The Annals of Statistics
1996, Vol. 24, No. 2, 449–465

LIKELIHOOD AND LINKAGE: FROM FISHER TO THE FUTURE[1]

BY E. A. THOMPSON

University of Washington

Genetic epidemiology is almost unique among the sciences in that computation of a likelihood function is the accepted approach to statistical inference. In the context of genetic linkage analysis, in which genes are mapped by analysing the dependence in inheritance of different traits, the use of likelihood dates back to the early work of Fisher and Haldane, and has seldom been seriously challenged. After introducing the underlying genetic concepts, this paper reviews the history of the statistics of linkage analysis, from 1913 to 1980, and its dependence on the development of likelihood inference.

With the sudden increase in genetic marker data deriving from new DNA technology, the potential for mapping the genes contributing to complex genetic traits is markedly increased, but the difficulties of likelihood analysis are also multiplied. With increasing complexity of models and the desire to make maximum use of available data on individuals not closely related, the likelihood approach to human linkage analysis faces new computational and methodological challenges. New methods are meeting some of these challenges; likelihood and linkage seem as closely interwoven as ever.

1. Introduction.

> ··· There is a widespread and urgent demand for competent mathematicians who understand that branch of mathematics known as theoretical statistics, but who are capable also of recognising situations in the real world to which such mathematics are applicable. [*R. A. Fisher: From a letter to John Wishart, dated October* 27, 1949, *agreeing to serve on a Faculty Board Committee to review a proposal for the Cambridge Diploma in Mathematical Statistics.*]

This paper was presented as the R. A. Fisher lecture at the Joint Statistical Meetings, Toronto, August 1994, and it is therefore appropriate to start with a quotation from Fisher. Not only does this quotation summarise his view as a statistical scientist, but it is also particularly appropriate to genetic analysis. Among all the real-world areas of science in which inference on the basis of a limited class of probability models is applied, the one where it is most applicable must be genetic linkage analysis. The model is very simple, but using it in the context of the real-world problem of finding the genes contributing to human disease susceptibility leads to many statistical challenges.

Received January 1995; revised September 1995.

[1] This paper is based on the R. A. Fisher Lecture, given at the Joint Statistical Meetings, Toronto, August 1994. Work supported in part by NIH Grant GM-046255.

AMS 1991 *subject classifications*. 62F03, 62F10, 92D10.

Key words and phrases. Genetic linkage, genome descent, Monte Carlo likelihood, multipoint location score, segregation indicators.

I am grateful to the Committee and the Organisers for their invitation to present the 1994 Fisher Lecture and for giving me the opportunity to talk about a subject that is central to my own past and future work. Above all I would like to acknowledge my Ph.D. and postdoctoral research advisors, Dr. A. W. F. Edwards and Professor L. L. Cavalli-Sforza. As Ph.D. student and postdoctoral researcher (respectively), each worked in Cambridge and was influenced by Fisher's teaching. Their admiration for his work has influenced my own. From Fisher to the future: it is hard to know where to draw the line, as research with each of my graduate students has influenced my thinking on likelihood or linkage or both. However, I would like particularly to acknowledge four of my former Ph.D. students whose work is most directly related to the content of this paper: Heike Bickeböller (Blossey), Kevin Donnelly, Charles Geyer and Shili Lin.

2. Foundations.

2.1. *The genetic model.* Modern genetics started with Mendel (1866), who postulated his two laws as a probability model. The following list summarises Mendel's laws in modern terminology.

1. Everyone has two genes (factors) controlling a given trait, one from the mother, one from the father.
2. When an individual has an offspring, a copy of a randomly chosen one of his two genes is copied (segregates) to the offspring,
3. Gene copying is independent of the other parent, independent for each child and independent for each trait (or locus).

Mendel's work was rediscovered in 1900, and not long thereafter geneticists realised that independence of segregations for different traits is not true. Instead, there are groups of traits, which are *linked*; the genes controlling them tending to be inherited by the child as a group, not independently. Fairly soon thereafter, geneticists associated this linkage (dependence) with the chromosomes, the linear DNA structures that can be seen in a cell nucleus. In the formation of the offspring chromosome, crossovers occur; these are points at which copying of the parental DNA switches from one parental chromosome to the other. For genes at any two given locations on a chromosome, recombination occurs if there are an odd number of crossover events between them. Then, at those locations, the offspring receives genes from different parental chromosomes—that is, deriving from different grandparents. The earliest genetic mapping started, by counting, in *Drosophila*, the combinations of types of genes inherited by offspring. Sturtevant (1913) showed the patterns were best explained by a linear arrangement of genes for different traits and he made the first *gene ordering* inference, by methods analogous to those still used today.

For those not familiar with genetics, Figure 1 shows a simple example, where estimation is a matter of counting. There are three genetic loci, labelled A, B and C, each with three corresponding alleles a_i, b_i and c_i, $i = 1, 2, 3$.

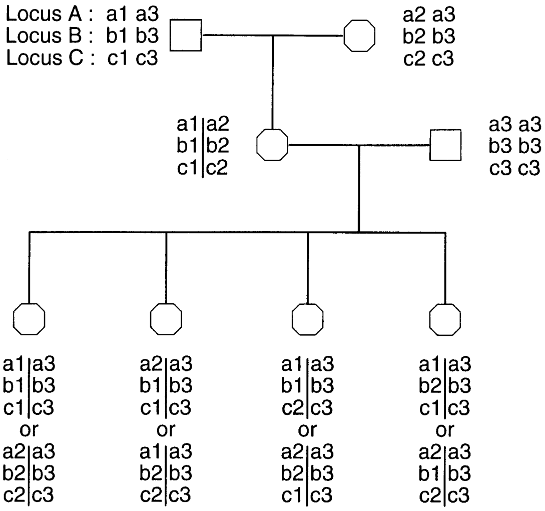

FIG. 1. *Segregation events in a small pedigree, where complete information on genotypes and grandparental gene origins can be inferred from the data on the types of alleles carried by individuals. There are three genetic loci, A, B, C, each with three corresponding alleles a_i, b_i, c_i ($i = 1, 2, 3$), here shown as being on a single chromosome in the order ABC.*

