Monte Carlo in Genetic Analysis
September, 1995

E. A. Thompson
Department of Statistics, University of Washington.

Background and Acknowledgements

These notes are based on parts of a course in **Genetic Epidemiology** given May 27 to June 3, 1995, at the Danish Cancer Research Centre, Copenhagen, Denmark. The course was taught jointly with Drs. Robert Elston and Duncan Thomas and was funded by the Nordic Council. Session numbers relate to that course schedule; other sessions were taught by Drs. Elston or Thomas. This technical report incorporates also the notes from a course given at Wageningen Agricultural University, Wageningen, Netherlands, May 16-20, 1994, on **Monte Carlo Likelihood in Genetic Analysis.** This course received funding from the E.E.C. I am grateful to Dr. Johan van Arendork (Wageningen) and Dr. Bendix Carstensen (Copenhagen) for their invitations to teach these courses.

Abreviated versions of the course, focussing on sessions 5 and 16 of the current notes were also given as 3-hour workshops at Mcgill University (June 6, 1995) and University of Adelaide (August 29, 1995). I am grateful to Professor Ken Morgan (McGill University) and Dr. Patty Solomon (University of Adelaide) for their hospitality, and that of their Departments, during the writing of this report.

This work was supported in part by NIH grant GM-46255, and by NSF grant BIR 9305835.

Table of contents

Sections 8.4, 8.5, the last part of 5.6 and the Exercises are from the Wageningen course; they were not included in the Copenhagen course. The Wageningen course consisted primarily of sessions 4, 5 and 8 of the Copenhagen course. The workshops at McGill and Adelaide Universities focussed on sessions 5 and 16.

Session 1 of the Copenhagen course was an introduction to genetics given by Dr. Elston. Although some basic terminology is repeated in these notes, no comprehensive introduction is given. For session 10 on ascertanment, the reader is referred to Thompson (1993). The notes for the sessions on classical linkage analysis (sessions 12 and 13) are also in outline only; Ott (1991) is the best text for this area, and there is a huge literature on more complex aspects.

Figure 1 is a pedigree structure originally deriving from a data set studied by Dr. Ellen Wijsman. Figure 3 is based on a graph in Irwin et al. (1994).

Session 4. Gene identity by descent and probabilities on pedigrees

4.1 Gene identity by descent

A *gene*, as opposed to an allele or a locus, is the *DNA segment that codes for a functional unit* (Elston; session 1). It is this DNA segment that is copied from parents to offspring. Underlying the patterns of phenotypes observed on related individuals are the *genotypes*, but underlying the genotypes are the patterns of gene identity by descent. Phenotypes of relatives are similar because they have similar genotypes and may share a common environment. Genotypes are similar because relatives share genes that are identical by descent (*ibd*) — identical copies of a gene segregating from a common ancestor within the defined pedigree. Disregarding mutation (which for modern microsattelite markers one probably shouldn't), genes that are *ibd* must be of the same allelic type, while genes that are not *ibd* are of independent allelic types.

Gene identity by descent is defined only within the context of a defined pedigree. A pedigree specifies the two parents of every non-founder individual. A founder has neither parent specified, and by definition the genes in founders are not *ibd*. It will often be convenient if a pedigree is ordered in such a way that every individual is preceded in the listing by his parents; this is clearly always possible.

Mendel's first law states that:

> a diploid individual receives at any given locus a copy of a randomly chosen one of the two genes in his father and (independently) a copy of a randomly chosen one of the two genes in his mother, and will pass on a copy of a randomly and independently chosen one of these two genes to each of his offspring.

This simple law leads to complex patterns of gene identity on an extended pedigree, due to the huge number of alternative events; 2^k for k segregations. The segregating genes determine the patterns of gene identity by descent on the pedigree, and hence the patterns of similarity among relatives.

