STATISTICAL ANALYSIS OF CHROMOSOME PAIRING IN INTERSPECIFIC HYBRIDS. I. THE PROBABILITY DISTRIBUTIONS¹

A. G. SFICAS²

North Carolina State College, Raleigh, North Carolina

Received May 8, 1962

THE intensive breeding programs for the development of new varieties of economic importance have forced plant breeders to look at the wild species as an additional source of desirable characters. The success of attempts to recombine desirable genes or chromosome segments of the wild species with those of the cultigen depends largely on pairing relationships of their chromosomes (GERSTEL 1948). Pairing of chromosomes in interspecific hybrids is also of great interest to the students of evolution. If no asynaptic genes are involved, pairing of chromosomes can be considered a measure of chromosome homology reflecting the degree of species relationships and their ancestral origin (GOODSPEED 1954). Hybrids which show a variable amount of pairing at MI of meiosis, and in which pairing chromosomes cannot be identified individually, pose several questions: Can all or only some of the chromosomes pair? Do these chromosomes pair in specific or random combinations and with the same or different probabilities of pairing? The effect of genetical and environmental factors on pairing distributions must also be considered. GAUL (1954) developed the following formula for determining the number of chromosomes p which are able to pair:

$$p = \frac{X^2 + X - B}{(2X - B)Z}$$

where X , B and Z are the total observed numbers of chiasmata, of paired chromosomes and the number of cells in the particular preparation, respectively. However, if the number of chiasmata per bivalent cannot be recognized, GAUL's formula is not applicable. The present paper will attempt to develop statistical distributions which are independent of the number of chiasmata per bivalent. Applications to hybrids of *Nicotiana* will be presented in the following paper.

RESULTS AND DISCUSSION

A probability distribution of the number of bivalents: Let p_i ($i = 1, 2, \dots, n$) be the probability that the i th chromosome of one species will pair with the same chromosome of the other species in the hybrid. It is assumed that only one pairing possibility exists for each chromosome in the species. Let $q_i = 1 - p_i$ be the

¹ Journal Series Paper No. 1352 of the North Carolina Agricultural Experiment Station. Part of a dissertation submitted to the Graduate Faculty of North Carolina State College for the degree of Doctor of Philosophy.

² Present address: Tobacco Institute, Drama, Greece.

probability that the i th chromosome will not pair with its partner. Since in the experiments of the following papers $\Sigma p_i < \Sigma q_i$ (low pairing) the derivation of the equations will be in terms of q_i to facilitate the computations and the solutions can be transformed back into p 's. The following situations arise with the above problem:

1. If all p 's are equal the distribution will be the binomial $(p + q)^n$ where n equals the number of maximum possible pairs in a cell and at a time. HALDANE (1941) discussed the fitting of the binomial law to the observed data by the method of moments. The number of trials (n) and the probability of success in a trial (p) are estimated from the mean (m) and the variance (v) by the following relations:

$$\hat{n} = \frac{m^2}{m - v} \text{ and } \hat{p} = \frac{m - v}{m} \quad (1)$$

2. An extension of the above formulas will be developed here. If there are two groups of possible bivalents n_1 and n_2 , and the probability of pairing in the first group is p_1 and for the second group is p_2 , where $p_1 \neq p_2$, the probability distribution function will be the product of two binomials

$$(p_1 + q_1)^{n_1} (p_2 + q_2)^{n_2} \quad (2)$$

If n is the total number of possible bivalents, the following relations are obtained by use of the mean and the variance:

$$\begin{aligned} n_1 p_1 + n_2 p_2 &= m \\ n_1 p_1 q_1 + n_2 p_2 q_2 &= v \end{aligned} \quad (3)$$

These equations can be solved for different values of n_1 and n_2 , with the restriction $n_1 + n_2 = n$. There are two sets of solutions (except for the case $n_1 = n_2$ where the two sets are identical) as follows:

$$p_2 = \frac{m}{n} \pm \sqrt{\frac{m^2}{n^2} - \frac{n_1}{nn_2} \left(\frac{m^2}{n_1} - m + v \right)} \quad (4)$$

and

$$p_1 = \frac{m - n_2 p_2}{n_1}$$

To estimate n_1 and n_2 , when their sum is known, the third moment about the origin should be calculated and equated to the expected one. The moment generating function of the above distribution is

$$m(t) = (p_1 e^{it} + q_1)^{n_1} (p_2 e^{it} + q_2)^{n_2}, \quad i^2 = -1 \quad (5)$$

and the expected third moment about the origin is

$$\begin{aligned} \left[\frac{\partial^3 m(t)}{\partial t^3} \right]_{t=0} &= n_1(n_1 - 1)(n_1 - 2)p_1^3 + 3n_1(n_1 - 1)p_1^2 + n_1 p_1 + \\ &\quad n_2(n_2 - 1)(n_2 - 2)p_2^3 + 3n_2(n_2 - 1)p_2^2 + n_2 p_2 + \\ &\quad 3n_1 n_2 p_1 p_2 [2 + (n_1 - 1)p_1 + (n_2 - 1)p_2] \end{aligned} \quad (6)$$