Because of the data on the mother's parents, we know the mother can only have $a_1b_1c_1$ on her maternal chromosome and $a_2b_2c_2$ on her paternal chromosome. The father must have $a_3b_3c_3$ on each of his two chromosomes and he passes this combination to each offspring, regardless of which paternal chromosome provides each offspring allele. From the mother there are four possibilities; a nonrecombinant offspring gets an intact maternal chromosome $a_1b_1c_1$ or $a_2b_2c_2$, while the other three possibilities consist of the allele at one locus deriving from a different maternal grandparent than the other two. For example, offspring maternal chromosome $a_1b_2c_2$ (or $a_2b_1c_1$) has alleles at locus A of different grandparental origin than those at B and C, and if the three loci are indeed on a single chromosome, in the order ABC, this implies recombination between locus A and locus B. From a large number of offspring of matings of this type, linkage (dependent segregation) can be tested for, recombination frequencies estimated and loci ordered—this last exercise depending on the fact that offspring in which the allele at the central locus is of different grandparental origin (e.g., $a_1b_2c_1$) will have much smaller frequency, since these can arise only as a result of recombination in both intervals (AB and BC).

Haldane (1919) extended the mathematical model, defining the distance along a chromosome as the expected number of crossover events. He related this additive *map distance* to recombination frequencies between loci on the assumption that crossovers occur as a homogeneous Poisson process. (This is the model of "no interference"; a given crossover event does not affect the probability distribution of the locations of other crossover events.) This sparked an exploration, which still continues, on the relationship between physical distance (actual DNA length) and map distance (the statistical measure based on the degree of dependence between genetic loci). However, the early statistical geneticists were clear that only recombination frequencies, and hence map distances, could be estimated from segregation data.

Fisher (1922b) used a slightly different relationship between recombination frequencies and map distance, assuming a crossover precluded any other crossover in the region of genome he considered. Under this model of "complete interference," recombination frequencies are themselves additive. With this simplifying assumption, he used the multinomial distribution of offspring counts as one of his first examples of maximum likelihood estimation, citing and drawing on his theoretical work published in the same year [Fisher (1922a)]. Thus from its earliest days, likelihood and linkage have been tightly connected.

2.2. *Human genetic linkage*: *the* 1930's. Linkage analysis in human genetics did not start until the 1930's, with the recognition that the same model used to analyze counts of offspring in experimental organisms could also be used to address data in human families—grandparents, parents and children (Figure 1). Several geneticists and statisticians were involved, but again it was Fisher and Haldane who furthered the likelihood approach—the approach of computing the probability of the observed data, under the probability model [Haldane, (1934); Fisher (1934)].

There are four main areas of discussion that are still active in the area of practical human genetic linkage analysis: each of these was recognised by the founders of this area, and particularly by R. A. Fisher.

1. Genetic counselling using observable genetic markers known to be linked to a locus determining a given genetic disease: Fisher (1934) foresaw the potential of linked markers for counselling and also foresaw some of the potential problems.
2. Power of a potential linkage study and assessment of the amount of trait data required to find a linkage: This was the focus of much of Fisher's work in this area [Fisher, (1934, 1935)].
3. Whereas putative linkages are often detected, for complex traits attempted confirmation by other studies often fails. In many cases, this may be due to genuine genetic heterogeneities—different genetic causes of apparently similar phenotypes. The possibility of linkage heterogeneity was noted by Fisher (1936). In this particular study he noted one family in which the recombination frequency was apparently different than in other families in the collection. Fisher analysed the evidence for linkage heterogeneity. He also cautioned against overinterpretation of the test statistic.
4. When now geneticists speak of finding a gene, they mean the physical DNA sequence. The relationship between physical and genetic distance now has practical implications that were irrelevant in the 1930's. However, the comments of Fisher and Haldane that only recombination patterns and map distances can be estimated by linkage analysis are no less true now.

Thus, although much has altered since the 1930's, the basic statistical framework, as developed by J. B. S. Haldane and R. A. Fisher remains. Over the next 50 years, there were, of course, further developments, and many contributions were made by many statistical geneticists. These led to better understanding of how inferences could be drawn [Haldane and Smith (1947); Morton (1955)], better methods for the computation of likelihoods [Elston and Stewart (1971)], and better understanding of their properties [Smith (1953)]. Ott (1991) covers many of these developments in his text.

3. Linkage analysis: 1935–1990.

3.1. *The likelihood function.* First, let us consider in more detail the form of the likelihood function required for linkage analysis. When we cannot observe or infer precisely which genes are on which chromosomes in all the relevant individuals, the likelihood is no longer a single multinomial. Instead, it is a sum **X** over all the latent possibilities:

$$P_\theta(\mathbf{Y}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y},\mathbf{X}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X}) P_\theta(\mathbf{X}). \tag{1}$$

Here **Y** are the observed data and **X** is everything else. Generally, the data on an individual depends only on his own genotype, and genes are transmitted from parents to offspring in accord with the Mendelian segregation probabilities [Mendel (1866)]. Thus the likelihood is most easily considered in the form

$$\begin{aligned} P_\theta(\mathbf{Y} \mid \mathbf{X}) &= \prod_{\text{observed}} P_\theta(Y_i \mid X_i) \\ P_\theta(\mathbf{X}) &= \prod_{\text{founders}} P_\theta(X_j) \prod_{\text{nonfounders}} P_\theta(X_j \mid X_{m_j}, X_{f_j}), \end{aligned} \tag{2}$$

where $\mathbf{X}_j$ now denotes the underlying types of the genes on the two chromosomes of individual j, and m_j and f_j are the parents of j. For data on linked loci, the probabilities of the combinations of genes in a child, given those in the parents, depend on the recombination frequencies; the linkage analysis parameters enter only through the final product of (2).