4.2 Kinship and inbreeding

We start with coefficients of *inbreeding* and *kinship*, since this provides a simple introduction to the basic ideas of gene identity by descent, to alternative computational approaches, and to Monte Carlo estimation of expectations. Kinship and inbreeding are best thought of as relationships between gametes rather than between individuals. The coefficient of kinship between two individuals B and C,

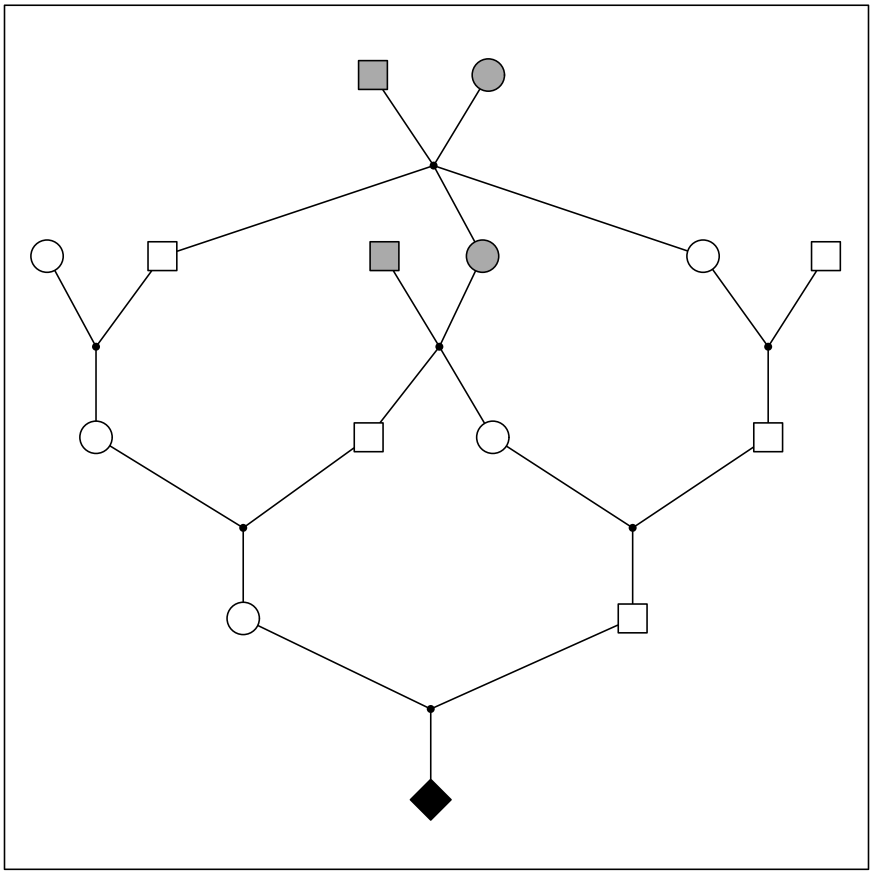

Figure 1: An example pedigree

$\psi(B, C)$, is the probability that homologous genes on gametes segregating from B and from C are *ibd*, while the inbreeding coefficient of an individual B, f_B is the probability that homologous genes on the two gametes uniting to form individual B are *ibd*. Hence

$$f_B \quad = \quad \psi(M_B, F_B)$$

where M_B and F_B are the parents of B. An individual is inbred if his parents are related. He is *autozygous* at a given locus if, at that locus, his two genes are *ibd*; his inbreeding coefficient is the prior probability of this event, based only on the pedigree.

