Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 4175–4178, June 1989
Genetics

Computer-simulation methods in human linkage analysis

JURG OTT

Columbia University, Departments of Genetics and Development and of Psychiatry, New York, NY 10032; and New York State Psychiatric Institute, New York, NY 10032

Communicated by Charles R. Cantor, March 9, 1989 (received for review December 29, 1988)

ABSTRACT **In human linkage analysis, many statistical problems without analytical solution could be solved by ad hoc Monte Carlo procedures were efficient computer-simulation methods available for members of family pedigrees. In this paper, a general method is described for randomly generating genotypes at one or more marker loci, given observed phenotypes at loci linked among themselves and with the markers. The method is based on a well-known expansion of the multivariate probability of genotypes, given phenotypes, into a product of conditional univariate probabilities that may be viewed as corresponding to conditionally independent univariate random variables. This representation allows a recursive evaluation of the univariate probabilities that can be implemented in a surprisingly simple manner by carrying out successive "risk calculations" with respect to marker genotypes, given observed phenotypes and marker genotypes already generated. Potential applications to various unresolved problems are discussed. The method is applied to 28 published families analyzed for genetic linkage between hereditary motor and sensory neuropathy I and the Duffy (*FY*) blood group locus and confirms heterogeneity of hereditary motor and sensory neuropathy I. An implementation of the simulation methods developed in the LINKAGE program package will be available later in 1989.**

In human linkage analysis, many statistical and operational problems exist that elude analytical solution. For example, significance tests for the presence of linkage or heterogeneity are applied. However, due to the particular data (nonindependent phenotypes seen in the members of family pedigrees), the test statistics must distribute differently in different families. This outcome is seen, for example, in table 3.11 of Ott (1), in which the line for $\theta = 0.5$ demonstrates that the p value associated with a given critical logarithm of odds (lod) score in the test for linkage differs considerably from one family type to another. In tests such as those for heterogeneity, asymptotic properties of the test statistics are often invoked, and test results are called approximate. This situation is unfortunate, as the amount of data, presumably, is rarely so large that asymptotic results are reliable. Computer-simulation (Monte Carlo) methods offer elegant solutions to many of these problems by simulating genetic mechanisms in the actual family data under investigation; the main advantage of these methods is that they can be tailored to a particular problem. The broad spectrum of applications of Monte Carlo methods will be discussed in detail in the *Applications* section.

To focus more specifically, consider a disease locus with a linked marker locus and assume known disease phenotypes but as-yet-unobserved marker phenotypes in the members of a family pedigreee. The problem is to predict with what probability the maximum lod score Z between disease and marker loci will exceed a value c. To calculate this conditional probability, $P(Z > c|\text{data})$, all possible phenotypes at the marker locus have to be evaluated for each individual along with their probabilities of occurrence. Then, the probabilities of those phenotype arrays associated with a lod score exceeding c are added, yielding the answer. However, the number of possible marker phenotype arrays is generally so large that complete enumeration is complex or impossible. The only feasible solution is to take a random sample of marker phenotypes (2). But the sample must be truly random, and this selection represents the main difficulty in such a computer-simulation approach.

Recently, Boehnke (2) introduced a computer-simulation method to estimate in a given set of families the maximum of the expected lod score for a disease versus a marker locus or, alternatively, to predict the probability that the maximum lod score will exceed a given value before marker typing of family members is actually done. To randomly generate (predict) marker genotypes, given disease phenotypes, one may randomly sample from the conditional distribution, $P(\mathbf{g}|\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_n)$ is the vector of *disease* phenotypes of length n, n being the number of family members and $\mathbf{g}$ is the vector of genotypes at the *marker* locus. When the disease is fully penetrant and the path of the disease gene can be tracked through the pedigree, taking a random sample from $P(\mathbf{g}|\mathbf{x})$ is straightforward. Essentially the marker genotypes of children then only depend on the disease and marker genotypes of the parents, so that simulation can efficiently proceed through a pedigree by always moving from parents to children (2).

An extension of Boehnke's (2) simulation method has been proposed by Sandkuyl and Ott (3), in which reduced penetrance at the disease locus is allowed for all individuals without offspring and then this method is applied to the determination of risk distributions in genetic counseling problems before marker data are obtained.