There are two ways to find the best solution for n_1 and n_2 (where $n_1 + n_2 = n$ known), both being equivalent, except probably for the amount of work required. The first way is simply to calculate all possible solutions for p_1 and p_2 from (4) for the different combinations of n_1 and n_2 for which a chi-square test is performed. The solution which yields the minimum chi-square is the best one.

The second way makes use of equation (6) from which integral positive values

of n_1 and n_2 with their respective probabilities p_1 and p_2 (where $0 \leq p_1, p_2 \leq 1$) and expected third moments are calculated. The solution which will give the third moment about the origin closer to the calculated one from the data will be the best solution. An exact solution for n_1 and n_2 can be obtained by substituting p_1 and p_2 of (4) and $n_1 = n - n_2$, which yields a cubic equation in n , but which is difficult to handle. The above suggested way will give the same solution, because integral values of n_1 and n_2 will be finally accepted.

3. The same argument as in the previous case can be followed for a further subdivision of n into more than two groups. The number of solutions increases and the computations become very difficult when the groups are more than two. All solutions will be identical for groups of equal numbers of possible pairs.

4. If each chromosome has a different probability of pairing, the bivalent distribution is generated by the expansion of the products

$$(p_1 + q_1)(p_2 + q_2) \dots (p_n + q_n) \quad (7)$$

where $p_i + q_i = 1$ and $i = 1, 2, \dots, n$. This is the general case of Bernoulli trials with unequal probabilities.

In expanding the above product a probability distribution is generated which corresponds to the number of bivalents formed in a cell. Let $B_0, B_1, \dots, B_n$ be the probabilities that 0, 1, $\dots, n$ pairs will be observed in a cell. This corresponds to the probability of a number of successes in n trials, where each trial has not necessarily the same probability of success. For simplicity in the expression of the distribution, let $Q_{(n-m)}^n P_m^n$ denote the product of $n-m$ different q 's times the product of m different p 's, which are also a different set from the q 's and let $\Sigma Q_{(n-m)}^n P_m^n$ be the sum over all such sets of p 's and q 's. The probability distribution will expand as follows:

$$\begin{aligned} B_0 &= \Sigma Q_n^n P_0^n = \Pi q_i \\ B_1 &= \Sigma Q_{(n-1)}^n P_1^n \\ B_m &= \Sigma Q_{(n-m)}^n P_m^n \\ B_n &= \Sigma Q_0^n P_n^n = \Pi p_i \\ \text{Sum } \frac{1}{1} &= \frac{1}{1} \end{aligned} \quad (8)$$

There are $n + 1$ nonlinear equations with n unknown q 's plus one restriction (the sum of either side should equal one), hence, there exist solutions for the system. The equations are of degree n , hence, there are n roots for any q . Since there are no particular coefficients for the different unknowns, but all products have the same coefficients and involve all unknowns precisely the same way, each of the q 's will have the same roots as any other q , hence, the n combinations of solutions are essentially the same, if order is not considered. Hence, there is only one general solution for the system. Positive roots with values between 0 and 1 are acceptable. Because of the high degree of the equations and the extremely large number of products involved, the solution of the system in its present form becomes practically impossible even for a relatively small number of unknowns. A method of transformation of the system is presented in the next section which simplifies its solution.