There are (at least) three methods for computing kinship coefficients. The early approach of *path-counting* (Wright, 1922) simply enumerates all the possibilities (in an efficient way). Each path from the individual, B, to common ancestor, A, of his parents, descending via a disjoint set of individuals to B again contributes a term $2^{-(n_M+n_F+1)}(1 + f_A)$ to the inbreeding coefficient f_B, where n_M and n_F are the number of segregations in the maternal and paternal lines of the path. For example, for the offspring of a first cousin marriage, there are 2 paths, one via each of the two grandparents shared by his parents, each having $n_M = n_F = 2$, providing an inbreeding coefficient of $2 \times 2^{-5} = 1/16$. In the complex pedigree of figure 1, the common ancestors of the parents of the final individual are shaded grey. The individual is the offspring of a first cousin marriage, but so also is each of his parents. Here there are two paths via his great-grandparents, each having $n_M = n_F = 2$ as for the simple cousin marrriage, and 3 paths via each of his parents' 2 shared great-grandparents, each with $n_M = n_F = 3$, providing a total inbreeding coefficient of $2 \times 2^{-5} + 2 \times 3 \times 2^{-7} = 7/64$.

Alternative approaches are based on the following equations. A gene segregating from A is with probability $\frac{1}{2}$ his maternal gene (the gene which segregated to him from his mother) and with probability $\frac{1}{2}$ his paternal gene. Thus, provided A is not B nor an ancestor of B

$$\begin{aligned}
\psi(A,\ B) &= \frac{1}{2}(\psi(M_A\,, B) + \psi(F_A\,, B)). \\
\text{Similarly we see} \quad \psi(A, A) &= \frac{1}{2}(1 + \psi(M_A, F_A)) \\
\text{and} \quad \psi(A,\ B) &= \psi(B,\ A) \\
\text{and} \quad \psi(A,\ B) &= 0 \text{ for a founder } A.
\end{aligned}$$

However, the equations need not determine the computational procedure. One way to implement these equations is via a top-down sequential brute-force matrix method, computing kinship coefficients between ancestors arriving finally at the descendant individuals of interest. This is computationally trivial, but expensive on store.

Alternatively, a recursive algorithm based on this equation is very easily implemented, and works well even on large and complex pedigrees. However, it is not necessarily computationally efficient; the same expansion may be repeated many times. In principle, this can be avoided, by saving key $\psi(C,\ D)$, but the simplicity of the method is then lost. All computation is a trade-off between time and store; there is no free lunch.

4.3 Relationships among individuals

For two non-inbred diploid individuals, there are three possible *gene identity states* at a single autosomal locus. That is, the individuals can have both genes *ibd*, or one, or neither. These events have probabilities $(k_2,\ k_1,\ k_0)$ say, $(k_2 + k_1 + k_0 = 1)$, determined by the pedigree. Individuals are related if $k_0 < 1$. Each relationship may thus be represented by a point in an equilateral triangle of unit height, the vertices corresponding to unrelated pairs $(k_0 = 1)$, parent-offspring $(k_1 = 1)$, and the identity (monozygous twin) relationship $(k_2 = 1)$. The kinship coefficient is the probability that homologous genes segregating from two individuals are identical by decent and thus $\psi = (2\ k_2 + k_1)/4$.

While each relationship determines a point $\mathbf{k}$, the converse is not true, not only may several relationships give the same point $\mathbf{k}$, but some points are not (even in the limit) attainable by any relationship. In fact, it can be shown that $k_1^2 \geq 4k_0k_2$. Relationships such as sibs and double-cousins of any given degree fall on the boundary parabola. Note that it is possible for the mother and father of each individual to be related to both the mother and the father of the other, without either individual being inbred. The simplest example is that of quadruple-half-first-cousins. This relationship gives $k_2 = 1/32,\ k_1 = 7/16,\ k_0 = 17/32$ and $\psi = 1/8$. The point in the triangle lies midway between that for half-sibs and for double-first-cousins, which also each have $\psi = 1/8$.

Between inbred individuals there are 15 states of gene identity at a single autosomal locus (Cotterman, 1940). These correspond simply to the number of partitions of the four genes into classes of genes that are *ibd*. Ordering the individuals, and the two genes within each, states can be labelled by labelling the first gene 1, and labelling each successive gene with the same label as any previously labelled gene to which it is *ibd*, and with the next available integer if it is not *ibd* to any previously labelled gene. However, there are only 9 genotypically relevant classes of states, since with regard to genotypes the maternal and paternal origins of genes are irrelevant.