3.2. *The DNA revolution.*

> ... the enormous amount of Statistics at this moment at their disposal is absolutely useless. Why? Because ... [they] have received no education whatever on the point on which all ... must ultimately be based. We do not want a *neat arithmetical sum*. We want to know *what we are doing*. What we want first is not ... an accumulation of facts, but to teach [them] the *uses* of facts, "Statistics", ... [*Florence Nightingale*: *From a letter to Benjamin Jowett*, 1891, *concerning the teaching of Statistics at the University of Oxford*, *Quinn and Prest* (1987).]

Although more genetic markers that could be typed in human individuals were gradually accumulated, in 1980 the entire biological framework changed with the advent of very large numbers of DNA markers [Botstein, White, Skolnick and Davis (1980)]. These are traits which first have to be mapped relative to each other and then can be used to map other traits of medical significance. Since 1980, the various kinds of these DNA markers have proliferated and the number available has exploded; for a view of the current human linkage map, see Murray et al. (1994). The mass of data creates many statistical problems, while also of course vastly increasing the potential power to map the genes contributing to complex traits—assuming that there are genes that make a significant contribution to the trait in question. Florence Nightingale, in the preceding quote, was writing about government Public Health statistics and Members of Parliament in 1891; her comments would apply also to genome data bases and molecular geneticists in 1991. There is no "neat arithmetical sum" that will extract a map from the mass of data currently available.

Rather than seeking to estimate recombination rates of perhaps 20%, it is necessary to order markers between which recombination rates are less than 1%. A marker map with 1% recombination frequencies is a goal of the Human Genome Project [Murray et al. (1994)]. To analyse these small recombination frequencies, far more data are required. Now genome scans are done to search for linkage; thus, we have the problem of multiple dependent tests. The method of map-specific multipoint linkage analysis is often used. That is, the marker map is assumed known and a log-likelihood difference (or "location score") is computed for each hypothesised location of the trait locus relative to the hypothesis that no trait locus is linked to this segment of the marker map. Because, even when the trait gene is in some map interval, there will often be strong evidence for recombination between the trait locus and an adjacent marker, the log-likelihood curve for the location of the trait locus will normally have multiple peaks with sharp decreases at the marker locations (Figure 2). The interpretation of such likelihood surfaces is not straightforward. Further, there may be uncertainties about the marker map, which can have a strong impact on the location log-likelihood curve. Differences in male and female genetic map distances may also have an impact. Moreover, the model of Haldane (1919), which assumes crossover events occur independently of each other and hence provides for a simple relationship between map distance and recombination frequencies, may no longer be adequate.

3.3. *Association tests for linkage.*

> Hypotheses which may be true may be rejected because they have not predicted observable results which have not occurred. [*Jeffreys* (1961), *page* 385]

Because of difficulties of likelihood computation, trait model uncertainties and the mass of marker data, often with different markers typed or informative in different families, a number of researchers moved away from likelihood analyses of linkage in the 1980's and instead developed a variety of association tests for linkage detection. There are two main classes of such tests: those at the population level and those at the family level. It is not within the scope of this paper to review the extensive literature on these tests; only sufficient information will be given to relate the ideas to the overall theme.

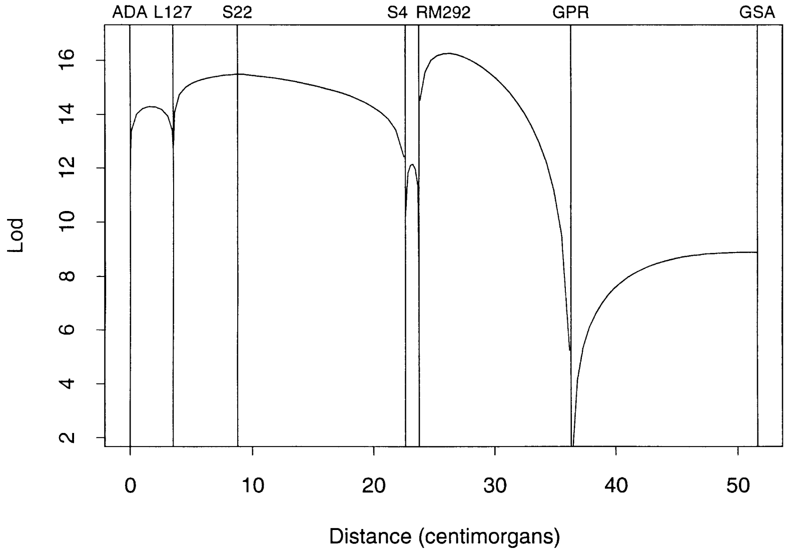

FIG. 2. *Nine-point location score for the MODY trait estimated by sequential imputation. [Figure* 3 *of Irwin, Cox and Kong* (1994), *reproduced with permission of the authors and the National Academy of Sciences from a graphics file provided by Mark Irwin.*]

Where there is very tight linkage, there will be no recombination within pedigrees and markers cannot be ordered. However, there will be population associations, due to a particular allele arising initially on a particular chromosomal background and recombination being very slow to break up this historical association. The earlier ideas of using the estimated magnitudes of pairwise associations to map loci at larger genetic distances have not proved useful. Unless recombination frequencies are very small, the residual effect of history rapidly declines. Moreover, the information in population samples as to the magnitude of associations depends on allele frequencies at the loci in question and can also be small.