Boehnke's (2) method has been further generalized by allowing for reduced penetrance and sporadic cases at the disease locus (4). This allowance has been achieved, in principle, by a two-step procedure: First, disease genotypes are simulated conditional on the observed disease phenotypes, and then marker genotypes are obtained, given disease genotypes (4).

At this time, to my knowledge, no general method for computer simulation in family pedigrees exists. The presently available methods suffer from major or minor restrictions; for example, none of these methods accommodates known partial typing at the marker locus.

A General Computer-Simulation Method. A very general procedure can be based on the following simple expansion of the conditional probability of genotype vectors. For the moment, retain the definitions introduced above, that is, $\mathbf{x} = (x_1, \ldots, x_n)$ shall denote the vector of disease phenotypes of length n, and $\mathbf{g}$ denotes the vector of genotypes at the marker

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: lod, logarithm of odds.

locus. Then, using basic probability calculus, the following equation can be written:

$$P(\mathbf{g}|\mathbf{x}) = P(g_1|\mathbf{x})P(g_2|g_1, \mathbf{x})P(g_3|g_1, g_2, \mathbf{x})P(g_4|g_1, g_2, g_3, \mathbf{x}) \ldots \quad [1]$$

The interpretation of Eq. 1 is that g_1, g_2, etc. are all conditionally independent random variables, where g_1 depends only on the disease phenotypes, g_2 depends on g_1 and the disease phenotypes, and so on. Consequently, a random vector $\mathbf{g}$ sampled from $P(\mathbf{g}|\mathbf{x})$ may be obtained by successively sampling g_1, g_2, etc. from the univariate distributions on the right side of Eq. 1, the ith element of which depends on the $i-1$ marker phenotypes previously obtained and on the disease phenotypes.

Specifically, to sample a marker genotype for the ith individual, the conditional genotype distribution, $P(g_i|g_{i-1}, g_{i-2}, \ldots, \mathbf{x})$ needs to be evaluated. This process involves making a "risk calculation" with respect to the marker genotypes for the ith individual, given marker genotypes for the $i-1$ previously sampled individuals and given disease phenotypes for all individuals. Such risk calculations are routinely done by standard multipoint linkage programs, for example, by the Mlink program of the LINKAGE program package (5). Because one risk calculation (an array of conditional probabilities) is necessary for each individual, the proposed random generation of marker genotypes, given disease phenotypes, would be more time-consuming on a computer than the original method by Boehnke (2). However, from the limited experience obtained so far, it is not usually generation of the random elements that produces the bottleneck in Monte Carlo procedures, but rather the subsequent analysis of the random elements.

One round of generating marker genotypes for each individual leads to one randomly generated genotype vector $\mathbf{g}$. Repeating the whole process m times, therefore, leads to m random replicates of $\mathbf{g}$—that is, m replicates of family pedigrees with disease phenotypes exactly the same in each replicate (as observed) but randomly different marker genotypes. Each of the m replicates may then be subjected to the same analysis as was the original set of family pedigrees, which, in fact, represents a sample of size 1 drawn from the same universe as the m replicates.

The derivation of the procedure described above did not require specifying a particular order for family members or a particular penetrance structure at disease or marker loci. Indeed, the procedure is completely general in these respects. The procedure will even accommodate situations more general than the one just described; by redefining the vector $\mathbf{x}$ as representing phenotypes at the disease and possible marker loci also, one may apply this procedure to families in which some individuals have been typed at the marker locus, whereas others have not been typed. Also, $\mathbf{x}$ may represent phenotypes at several loci and may even comprise observations on concomitant (biological) variables that, together with disease status, form a multidimensional disease phenotype. Furthermore, it is clear that this procedure will generate genotypes at several linked markers by iteratively generating genotypes at one locus after another, the genotypes previously generated taken as given observations.

The algorithm developed above will be implemented in the LINKAGE program package (5) and made available at no cost to interested researchers toward the end of 1989.

Applications. In principle, Monte Carlo procedures as the one described above may be applied to two different situations: (*i*) to predict genotypes or risks before collecting the relevant information and (*ii*) to evaluate and assess the significance of information collected.

