{1}{2}\log_e(1-2y)$, where y is the recombination fraction and x the map distance, in cM.

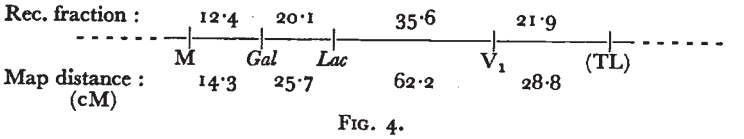

FIG. 4.

Since the observations have 7 degrees of freedom and we have estimated only 4 parameters, there are 3 degrees of freedom remaining with which to test how well the hypotheses fit the data. The expected numbers are shown in table 5*a*. The chief source of variation is

between the triple-recombination classes. As the numbers in classes (124) and (134) are somewhat small, let us group these together before calculating a χ^2 with 2 degrees of freedom. We find $\chi_2^2 = 4{\cdot}84$, and P = 8·9 per cent. We may regard this as a sufficiently good fit.

The total map distance between M and (TL) works out at 131·0 cM. It is of some interest to find the standard error of this quantity. It will not be very small because the estimates for the individual segments are positively correlated.

Suppose that the total map distance is X, and that the values for the individual segments are x_1, x_2, x_3 and x_4. Then :

$$X = x_1+x_2+x_3+x_4 = -\tfrac{1}{2}\Sigma\log(1-2\kappa) = -\tfrac{1}{2}\log(1-2R)$$

Therefore $$\frac{\partial X}{\partial \kappa} = \frac{1}{1-2\kappa} \text{ etc.}$$

Thus, in large samples,

$$\begin{aligned} \text{var } X &= \Sigma \text{var } x_1 + 2\Sigma \text{cov}(x_1 x_2) \\ &= \Sigma\left\{\frac{\partial X}{\partial \kappa}\right\}^2 \text{var } \kappa + 2\Sigma\left\{\frac{\partial X}{\partial \kappa}\right\}\left\{\frac{\partial X}{\partial \lambda}\right\}\text{cov }(\kappa,\lambda) \end{aligned}$$

Therefore $$\text{var } X = l'.I^{-1}.l \quad . \quad . \quad . \quad . \quad . \quad . \quad (6.14)$$

where $$l = \begin{Bmatrix} 1/(1-2\kappa) \\ 1/(1-2\lambda) \\ 1/(1-2\mu) \\ 1/(1-2\nu) \end{Bmatrix} \quad . \quad . \quad . \quad . \quad . \quad (6.15)$$

Using the estimates in (6.13) we have $$l = \begin{Bmatrix} 1{\cdot}3298 \\ 1{\cdot}6722 \\ 3{\cdot}4722 \\ 1{\cdot}7794 \end{Bmatrix} \quad . \quad . \quad (6.16)$$

Substituting (6.10) and (6.16) in (6.14) gives :

$$\left.\begin{aligned} \text{var } X &= 207{\cdot}68 \\ \text{therefore} \quad X &= 131{\cdot}0 \pm 14{\cdot}4 \text{ cM} \end{aligned}\right\} \quad . \quad . \quad (6.17)$$

A similar analysis can, of course, be carried out with data from a four-point cross.

The recombination fractions shown in fig. 4 are appreciably higher than the corresponding values obtained by Lederberg. However, Lederberg's estimates do not take account of differential viabilities by which they are almost certainly influenced, while in the above treatment of Cavalli's five-point material it was not possible to allow for this kind of disturbance. Thus we are at present unable to decide whether the two bodies of data are in agreement or not.

7. SUMMARY

With bacteria linkage data is obtained from the crossing of complementary types by classifying organisms, for which recombination over an odd number of segments has been obligatory. If the loci,

for which both members of a contrasting pair of characters are scorable, are distal to the segment over which recombination is compulsory, then, on the assumption of no interference, the data may be analysed by the methods already used in experimental linkage work.

If, on the other hand, these loci lie within the segment, then special methods of analysis are required. The present paper develops the appropriate maximum likelihood treatment of a selection of such problems arising in three, four and five-point crosses, and gives some worked examples.

REFERENCES

BAILEY, NORMAN T. J. 1949. A method of allowing for differential viability in estimating linkage from backcross matings in coupling only or repulsion only. *Heredity*, *3*, 225-228.

FISHER, R. A. 1935-49. *The Design of Experiments*. Edinburgh : Oliver and Boyd.

LEDERBERG, JOSHUA. 1947. Gene recombination and linked segregations in *Escherichia coli*. *Genetics*, *32*, 505-525.

LEDERBERG, J., AND TATUM, E. L. 1946*a*. Detection of biochemical mutants of micro-organisms. *J. Biol. Chem.*, *165*, 381-382.

LEDERBERG, J., AND TATUM, E. L. 1946*b*. Novel genotypes in mixed cultures of biochemical mutants of bacteria. *Cold Spring Harbor, Symp. Quant. Biol.*, *11*, 113-114.

LEDERBERG, J., AND TATUM, E. L. 1946*c*. Gene recombination in *Escherichia coli*. *Nature*, *158*, 558.

TATUM, E. L., AND LEDERBERG, J. 1947. Gene recombination in the bacterium, *Escherichia coli*. *J. Bact.*, *53*, 673-684.