However, "disequilibrium mapping" can be very useful in ordering the loci of a tightly linked group, particularly where the trait allele is rare and recent (on an evolutionary time scale). The combinations of marker alleles on chromosomes which carry a given trait allele provide an indication of the recombinations that have occurred over the generations since the trait allele first arose. Thus, far more segregations (and opportunities for recombination) are implicitly observed than can be observed in an analysis of marker and trait segregations in a pedigree. Of interest here is that the first such locus ordering on the basis of population frequency data was that of Fisher (1947). Fisher regarded his analysis of the Rhesus blood group system as a prime example of scientific inference; the final step of his analysis was to infer the order of the three main loci of the system from the population frequencies of the eight possible combinations of alleles at the three loci.

Within a pedigree there are also associations that indicate linkage. Relatives with similar trait values or affected for the same disease are likely to share genes at loci affecting that trait or disease, these genes being copies of a single gene in a recent common ancestor. In this case, they will share genes also at closely linked loci. Hence it is possible to construct a test statistic to test for gene sharing at marker loci, on the basis of sharing of marker alleles, and hence to locate genes contributing to a trait. Under the null hypothesis of no linkage, test statistics have distributions independent of the trait model. Thus they can provide a useful screening tool. However, not only do these tests generally lack power, but power is dependent on a trait model. Moreover, in the context of genome-wide searches for markers linked to genes contributing to complex traits, even the P-value for the rejection of absence of linkage (the "hypothesis which may be true, but may be rejected") needs careful consideration. If there is a gene contributing to the trait, it must be located somewhere in the genome, but any framework for the interpretation of results must consider not only trait gene location, but trait gene existence. The trait model cannot be ignored.

In fact, once probabilities under linkage alternatives are computed, we return to likelihood inference, although possibly on a subset of the available data. A specific instance is provided by Smith (1953), who considered the maternal and paternal marker allele similarities in inbred individuals affected by a rare recessive disease, and hence having this disease allele on both the maternal and paternal chromosome. At one level, this is an association test based on the shared marker types on paternal and maternal chromosomes ("homozygosity mapping"), but also it is a likelihood analysis based on the marker and trait data on unrelated inbred individuals.

While much is written about the advantages of so-called model-free association tests, when a linkage likelihood can be obtained, practitioners want it. However, the computational methods that could be used before 1980 are no longer practical for many of the data sets of today.

4. Monte Carlo estimation of linkage likelihoods.

4.1. *Importance sampling.* Due to the increasing demands of the available data, there has been much recent effort directed toward improving computational methods. Some of this effort has been toward exact methods [Cottingham, Idury and Schäffer (1993); Lander and Green (1987)], but also there has been an explosion of ideas for the use of Monte Carlo estimation methods. We consider here only Monte Carlo methods for the estimation of likelihoods or location scores; there have also been recent advances in simulation methods for studying the distributional properties of log-likelihoods, conditional on partial data.

The likelihood (1) is a sum over a huge space of $\mathbf{X}$-values; Monte Carlo integration is an obvious route, sampling from some probability distribution $h(\cdot)$ on this space. The simplest form of the likelihood as an expectation dates back to Ott (1979):

$$\begin{aligned} L(\theta) = P_\theta(\mathbf{Y}) &= \sum_{\mathbf{X}} P_\theta(\mathbf{Y},\mathbf{X}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X})\ P(\mathbf{X}) \\ &= \mathrm{E}_\theta(P_\theta(\mathbf{Y} \mid \mathbf{X})) = \mathrm{E}_{\theta_0}\left(P_\theta(\mathbf{Y} \mid \mathbf{X}) \frac{P_\theta(\mathbf{X})}{P_{\theta_0}(\mathbf{X})} \right), \end{aligned} \tag{3}$$

the last equation being due to K. Lange [Ott (1979)]. On large pedigrees or for multiple loci, sampling from prior distributions $P_\theta(\mathbf{X})$ or $P_{\theta_0}(\mathbf{X})$ is ineffective, since these prior probabilities bear no relation to the observed data $\mathbf{Y}$. However, the formulation (3) does introduce three key ideas. The first is use of Monte Carlo integration, the second importance sampling and the third, estimation of a function of θ by simulation at a single θ_0.

For importance sampling, we want the sampling distribution to mimic the integrand. Since

$$P_\theta(\mathbf{Y},\mathbf{X}) \;=\; P_\theta(\mathbf{X} \mid \mathbf{Y}) P_\theta(\mathbf{Y}) \propto P_\theta(\mathbf{X} \mid \mathbf{Y}),$$

this means we must sample from something close to the conditional distribution of underlying genotypes ($\mathbf{X}$-values) given the data. Direct Monte Carlo from exactly this distribution is impossible; if a probability distribution proportional to the integrand were explicitly known, so also would be the value of the integral, and the Monte Carlo would be unnecessary [Hammersley and Handscomb (1964)]. However, there are at least two ways to come close to the required sampling.

4.2. *Sequential imputation.* One approach is that of sequential imputation [Kong, Liu and Wong (1994)]. The development for the case of multilocus linkage likelihoods is given by Irwin, Cox and Kong (1994) and, in brief, is as follows. Suppose we have data on m genetic loci (say $m-1$ markers and a disease). Let Y_l now denote the data for locus l and let X_l denote the underlying genotypes at that locus. A realisation X_l^* is obtained for each locus in turn from the distribution

$$P_{\theta_0}(X_l \mid X_1^*, \ldots, X_{l-1}^*, Y_1, \ldots, Y_{l-1}, Y_l).$$

Note that this probability can be rewritten as

$$\frac{P_{\theta_0}(X_l, Y_l \mid X_1^*, \ldots, X_{l-1}^*, Y_1, \ldots, Y_{l-1})}{P_{\theta_0}(Y_l \mid X_1^*, \ldots, X_{l-1}^*, Y_1, \ldots, Y_{l-1})}.$$