A method of transformation and simplification of the distribution: The method

will be presented in an example of five possible pairs with probabilities p_1, p_2, p_3, p_4 and p_5 that the specific pair will be formed and q_1, q_2, q_3, q_4 and q_5 that it will not be formed. Using (8) with $n = 5$ the probability distribution is as follows:

$$\begin{aligned} B_0 &= \Pi q_i = q_1 q_2 q_3 q_4 q_5 \\ B_1 &= \Sigma Q_1^5 P_1^5 = q_1 q_2 q_3 q_4 p_5 + q_1 q_2 q_3 p_4 p_5 + \dots \\ B_2 &= \Sigma Q_2^5 P_2^5 = q_1 q_2 q_3 p_4 p_5 + q_1 q_2 p_3 p_4 p_5 + \dots \\ B_3 &= \Sigma Q_3^5 P_3^5 = q_1 q_2 p_3 p_4 p_5 + q_1 p_2 p_3 p_4 p_5 + \dots \\ B_4 &= \Sigma Q_4^5 P_4^5 = q_1 p_2 p_3 p_4 p_5 + p_1 q_2 p_3 p_4 p_5 + \dots \\ B_5 &= \Pi p_i = p_1 p_2 p_3 p_4 p_5 \end{aligned} \quad (9)$$

If $1 - q_i$ is substituted for p_i in (9) and let:

$$B = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -5 & 1 & 0 & 0 & 0 & 0 \\ 10 & -4 & 1 & 0 & 0 & 0 \\ -10 & 6 & -3 & 1 & 0 & 0 \\ 5 & -4 & 3 & -2 & 1 & 0 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix},$$

$$Q = \begin{bmatrix} \Pi q_i \\ \Sigma Q_1^5 \\ \Sigma Q_2^5 \\ \Sigma Q_3^5 \\ \Sigma Q_4^5 \\ 1 \end{bmatrix}, \quad \text{and } A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 1 & 0 & 0 & 0 & 0 \\ 10 & 4 & 1 & 0 & 0 & 0 \\ 10 & 6 & 3 & 1 & 0 & 0 \\ 5 & 4 & 3 & 2 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

then the relation (9) is equivalent to:

$$B = AQ \quad (10)$$

and solving for Q :

$$Q = A^{-1}B = T = \begin{bmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{bmatrix} \quad (11)$$

where T will be the new set of constant terms in the equations derived from the observations B of the experiment.

The above system of equations has the solution of the polynomial:

$$q^5 - T_4 q^4 + T_3 q^3 - T_2 q^2 + T_1 q - T_0 = 0 \quad (12)$$

the coefficients of which are the T 's of the column vector (11). This is proved as follows: Let q_i be the i th root of the polynomial

$$q^5 - C_4 q^4 + C_3 q^3 - C_2 q^2 + C_1 q - C_0 = 0 \quad (13)$$

The polynomial (13) can be factored as follows:

$$\prod_{i=1}^5 (q - q_i) = 0 \quad (14)$$

by expanding (14) we obtain the coefficients of q^i ($i = 0, 1, \dots, 5$) which are exactly the elements of the column vector Q which in turn are exactly equal to the corresponding elements of T in equations (11). Hence, $T_i = C_i$ ($i = 1, 2, \dots, 5$) and the system of equations (11) has the unique solution of polynomial (12).

If n q 's are involved in a similar system, the system can be reduced to the polynomial

$$q^n - T_{n-1}q^{n-1} + T_{n-2}q^{n-2} \dots = 0 \quad (15)$$

which has n roots corresponding to the solution of the system (8).

The estimation of the parameters: Matrix techniques can be used to get information on the roots of the polynomial (15) (See BROWNE 1958). The graphing of the polynomial can be used for the same purpose. The roots could be found simply by trial and error. A difficulty with this method is that a large number of significant figures must be carried in the calculations because small rounding errors at the beginning have a large multiplicative effect on the successive solutions. This effect can be reduced when the roots of smaller magnitude are first obtained.

A method to locate the roots of the polynomial especially if there is reason to believe that they constitute clusters or groups whose members differ slightly but the difference among groups is appreciable, can be applied by making use of Descartes rule of signs. According to this rule a polynomial with all its roots real will have as many positive roots as are variations in sign, and the remaining roots are negative. If we make the substitution $q = \gamma + h$ in the polynomial where h would be any value in the range of values of roots then the new polynomial in γ will have the roots of q minus the quantity h . By using the rule of signs we find how many roots become negative in the new polynomial, and these roots will be in the range from zero to h . By varying the arbitrary value h the interval in which different number of roots are located can be found. For example in the second substitution $q = \gamma + h'$ where $h' = h + k$, the difference in negative roots found in the new polynomial from what were found in the polynomial with the substitution $q = \gamma + h$, would give the number of roots in the interval from h to $h' = h + k$ and so on.