Prediction. Before marker typing is done, as proposed by Boehnke (2) and Ploughman and Boehnke (4), a prediction of the potential for linkage (between the disease gene segregating in a set of families and a marker locus at a certain distance from that disease locus) is desirable. Therefore, the genotype vectors for the marker locus may be randomly sampled to yield a number m of replicates of sets of families; each replicate consists of exactly the same structure of families with identical disease phenotypes but with randomly different marker genotypes. Each replicate could now be analyzed as though it were an observed set of families; for example, the fraction of replicates in which the maximum lod score Z exceeds a certain constant, such as $c = 2$, can be determined. This fraction represents the predicted probability that the lod score will exceed c when marker data is collected. Collecting marker data yields a sample of size 1 from the conditional distribution $P(\mathbf{g}|\mathbf{x})$ so that confidence bounds must be reported in addition to the estimate of probability $P(Z > c)$. Incidentally, computing an average of the logarithm likelihood over replicates allows a simple test of the simulation procedure: As the maximum of the expected logarithm likelihood occurs at the true value θ of the recombination fraction (6), the maximum of the averaged logarithm likelihood curves over replicates should be close to θ—that is, within two or three SDs from θ, where the SD may be obtained numerically from the curvature of the average lod score.

As another example, imagine a disease locus for which flanking marker loci have been found. These markers allow the position of the disease locus to be estimated on the gene map with a certain accuracy and confidence. The question then arises whether additional markers can increase the lod score thus far obtained or whether the presently collected marker data are already almost fully informative. Although an approximate solution to this problem has been offered by Boehnke (2), the procedure described here allows a completely general treatment of such questions, which is particularly important in situations of rare Mendelian diseases where only a small number of families can be collected.

Evaluation. In linkage analysis, various statistical tests are applied, for example, to determine whether a new locus belongs to a linkage group of a known map of marker loci. In two-point linkage analysis, significance is declared when the maximum lod score exceeds 3. However, as outlined in the introduction, depending on family types analyzed, this criterion is associated with very different significance levels, so that its interpretation cannot be the same in different situations. By use of a general Monte Carlo procedure, it is easy to generate an approximation to the null distribution (under H_0: $\theta = 1/2$) of the maximum lod score, which furnishes the empirical significance level associated with any observed maximum lod score. The question to be answered is simply the following: With the disease phenotypes seen in the family members, assuming markers with the same characteristics as those used and recombination fraction of 50% between disease and marker loci, what is the probability that the maximum lod score is at least 3? Such an evaluation of linkage results appears particularly important in multi-point analysis, where the expected lod score has multiple maxima, so that the statistical properties of map-distance estimates and test criteria are difficult to assess (7).

In tests for heterogeneity, the usual likelihood ratio test statistics may asymptotically follow χ^2 distributions, but even that is not guaranteed for a mixture of families in which different disease genes leading to the same phenotype may be linked with different genetic markers. Again, Monte Carlo procedures will elegantly find the distribution of a test statistic under the null hypothesis of homogeneity and, thus, will approximate the proper empirical significance level associated with some observed criterion for heterogeneity. Clearly, considerable computer resources may be required

for obtaining the distribution of a particular statistic. Because the statistic obtained in a given replicate represents one sample point in the distribution of that statistic, complete data analysis is generally required for each replicate. Such Monte Carlo methods are, therefore, only feasible to simulate analyses that do not take an inordinate amount of computer time. An example has been calculated in the following way.

Recently, a linkage analysis (for previously published and new families) between hereditary motor and sensory neuropathy type I versus the Duffy blood group locus on chromosome 1 was published (8). A few families show good evidence for linkage, whereas in most families, the two loci appear unlinked. Heterogeneity was tested by use of the HOMOG program (1), which estimates the proportion α of families with linkage and the recombination fraction θ in these families. This test showed highly significant heterogeneity. The authors conclude that "the total lod scores from many families are now overwhelmingly negative for linkage between hereditary motor and sensory neuropathy I and *FY*, but HOMOG testing for heterogeneity still suggests that there may be heterogeneity".

I reanalyzed 28 informative families from the literature for heterogeneity of linkage between hereditary motor and sensory neuropathy I and *FY* for which the HOMOG program yields a χ^2 value (1 df) of $\chi^2 = 8.45$ for heterogeneity, with estimates for α and θ of 0.20 and 0.05, respectively. Under the null hypothesis of homogeneity, asymptotically, $\chi^2 = 8.45$ is only exceeded with a probability (empirical significance level) of 0.0036, and when the test is considered one-sided, the empirical significance level is only 0.0018.