Hence it is readily shown that the joint simulation distribution for $\mathbf{X}^* = (X_1^*, \ldots, X_m^*)$ is

$$P^*(\mathbf{X}^*) = P_{\theta_0}(\mathbf{Y},\mathbf{X}^*)/w_m(\mathbf{X}^*),$$

where

$$w_m(\mathbf{X}^*) = \prod_{l=1}^{m} P_{\theta_0}(Y_l \mid Y_1, \ldots, Y_{l-1}, X_1^*, \ldots X_{l-1}^*).$$

Thus, in addition to computing the successive simulation distribution for each X_l^*, the predictive probabilities of the observed data Y_l conditional on genotypes (and phenotypes) at preceding loci

$$P_{\theta_0}(Y_l \mid Y_1, \ldots, Y_{l-1}, X_1^*, \ldots, X_{l-1}^*)$$

must also be evaluated. The product of these predictive weights is accumulated:

$$w_i = \prod_{l=1}^{i} P_{\theta_0}(Y_l \mid Y_1, \ldots, Y_{l-1}, X_1^*, \ldots, X_{l-1}^*)$$

to provide finally the denominator $w_m(\mathbf{X}^*)$.

Then

$$E_{P^*}(w_m(\mathbf{X}^*)) = \sum_{\mathbf{X}^*} w_m(\mathbf{X}^*) P^*(\mathbf{X}^*) = P_{\theta_0}(\mathbf{Y}) \tag{4}$$

and thus a Monte Carlo estimate of $L(\theta_0) = P_{\theta_0}(\mathbf{Y})$ is given by the mean value of $w_m(\mathbf{X}^*)$, over repeated independent repetitions of the sequential imputation process. Repeating the process for different disease locus locations, one obtains an estimated location log-likelihood curve. Figure 2 shows the resulting location curve from Irwin, Cox and Kong (1994)—a very typical location score curve for the disease locus against a known map of eight markers at seven distinct locations. There is strong evidence for a gene, although precise interpretation of such curves is unclear. On this large pedigree exact computation would be impossible. In some cases, the sequential imputation computations are also impossible, in particular for complex pedigrees. However, in many cases this approach will be both feasible and successful.

Very often, one wants to compute conditional probabilities, given the data, with respect to some particular model $P_{\theta_0}(\cdot)$: Where are the recombinations? Whom should we sample to obtain most information? Where are the biggest uncertainties in underlying marker genotypes? How would it affect inferences to reduce such uncertainty? In principle, such expectations can be readily estimated, using the sequential imputation probability distribution P^* and computed weights w_m:

$$\begin{aligned} E_{\theta_0}(g(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Y}) &= \sum_{\mathbf{X}} g(\mathbf{X}, \mathbf{Y}) P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \qquad (5) \\ &= \sum_{\mathbf{X}} g(\mathbf{X}, \mathbf{Y}) P^*(\mathbf{X}) w_m(\mathbf{X}) / P_{\theta_0}(\mathbf{Y}) \\ &= E_{P^*}(g(\mathbf{X}, \mathbf{Y}) w_m(\mathbf{X})) / P_{\theta_0}(\mathbf{Y}). \qquad (6) \end{aligned}$$

Equation (4) provides a Monte Carlo estimate of $P_{\theta_0}(\mathbf{Y})$ so that

$$E_{\theta_0}(g(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Y}) = E_{P^*}(g(\mathbf{X}, \mathbf{Y}) w_m(\mathbf{X})) / E_{P^*}(w_m(\mathbf{X})).$$

However, the numerator and denominator here require separate Monte Carlo estimates, which are then combined in a ratio estimate.

4.3. *Markov chain Monte Carlo.* An alternative approach avoids use of a ratio estimator. The expectation (5) could be estimated directly if realisations from $P_{\theta_0}(\cdot \mid \mathbf{Y})$ were available. Direct simulation is not possible since, again, this probability distribution is known only up to the unknown normalising factor $P_{\theta_0}(\mathbf{Y})$, but the realisations can be obtained by Markov chain Monte Carlo (MCMC). Here is not the place for even a brief review of MCMC. Suffice it to say that by using the Gibbs sampler, or any of the broad class of Metropolis–Hastings algorithms [Hastings (1970)], one obtains realisations from a Markov chain whose equilibrium distribution (the desired distribution) is known only up to such an unknown normalising factor. They are dependent realisations, which complicates analysis as compared to the independent realisations of sequential imputation. Further, only when effects of the starting configuration have decayed are the realisations from the desired equilibrium distribution of the Markov chain. On the other hand, there is no reweighting, no weights to be computed and no separate estimation of the normalising factor. Whereas sequential imputation could be regarded as sampling from approximately the right distribution, MCMC is sampling approximately from the right distribution.

Of all expectations we could estimate, a key one is the likelihood ratio itself:

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X})}{P_{\theta_0}(\mathbf{Y},\mathbf{X})} \;\middle|\; \mathbf{Y} \right) \tag{7}$$

[Thompson and Guo (1991)]. That is, with observed data $\mathbf{Y}$ and latent variables (genotypes) $\mathbf{X}$, the likelihood ratio is the expected complete-data likelihood ratio, where genotypes $\mathbf{X}$ are sampled conditional on $\mathbf{Y}$. Hence we can estimate a likelihood ratio function by the average of complete-data likelihood ratios at realised $\mathbf{X}$-values:

$$\frac{1}{N}\sum_{l=1}^{N}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X}^{(l)})}{P_{\theta_0}(\mathbf{Y},\mathbf{X}^{(l)})} \right). \tag{8}$$

Locally (for $\theta \approx \theta_0$) the estimate will be "good" in an importance sampling sense. Further, we have an estimate as a function of θ from realisations at a single θ_0. [Lange and Sobel (1991) also develop a MCMC approach to genetic linkage location score curves. Their method uses a different space of latent variables $\mathbf{X}$, a different MCMC algorithm and a different expectation equation for estimation of the likelihood. Each method has its advantages.]