A scheme will be given which can be applied for the substitution for any value of h and any degree of the polynomial. For a polynomial $q^3 - T_2q^2 + T_1q - T_0 = 0$, the new polynomial in γ will be

$$\gamma^3(1) + \gamma^2(3h - T_2) + \gamma(3h^2 - 2T_2h + T_1) + (h^3 - T_2h^2 + T_1h - T_0) \quad (16)$$

The numerical coefficients in the parentheses are (except for sign) the elements of the matrix A^{-1} which has been used in the transformation. Hence, the coefficients in the new polynomial in γ will be polynomials in h , having the obvious pattern shown in the example above with coefficients from the matrix A^{-1} and the values of the transformation matrix T , and it can be extended for any number of n and can be used for any value of h . This method can also be used to find the roots by narrowing the interval where they belong. By finding each root individually and independently of the others, some of the pitfalls of the successive solution can be avoided.

The probability distribution of (7) has mean and variance equal to $n_1p_1 + n_2p_2 + \dots = n\bar{p}$, and $n\bar{p}q - \Sigma(p_i - \bar{p})^2 = \Sigma p_i q_i$, respectively (RIETZ 1927). When the p 's are small or the differences among the p 's are small, or both, the variance of the distribution approaches the variance of the binomial.

The solution obtained in the previous section may fit the data completely. But

it is possible for small differences in the estimated parameters to be due to sampling errors. Some hypotheses can be tested concerning equality of groups of probabilities. One hypothesis would be to consider all probabilities equal and test the goodness of fit of the binomial. Also group averages can be tested as in (2), for example.

FELLER (1957, p. 264) shows that the case of Bernoulli trials with unequal probabilities can be approximated by the Poisson distribution when n increases and the probabilities depend on n in such a way that the largest tends to zero, but the sum $p_1 + p_2 + \dots + p_k = m$ remains constant. In the case of the distribution of the number of bivalents per cell, m can be estimated by the mean number of bivalents. The number n can be large when many homologous segments are scattered in the chromosomes of the different genomes and this gives a large number of possible pair combinations with a small probability for a specific bivalent to be formed, so that a Poisson distribution might arise.

Statistical treatment of other variation: Let there be k locations sampled in a single anther. Let p_t be the probability of pairing in the t th location and let it be constant for all n possible bivalents in a cell and for all cells within the t th location, but different from location to location of that anther. If we sample one cell or n potential chromosome pairs from each location, the expected total number of bivalents is $np_1 + np_2 + \dots + np_k = kn\bar{p}$ where $\bar{p} = (1/k)(p_1 + p_2 + \dots + p_k)$. Therefore, the expected number of bivalents in a cell from a random location is $n\bar{p}$, which is equal to the expected number of bivalents in a cell when the probability of pairing is constant for all locations and is equal to $\bar{p}$.

The expected value of the square of the standard deviation of a number of possible pairs in a cell is $np_t q_t$. If the deviations were measured from $n\bar{p}$ instead of np_t , the mean square of the deviations would be (RIETZ 1927) $np_t q_t + (np_t - n\bar{p})^2$. If N cells were taken in which the probability p_t is the same then

$$Nnp_t q_t + Nn^2(p_t - \bar{p})^2 \quad (17)$$

would be the expected value of the sum of squares of the deviations from $n\bar{p}$ in the N cells taken from the same location of the anther. By adding together the expression (17) for $t = 1, 2, \dots, k$ we have

$$Nn \sum_{t=1}^k p_t q_t + Nn^2 \sum_{t=1}^k (p_t - \bar{p})^2 \quad (18)$$

for the expected sum of squares from $n\bar{p}$ (mean bivalents per cell) from the k locations.

The mean square deviation from $n\bar{p}$ of the number of bivalents, n being possible in a cell taken one cell from each of the k locations, is obtained by dividing (18) by the number of sets Nk , which gives

$$\sigma_L^2 = \frac{n}{k} \sum_{t=1}^k p_t q_t + \frac{n^2}{k} \sum_{t=1}^k (p_t - \bar{p})^2 \quad (19)$$

From (18) and (19), and $\sigma_B^2 = n\bar{p}q$ (the variance of the binomial)

$$\sigma_L^2 = \sigma_B^2 + \frac{n^2 - n}{k} \sum_{t=1}^k (p_t - \bar{p})^2 \quad (20)$$

Hence, the above distribution, called Lexies distribution (RIETZ 1927) has a

variance (σ_L^2) greater than that of a binomial based on a constant probability

$$\bar{p} = \frac{1}{k}(p_1 + p_2 + \dots + p_k).$$

Probability of pairing (p) is likely to vary with the environment. Therefore, a nonuniform sample would give a higher variance; hence an upwards binomial estimate for n will be obtained in (1).