A computer simulation was then applied to determine the actual probability, under the null hypothesis, that for these 28 families the HOMOG program will yield a value of χ^2 exceeding 8.45. As the raw family data were not readily available, the lod scores for each family were converted to numbers of recombinants and nonrecombinants, yielding approximately the same maximum lod score at the same recombination fraction. The derived family sizes range from 1 through 20 opportunities for recombination. These data for the 28 families when analyzed by the HOMOG program lead to the same estimates of α and θ as obtained from the published lod scores and to $\chi^2 = 8.18$, which is significant at the 0.0021 level (one-sided). Clearly, counts of recombinants and nonrecombinants in these artificially created 28 families will have somewhat different statistical properties as the original 28 families, but it is hoped that they are nevertheless comparable with respect to the test for heterogeneity.

A computer program was written to randomly generate recombinants and nonrecombinants in 28 families with the same number of opportunities for recombination as the ones derived above. Because homogeneity is specified by a compound rather than a simple hypothesis, a range of different θ values between 0.05 and 0.50 was used. For a given θ, 10,000 random samples were generated, each sample was analyzed by the HOMOG program, and the proportion of samples with $\chi^2 > c$ was determined, which approximates the empirical significance level of the homogeneity test in these families. For the constant c three values were considered: 3.84 (nominal 5% level, two-sided), 6.63 (nominal 1% level, two-sided), and 8.18 (family data, nominal 0.4% level, two-sided). The results, shown in Table 1, demonstrate that the significance level depends on the true value of the recombination fraction θ and has a maximum around $\theta = 0.30$. The test is conservative when applied in a two-sided manner and may be slightly nonconservative when used in a one-sided manner. The observed χ^2 value of 8.18 is exceeded with a probability of, at most, 0.0021. One may, thus, conclude that the heterogeneity in the linkage between hereditary motor and sensory neuropathy I and *FY* is significant. These results also show that the test may reliably be used in a one-sided manner.

Table 1. Proportions of 10,000 random samples of 28 families each in which, at given recombination fraction θ, the critical value c of χ^2 is exceeded

	Critical value c of χ^2		
θ	3.84 (5%)*	6.63 (1%)*	8.18 (0.4%)*
0.05	0.0101	0.0019	0.0009
0.10	0.0163	0.0034	0.0014
0.20	0.0178	0.0039	0.0014
0.30	0.0258	0.0052	0.0021
0.40	0.0268	0.0048	0.0014
0.50	0.0136	0.0022	0.0008

*% nominal two-sided significance level.

A desirable next step in analyzing the properties of the homogeneity test is to determine its statistical power, a possible bias in the estimates of α and θ in real family data, and the effects of the presence of families with different mating types.

Once linkage of a new locus with a map of known marker loci has been established, the question arises whether the information is sufficient to confidently determine the interval on the map containing the new locus. Some statistical tests regarding order of loci have been developed (9, 10), but these tests are applicable only in special circumstances. The commonly used criterion for judging the significance of one order of loci against another order is the difference in maximum logarithm likelihood achieved under the two orders (11), where a difference of 1 unit of logarithm (likelihood) to base 10, corresponding to a likelihood ratio of 10:1, is usually considered important. The most relevant test statistic—that is, the difference in logarithm likelihood between the best and second best order—is difficult to assess statistically, as it does not correspond to a likelihood ratio test because different locus orders correspond to different regions in the *same* parameter space (1). A generally applicable solution, based on computer simulation, resembles previously published methods (9) and is proposed as follows.