In practice, more has to be done to obtain a useful method from (7). The first problem is the sampling at a single θ_0. This will be useless if we require a likelihood over a large range of genetic maps or trait models. Geyer (1991) provided a solution; without the technical detail, his solution is to sample at a large number of different θ_j and combine the estimates in accordance with the appropriate importance sampling weights. Second, the Gibbs sampler will not sample effectively the huge space of multilocus genotype configurations

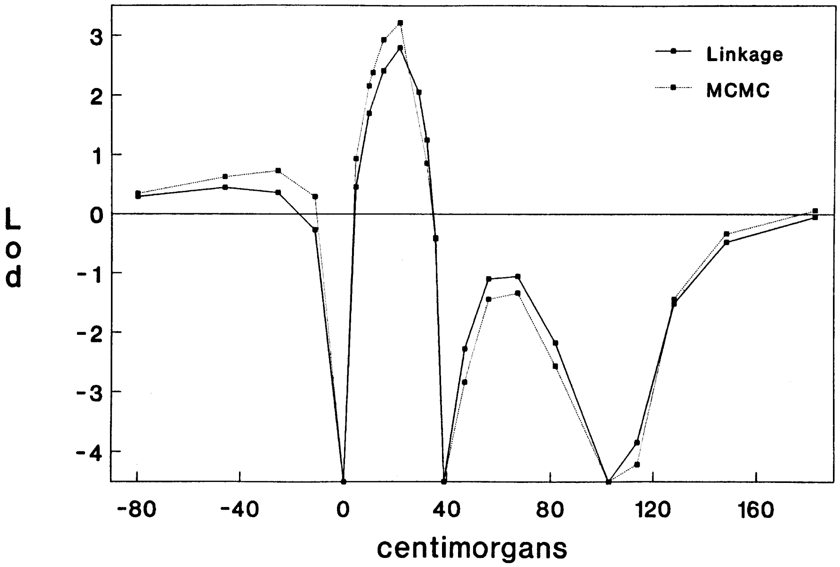

FIG. 3. *Five-point location score for a set of markers obtained by MCMC, with exact values for comparison. [The figure is due to Lin and Wijsman* (1994), *computed using the method of Lin* (1995), *using marker data from Palmer, Dale, Livingston, Wijsman and Stephens* (1994).]

for multiallelic markers on extended pedigrees. However, samplers that update more than one individual's genotype at a time are hard to implement, because of the constraints of Mendelian segregation. The marker data impose additional constraints that make it difficult (or sometimes impossible) for a single-site updating method to reach all feasible genotypic configurations. For the particular case of multilocus linkage analysis, the work of Lin [Lin, Thompson and Wijsman (1994); Lin, (1995)] has resolved many of these difficulties. Figure 3 shows a location score curve obtained by MCMC for a problem for which exact computation is still feasible (and the exact solution also shown), but for which MCMC requires 15 times less CPU time [Lin and Wijsman (1994)]. With an additional marker, exact computation would no longer be practical.

5. Segregation indicators and genome descent.

5.1. *Segregation indicators as latent variables.* Another sampling design for linkage analysis, first considered by Smith (1953) but recently gaining in popularity, is that of data on a single individual who has two copies of a rare disease gene. If the trait is sufficiently rare, the posterior probability is high that these are two copies of a single gene in a recent common ancestor of his parents. Figure 4 shows a pedigree arising in a recent study of a rare recessive trait [Nakura et al. (1994)]; any of the three marked founder ancestors could contribute two copies of a gene to the affected individual. If the individual

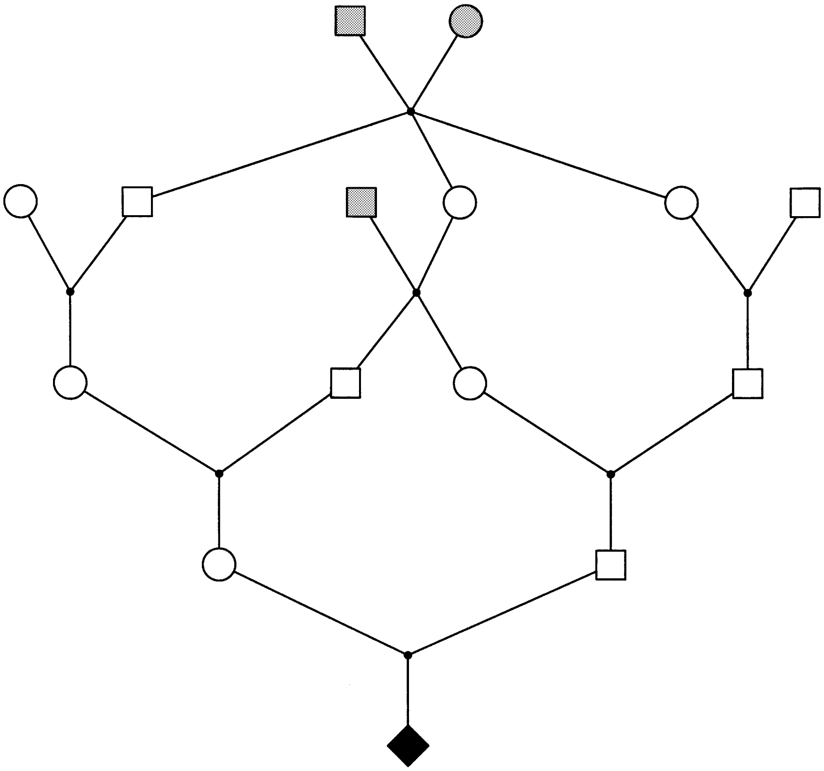

FIG. 4. *The complex ancestry of an individual affected with a rare recessive disease. This pedigree is a part of a study described in Nakura et al.* (1994). *Any of the three founder individuals shaded gray can contribute two copies of a single gene to the final affected individual* (*the black diamond*).

has two copies of a single gene at the disease locus, he likely does so also at nearby loci—patches of homozygosity are evidence for linkage. Again we want to know the likelihood—the probability of observed data under alternative hypotheses. On a complex pedigree, such as that of Figure 4, exact computation methods fail with more than a very few loci. The Monte Carlo methods of the previous section are also not satisfactory. The sequential imputation computations are nontrivial on a pedigree of this complexity and the alternative multilocus genotypic patterns on all the unobserved ancestors are not easy to sample with a MCMC method.