FISHER (1941) discusses the problem of nonuniformity of the sampled population in another way. He points out that the Poisson distribution arises when equal samples are taken from perfectly homogeneous material, which is completely determined by the mean (m). When unequal samples are taken, or the material is not perfectly homogeneous, the value of m would vary from sample to sample and a negative binomial distribution arises. The variance of m always increases the variance of the variable x . Since the binomial has as a limit the Poisson distribution when p is small and n is large, the Lexies distribution will have as a limit the negative binomial distribution under the same conditions.

Another approach to the generation of the negative binomial is discussed by JAIN (1959). If the probabilities of 1, 2, . . . , n bivalents to form in a cell are not homogeneous, a transition function is obtained, similar to the one obtained by JACKSON and BARBER (1958) for representing the patterns of chromosome breakage in aged and irradiated material. This function gives rise to a typical model of negative binomial. He interprets it in terms of contagious distribution or a type of interaction among the different bivalents formed in a cell.

SUMMARY

In interspecific hybrids it is often not possible to recognize which chromosomes are involved in the pairs which are formed. A question arises whether this pairing involves all the chromosomes or only some of them, and whether these chromosomes pair in specific or random combinations and with the same or different probabilities of pairing.

If particular chromosomes of one genome can pair only with particular and always the same chromosomes of the other genome, a probability distribution is generated by extending the binomial concept for the case in which pairing probabilities are not necessarily the same. This probability distribution can also be approximated when other pairing possibilities are negligible, which can be justified from observations of the degree of pairing in the haploid species and/or the frequency of trivalents and multivalents in the hybrid.

The probability distribution with unequal probabilities is the expansion of the product $(p_1 + q_1)(p_2 + q_2) \dots (p_n + q_n) = B_0 + B_1 + B_2 + \dots + B_n$ where p_i ($i = 1, 2, \dots, n$) is the probability of a specific pair to occur in a cell, $q_i = 1 - p_i$ and B_i ($i = 0, 1, \dots, n$) is the probability of observing i bivalents in a cell estimated from the data. The solution for p_i or q_i became possible by the development of a matrix transformation which reduces the problem to the finding of the roots of the polynomial $q^n - T_{n-1}q^{n-1} + T_{n-2}q^{n-2} - \dots = 0$, where T_i ($i = 0, 1, 2, \dots, n-1$) are the transformed B 's. An estimation procedure is discussed together with the test of different hypotheses. If the probabilities are small as n

increases while their sum is constant the above distribution will approximate a Poisson distribution. When the sampled population is not homogeneous, the variance becomes greater than that of the binomial and can be greater than the mean, and a Lexies series which has as a limit the negative binomial distribution is generated.

Applications to specific examples of interspecific hybrids will be discussed in following papers of this series.

ACKNOWLEDGMENTS

The author is indebted to PROFESSORS D. U. GERSTEL and C. C. COCKERHAM for review and suggestions concerning the preparation of the manuscript. He owes gratitude to the National Tobacco Board of Greece for a scholarship which enabled him to work in North Carolina.

LITERATURE CITED

- BROWNE, E. T., 1958 *Introduction to the Theory of Determinants and Matrices*. The University of N. C. Press. Chapel Hill, N. C.
- FELLER, W., 1957 *An Introduction to Probability Theory and Its Applications*. Vol. I. Second edition, John Wiley and Sons, Inc. New York, N.Y.
- FISHER, R. A., 1941 The negative binomial distribution. *Ann. Eug.* **11**: 182-187.
- GAUL, H., 1954 Asynapsis und ihre Bedeutung für die Genomanalyse. *Z. Ind. Abst. Vererb.* **86**: 69-100.
- GERSTEL, D. U., 1948 Transfer of the mosaic-resistance factor between H-chromosomes of *N. glutinosa* and *N. tabacum*. *J. Agr. Res.* **76**: 219-223.
- GOODSPEED, T. H., 1954 *The Genus Nicotiana*. Chronica Botanica Co. Waltham, Mass.
- HALDANE, J. B. S., 1941 The fitting of the binomial distributions. *Ann. Eug.* **11**: 179-181.
- JACKSON, W. D., and H. N. BARBER, 1958 Patterns of chromosome breakage after irradiation and aging. *Heredity* **12**: 1-26.
- JAIN, S. K., 1959 Fitting the negative binomial distribution to some data on asynaptic behavior of chromosomes. *Genetica* **30**: 108-122.
- RIETZ, H. L., 1927 *Mathematical Statistics*. Math. Assoc. of America. Open Court Pub. Co. Chicago, Illinois.