As is well-known, there is a close connection between the confidence interval for an unknown parameter around its maximum likelihood estimate and tests of hypotheses (6). One view of this relationship is to recognize that variation of the unknown parameter generates a continuum of hypotheses, each of which might be regarded as a null hypothesis. The portion of this continuum that does not lead to rejection of the null hypothesis constitutes the confidence interval (12). Analogously, in multi-point linkage analysis, when the data has led to a maximum likelihood order of loci (the observed "best order"), a confidence set of locus orders associated with this best order should be determined. The particular solution proposed applies a randomization test as follows. A certain locus order is picked, and replicates of family-marker data under this order are generated as the null hypothesis. As in the homogeneity-test example discussed above, one also has to determine a reasonable set of interlocus recombination fractions under which the replicates are to be generated. For each replicate, the likelihood associated with the null order and the best order referred to above must be calculated (such calculations will generally involve estimating map distances). Then the proportion p of replicates in which the null order has lower likelihood than the best order is determined. If $p > 0.10$, the null order should be included in the confidence set. Repeating this procedure for (all) other orders leads to a confidence set with confidence coefficient of 90%. Generally, it is not necessary to try all orders as null orders. If one has started with the orders closest to the best order in terms of logarithm likelihood difference, presumably the first few orders will be included in the confidence set, whereas later, orders will be excluded, so that after a certain period of only exclusions the procedure is complete.

Multi-point linkage analysis raises many statistical problems nonexistent in classical linkage analysis. A particularly disturbing problem is that of the influence of different information from marker loci. For example, consider two marker loci, marker 1 to the left of marker 2, at a known genetic distance from each other, and assume a new locus the position d of which relative to the two marker loci is to be estimated. If its true position, d_0, is exactly midway between the marker loci, then the expected logarithm likelihood will generally have three maxima, an absolute maximum at d_0 and two local maxima—one at d_1 to the left of marker 1 and the other at d_2 to the right of marker 2. The two local maxima will be of equal height, provided that the two markers are equally informative. However, take as an example a situation where marker 2 has four alleles with frequencies of 0.25 each and marker 1 has two alleles with frequencies of 0.5 each. More matings are informative for marker 2 than for marker 1 and the local maximum at d_2 is higher than that at d_1. The implications of this finding are yet to be investigated. Presumably, markers that are differently informative lead to biases toward particular locus orders. However, because differently informative markers are part of the simulation procedure, the null distribution of the particular statistic used will automatically take unequal marker heterozygosity into account.

To implement evaluating the null distribution of test statistics by Monte Carlo procedures, problems of ascertainment have to be carefully considered. For example, when a rare allele of a two-allele marker segregates in a pedigree, one may want to force that allele to be present in each replicate—that is, simulations will be done conditional on the pedigree being informative for linkage. This is a meaningful strategy because, statistically speaking, uninformative pedigrees are clearly not part of the sampling space. Generally, however, the choice of conditions for simulations will affect the empirical significance, so that it is mandatory to choose conditions that accord with the actual mode of ascertainment used.

Stimulating discussions with Drs. M. Boehnke and L. Sandkuyl are gratefully acknowledged. This work has been supported by U.S. Public Health Service Research Grant MH44292 and by the W. M. Keck Foundation.

1. Ott, J. (1985) *Analysis of Human Genetic Linkage* (Johns Hopkins Univ. Press, Baltimore).
2. Boehnke, M. (1986) *Am. J. Hum. Genet.* **39,** 513–527.
3. Sandkuyl, L. A. & Ott, J. (1989) *Hum. Genet.,* in press.
4. Ploughman, L. M. & Boehnke, M. (1989) *Am. J. Hum. Genet.* **44,** 543–551.
5. Lathrop, G. M. & Lalouel, J. M. (1988) *Am. J. Hum. Genet.* **42,** 498–505.
6. Rao, C. R. (1973) *Linear Statistical Inference and Its Applications* (Wiley, New York).
7. Ott, J. & Lathrop, G. M. (1987) *Cytogenet. Cell Genet.* **46,** 674.
8. Middleton-Price, H. R., Harding, A. E., Berciano, J., Pastor, J. M., Huson, S. M. & Malcolm, S. (1989) *Genomics* **4,** 192–197.
9. Lathrop, G. M., Chotai, J., Ott, J. & Lalouel, J. M. (1987) *Ann. Hum. Genet.* **51,** 235–249.
10. Ott, J. & Lathrop, G. M. (1987) *Genet. Epidemiol.* **4,** 51–57.
11. Bishop, D. T. (1985) *Genet. Epidemiol.* **2,** 349–361.
12. Fisher, R. A. (1960) *The Design of Experiments* (Oliver & Boyd, Edinburgh).