From the earliest linkage analyses, perhaps even dating from Mendel's (1866) latent "factors," there has been a tendency to consider genotypes as the underlying latent variables. The form of the likelihood (1) and (2) expresses this view. However, there is a more basic specification of gene descent. The paths of descent of genes are determined by a specification, for every point along a transmitted chromosome, of whether the gene derives from the parent's paternal or maternal chromosome [Donnelly (1983)]. Thus we can define segregation indicators:

$$S_{ij} = \left\{ \begin{matrix} 1 \\ 0 \end{matrix} \right\} \text{ if segregation } i \text{ at locus } j \text{ is parent's } \left\{ \begin{matrix} \text{paternal} \\ \text{maternal} \end{matrix} \right\} \text{ gene}$$

The space of segregation indicators, although large, is very much smaller than the space of individual multilocus genotypes. The segregation indicators in different segregations are independent. If absence of interference is assumed, the probability of a single S_{ij} conditional on the remaining indicators depends only on the values for the same segregation for the two adjacent loci. Further, where only one, or a very few, individuals are observed on a pedigree, the probability of observed data conditional on specified values of the segregation indicators can be computed very rapidly. It is therefore possible to implement an effective MCMC method for sampling the $\mathbf{S} = \{S_{ij}\}$ conditional on observed data for this type of linkage analysis design, and hence to obtain Monte Carlo estimates of likelihood ratios for linkage, analogous to (6) and (7), with **S** replacing **X**. Just as for the earlier discussion, sampling from the conditional distribution provides effective Monte Carlo estimates, whereas sampling from a prior distribution often cannot. Details are given elsewhere [Thompson (1994a, b)].

5.2. *Genome sharing and Fisher's theory of junctions.* In the 1980's, with the increasing amounts of genetic marker data, association tests for the detection of linkage gained in popularity. In Section 3.3 we found that, first, to interpret results of such tests one needs the probabilities of data under alternative hypotheses (the likelihood function) and, second, that Fisher (1947) had considered the information for mapping available in population associations at tightly linked loci.

Now history is repeating itself. One current enthusiasm in genetic mapping is genome matching [Nelson, McCusker, Sander, Kee, Modrish and Brown (1993)], in which the genomes of individuals having particular characteristics of interest are compared. However, to interpret the results of such a comparison a likelihood is still required, and conditional probabilities of genome sharing, given observed marker data, are not easily computed. The case of shared genome between the two chromosomes of an individual having two copies of a rare allele, considered in the preceding text, is the simplest possible case. Donnelly (1983) provides a framework for such computations in the continuous genome setting and considered pairs of relatives connected by a single descent path and descent from a parent to a set of offspring. This framework has been used by Blossey (1993) to address more complex questions of genome descent. Figure 5 shows the complexity that arises as soon as more than one descent path or more than one generation of segregations is considered, even with the simplest model of recombinations occurring as a Poisson process along a chromosome. The patterns of grandparental genome among sibs are constrained by the genome that the parent receives from the grandparent.

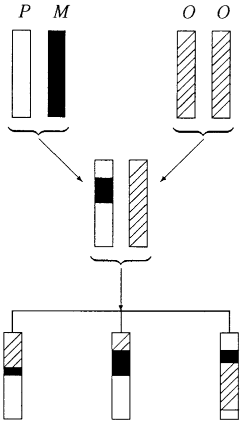

FIG. 5. *Segregation of chromosome seqments in a three-generation pedigree, showing the dependence among offspring resulting from their shared history in the genome received by their parent from her parents.* [*Figure from Blossey* (1993); *reproduced from a graphics file provided by Heike Bickeböller* (*Blossey*)].

Which finally brings us back to Fisher (1949), 45 years ago. In that year Fisher published his small monograph on the *Theory of Inbreeding*, intended, according to the Preface, as "a practical handbook for animal breeders." While it is doubtful it served this purpose, the book contains much interesting material on gene descent and a chapter in which the descent of genome is considered. Fisher performed (by hand, using tables of random digits) a Monte Carlo simulation of the descent patterns of chromosome segments in a sib-mating system. In fact, he did a conditional simulation, conditioning on continuing variation at specified points in the genome. Now that we have fast workstations, simulation conditional on data, and large scale Monte Carlo likelihood analyses become a practical proposition, even for the more complex problems resulting from a dense genetic map that provides information on shared genome.

Acknowledgments. I am grateful to Dr. A. Kong and the National Academy of Sciences for permission to use Figure 2, to Dr. S. Lin and Dr. E. Wijsman for Figure 3, to Dr. E. Wijsman for the information used in Figure 4, which exhibits a pedigree structure obtained in a study supported by NIH grant AG 08303, and to Dr. Heike Bickeböller for Figure 5. I am also grateful to Dr. Mike Boehnke, Dr. Charles Geyer, Dr. Shili Lin, Dr. Ellen Wijsman and to two referees for comments on an earlier draft.

REFERENCES

BLOSSEY, H. (1993). The Poisson clumping heuristic and the survival of genome in small pedigrees. Ph.D. dissertation, Dept. Statistics, Univ. Washington.

BOTSTEIN, D., WHITE, R. L., SKOLNICK, M. H. and DAVIS, R. W. (1980). Construction of a linkage map in man using restriction fragment polymorphism. *American Journal of Human Genetics* **32** 314–331.

COTTINGHAM, R. W., IDURY, R. M. and SCHÄFFER, A. A. (1993). Faster sequential genetic linkage computations. *American Journal of Human Genetics* **53** 252–263.