Among larger sets of individuals, the number of possible states of *gibd* increases rapidly (Thompson, 1974). For the 12 genes of 6 individuals, there are more than 4 million gene identity states (partitions of 12 ordered objects). However, there are only just over 198,000 genotypically distinct classes of states. Although this is not a small number, with modern computers and an efficient indexing of state classes it is not impossible to consider all the possible state classes given data on 6 individuals.

4.4 Gene identity at linked loci

The basics of linkage and recombination were introduced by Dr. Elston in an earlier session; for some teminology, see session 12. At linked loci, there are also many more possible gene identity patterns (Denniston, 1975). Some relationships which have identical *gibd* probabilities at a single locus can, in principle, be distinguished by data at linked loci (Thompson, 1986). The simplest example is for the three unilineal ($k_2 = 0$) pairwise relationships of grandmother-granddaughter (G), half-sisters (H), and aunt-niece (N). For such relationships, gene identity at two linked loci is defined by

$$k_{11}(r) \quad = \quad P(share\ 1\ gene\ ibd\ at\ each\ of\ 2\ loci\ recombination\ r)$$

For the three relationships above we have

$$\begin{aligned} G \ &: \ k_{11}(r) \ = \ \frac{1}{2}(1-r) \\ H \ &: \ k_{11}(r) \ = \ \frac{1}{2}(r^2 + (1-r)^2) \ = \ \frac{1}{2}R \\ G \ &: \ k_{11}(r) \ = \ \frac{1}{2}((1-r)R \ + \ r/2) \end{aligned}$$

Thus the relationships are identifiable of the basis of data at two linked loci ($0 < r < \frac{1}{2}$), but not on the basis of data at unlinked loci. All the three relationships have $k_{11}(0) = \frac{1}{2}$ and $k_{11}(\frac{1}{2}) = \frac{1}{4}$.

Some relationships which provide identical two-locus kinship coefficients have different three-locus kinship coefficients (Thompson, 1988). Gene identity at multiple linked loci is a complex process. It is most easily considered by using segregation indicators

$$\begin{aligned} S_{ij} \ &= \ 1 \text{ if segregation } i \text{ locus } j \text{ is grandpaternal} \\ &= \ 0 \text{ if segregation } i \text{ locus } j \text{ is grandmaternal} \end{aligned}$$

The process of $\{S_{ij}\}$ is independent over segregations i, and in the absence of interference is Markov over a sequence of loci j along a chromosome. However, the agglomerated process of identity/non-identity will in general not be Markov along the chromosome (Thompson, 1994a).

Beyond linked loci, one can model a genome continuum, where the segregation indicator $S_i(x)$ is a function of the position x along a chromosome. Under the simplest model of offspring chromosome formation (Haldane, 1919; see also section 12.1) points x at which $S_i(x)$ switches from 0 to 1, or from 1 to 0, occur as a Poisson process. A framework for analysis of genome sharing under this model by different types of relatives was provided by Donnelly (1983).

4.5 Recursive computation of gibd probabilities

The method can be extended to larger number of genes. Karigl (1981) defined kinship coefficients among several individuals, probabilities that genes segregating from each of these individuals are all *ibd*. For example, if B_1 is distinct from, and not ancestral to, $B_2, ..., B_k$,

$$\psi(B_1, ..., B_1, B_2, ..., B_k) \;=\; 2^{1-r}\psi(B_1, B_2, ..., B_k) \;+\; (1-2^{1-r})\psi(M_1, F_1, B_2, ..., B_k)$$

where r is the number of times B_1 is repeated in the lhs, and M_1 and F_1 are the parents of B_1. Karigl (1981) used these multiindividual kinships to determine *gibd* state probabilities between possibly inbred individuals. There are simple linear relationships between any multiple kinship of up to 4 genes segregating from 2 distinct individuals and the state probabilities between four genes.