DONNELLY, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoret. Population Biol.* **23** 34–64.

ELSTON, R. C. and STEWART, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21** 523–542.

FISHER, R. A. (1922a). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

FISHER, R. A. (1922b). On the systematic location of genes by means of crossover observations. *American Naturalist* **56** 406–411.

FISHER, R. A. (1934). The amount of information supplied by records of families as a function of the linkage in the population sampled. *Annals of Eugenics* **6** 66–70.

FISHER, R. A. (1935). The detection of linkage with "dominant" abnormalities. *Annals of Eugenics* **6** 187–201.

FISHER, R. A. (1936). Heterogeneity of linkage data for Friedrich's ataxia and the spontaneous antigens. *Annals of Eugenics* **7** 17–21.

FISHER, R. A. (1947). The *Rhesus* factor: a study in scientific method. *Amer. Sci.* **35** 95–102, 113.

FISHER, R. A. (1949). *The Theory of Inbreeding.* Oliver & Boyd, Edinburgh.

GEYER, C. J. (1991). Reweighting Monte Carlo mixtures. Technical Report 568, School of Statistics, Univ. Minnesota.

HALDANE, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8** 299–309.

HALDANE, J. B. S. (1934). Methods for the detection of autosomal linkage in man. *Annals of Eugenics* **6** 26–65.

HALDANE, J. B. S. and SMITH, C. A. B. (1947). A new estimate of the likage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* **14** 10–31.

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1964). *Monte Carlo Methods.* Methuen, London.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

IRWIN, M., COX, N. and KONG, A. (1994). Sequential imputation for multilocus linkage analysis. *Proc. Nat. Acad. Sci. USA* **91** 11,684–11,688.

JEFFREYS, H. (1961). *The Theory of Probability*, 3rd ed. Clarendon, Oxford.

KONG, A., LIU, J. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

LANDER, E. S. and GREEN, P. (1987). Construction of multilocus linkage maps in humans. *Proc. Nat. Acad. Sci. USA* **84** 2363–2367.

LANGE, K. and SOBEL, E. (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics* **49** 1320–1334.

LIN, S. (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics* **51** 318–322.

LIN, S., THOMPSON, E. A. and WIJSMAN, E. M. (1994). A faster mixing algorithm for Hastings–Metropolis updates on complex pedigrees. *Annals of Human Genetics* **58** 343–357.

LIN, S. and WIJSMAN, E. M. (1994). Monte Carlo multipoint linkage analysis. *American Journal of Human Genetics* **55** A40.

MENDEL, G. (1866). Experiments in plant hybridisation. [Mendel's original paper in English translation, with a commentary by R. A. Fisher, published by Oliver and Boyd, Edinburgh, 1965.]

MORTON, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7** 277–318.

MURRAY, J. C., BUETOW, K. H., WEBER, J. L., LUDWIGSON, S., SCHERPIER-HEDDEMA, T., MANION, F., QUILLEN, J., SHEFFIELD, V. C., SUNDEN, S., DUYK, G. M., WEISSENBACH, J., GYAPAY, G., DIB, C., MORRISSETTE, J., LATHROP, G. M., VIGNAL, A., WHITE, R., MATSUNAMI, N., GERKEN, S., MELIS, R., ALBERTSEN, H., PLAETKE, R., ODELBERG, S., WARD, D., DAUSSET, J., COHEN, D. and CANN, H. (1994). A comprehensive human linkage map with centimorgan density. *Science* **265** 2049–2064.

NAKURA, J., WIJSMAN, E. M., MIKI, T., KAMINO, K., YU, C-E, OSHIMA, J., FUKUCHI, K., WEBER, J. L., PIUSSAN, C., MALARAGNO, M. I., EPSTEIN, C. J., SCAPPATICCI, S., FRACCARO, M., MATSUMURA, T., MURANO, S., YOSHIDA, S., FUJIWARA, Y., SAIDA, T., OGIHARA, T., MARTIN, G. M. and SCHELLENBERG, G. D. (1994). Homozygosity mapping of the Werner's syndrome locus (WRN). *Genomics* **23** 600–608.

NELSON, S. F., MCCUSKER, J. H., SANDER, M. A., KEE, Y., MODRISH, P. and BROWN, P. O. (1993). Genomic mismatch scanning; a new approach to genetic linkage mapping. *Nature Genetics* **4** 11–18.

OTT, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31** 161–175.

OTT, J. (1991). *Analysis of Human Genetic Linkage*, 2nd ed. Johns Hopkins Univ. Press.

PALMER, S. E., DALE, D. C., LIVINGSTON, R. J., WIJSMAN, E. M. and STEPHENS, K. (1994). Autosomal dominant cyclic hematopoiesis: exclusion of linkage to the major hematopoietic regulatory gene cluster on chromosome 5. *Human Genetics* **93** 195–197.

QUINN, E. V. and PREST, J. M. (1987). *Dear Miss Nightingale* (*A Selection of Benjamin Jowett's Letters* 1860–1893). Clarendon, Oxford.

SMITH, C. A. B. (1953). Detection of linkage in human genetics. *J. Roy. Statist. Soc. Ser. B* **15** 153–192.

STURTEVANT, A. H. (1913). The linear association of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14** 43–59.

THOMPSON, E. A. (1994a). Monte Carlo likelihood in linkage analysis. *Statist. Sci.* **9** 355–366.

THOMPSON, E. A. (1994b). Monte Carlo estimation of multilocus autozygosity probabilities. In *Proceedings of the 1994 Interface Conference* (J. Sall and A. Lehman, eds.) 498–506. Interface Foundation of North America, Fairfax Station, VA.

THOMPSON, E. A. and GUO, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.* **8** 149–169.

DEPARTMENT OF STATISTICS
GN-22
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98115