The idea extends naturally to "descent probabilities", the probability that a set of genes segregating from a set of individuals all descent from a specified set S of founder genes (Thompson 1983; Thompson & Morgan, 1989):

$$\alpha_S(B_1, ..., B_1, B_2, ..., B_k) \;=\; 2^{1-r}\alpha_S(B_1, B_2, ..., B_k) \;+\; (1-2^{1-r})\alpha_S(M_1, F_1, B_2, ..., B_k)$$

The form of the recursive equation is identical to that for ψ, although the boundary conditions of course are slightly different. This is one approach to ancestral inference, but provides only an approximate answer in some rather special cases.

4.6 Descent probabilities and ancestral inference

What one really wants for ancestral inference are posterior probabilities of genotypes given phenotypic data.

$$P(genes \mid data) \;=\; \frac{P(data \mid genes)P(genes)}{P(data)}$$

Leaving aside the problem of computing $P(data)$, consider just the numerator. If the "genes" is a complete specification of genotypes of all pedigree members, the numerator is straightforward.

$$\begin{aligned} P(data \mid genes) &= \prod_{observed\ i} P(data_i \mid genotype_i) \\ P(genes) &= \prod_{founders\ j} P(genotype_j) \prod_{non-founders\ i} P(geno(i) \mid geno(M_i), geno(F_i)) \end{aligned}$$

However, if only a subset of genotypes are specified—for example, genotypes of founders, then others must be summed over to obtain this marginal distribution.

Another very closely related probability is a founder-gene extinction probability; in problems of conservation genetics one may be interested in whether a particular set of founder genes are all extinct in the extant population. Suppose one specified a founder genotype configuration in which all the genes of the set are allele a_1 and all other founder genes are allele a_2. Then the genes of the set are extinct if and only if all extant individuals are genotype a_2a_2:

$$P(specified\ founder\ genes\ extinct) = \\ P(\text{“}data\text{''}\ a_2a_2 \mid founder\ gene\ types\ a_1,\ others\ a_2)$$

4.7 Probabilities of phenotypic observations

In genetic epidemiology, the probability most often required is the likelihood of a genetic model – the probability of phenotypes observed on (a subset of) the members of the pedigree under the linkage or segregation analysis model. This is also the denominator $P(data)$ of the previous section. Where there are data on only a very few individuals, it may be possible to consider the alternative states of gene identity by descent:

$$P(data) = \sum_s P(data \mid s)\ P(gibd\ state\ s)$$

For example, for two individuals the probability of single-locus genotypes under each of the nine states of gene identity are easily tabulated.

However, for more than a very few individuals this is not feasible. Instead, methods that consider the set of underlying genotypes on the pedigree are normally used.

$$P(data) = \sum_g P(data \mid g)\ P(genotypic\ configuration\ g)$$

The question is then how to sum over all possible underlying genotypic configurations. Methods to do this, sometimes known as “peeling”, will be considered in session 8.

As a final note, consider the conditional probability format presented by Robert Elston (the ITU method) which in the current notation is

$$P(B_2 \mid B_1) = \sum_{i=0}^{2} k_i\ P(B_2 \mid B_1, i\ genes\ ibd)$$

In general

$$P(B_2 \mid B_1) = \sum_s P(B_2 \mid B_1, s)\ P(gibd\ state\ s \mid B_1)$$

Note however that $P(s \mid B_1)$ is not the prior state probability $P(s)$; for example, if B_1 is heterozygous states in which he is autozygous are excluded. It is generally much safer to work with joint probabilities

$$P(B_2, B_1) \;=\; \sum_s P(B_2, B_1 \mid s)\; P(s)$$

and to normalise by any marginal probability afterwards.

References for this section

Cotterman (1940)	Denniston (1975)	Donnelly (1983)
Haldane (1919)	Karigl (1981)	Thompson (1974)
Thompson (1983)	Thompson (1986)	Thompson (1988)
Thompson & Morgan (1989)	Thompson (1994a)	Wright (1922)

Session 5. Monte Carlo integration and Monte Carlo likelihood.

5.1 Monte Carlo estimation of expectations.

Where exact probabilities cannot be computed, Monte Carlo simulation is an alternative. Simulation is a method of estimating expectations. If we wish to estimate $\mathrm{E}_P(g(\mathbf{X}))$ then if $X^{(1)}, ..., X^{(N)}$ are simulated from the distribution $P(\cdot)$

$$\frac{1}{N}\sum_{t=1}^{N} \; g(\mathbf{X}^{(t)})$$

is an unbiased estimator of the expectation. Of course, it may not be a very good estimator; in fact it may be a very bad estimator. The art of Monte Carlo is finding good distributions to simulate from, and good ways of simulating from them, in order to get good estimators. A "good" estimator is one with small variance. Note this is not the standard statistical paradigm where we talk about parameter estimation from data: the variances we refer to are Monte Carlo variances.

Note also the difference between a "simulation study" and a "Monte Carlo analysis". Simulation studies are typically undertaken to discover empirically the distribution of a test statistic, or to assess the potential power of a study design. It involves the simulation of data random variables under a model of interest. In a Monte Carlo analysis, integrals, sums, or expectations are estimated by simulating random variables from some distribution, but the random variables are not the data random variables (the data are fixed) and the distribution is simply a tool to provide good estimates of the required expectations. In practice, the difference may be slight. The probability distribution we simulate in a Monte Carlo estimation problem from may often be closely related to the probability model underlying the data in a statistical problem, while the one we use in a simulation study could equally be a convenient tool, with reweighting (see section 5.3) used to adjust the realisations to the distribution of interest. In a Monte Carlo analysis we shall normally simulate conditional on fixed data, but in a simulation study it may sometimes also be desirable to simulate potential data conditional on partial data already obtained.

5.2 Gene-drop Monte Carlo on pedigrees.

The simplest form of Monte Carlo is where we simulate independent, identically distributed (i.i.d.) realisations from some distribution $P(\cdot)$. On pedigrees, the simplest distribution to simulate from is the prior distribution on genotypes, which is done by "gene dropping". Genes are assigned to the founders of the pedigree, segregation of genes down the pedigree is simulated, and the required statistics relating to the resultant current genes are computed.

This approach was used to provide estimates of structural parameters such as inbreeding and kinship coefficients, in one of the very earliest uses of Monte Carlo in pedigree analysis using a digital computer (Edwards, 1967). For each individual, this gives independent Bernoulli trials; the number of trials in which the individual has two *ibd* genes is binomial $B(N, f)$ and the proportion of such trials is an unbiased estimator of f, with variance $f(1-f)/N$. Of course, if we use the same trials to estimate inbreeding coefficients of several related individuals, these estimators are dependent. This may not be a bad thing – if fact it may be very efficient – but the dependence has to be recognised.

Gene-drop has also been used in studying the loss of genes in the pedigrees of endangered species (MacCluer et al., 1986). Exact methods for the computation of probabilities of loss of specified sets of founder genes exist (see sections 4.6, 8.1), but are not always feasible on very complex pedigrees. Moreover, questions of interest relate not only to which founder genes survive in an extant population, but also to how many copies of a given founder gene currently exist. Although means for such random variables are easily computed, the probability distribution is most easily studied by simulating the descent of genes in pedigrees.

Gene-drop can also be used to simulate potential data on pedigrees, and hence to assess the power of a particular pedigree structure with regard to some genetic analysis. Genes can be simulated by gene-drop at all the relevant genetic loci, and then potential data observations simulated given the underlying genes. For each data realisation the required analysis can then be carried out. Gene-drop was used in the earliest assessment of the potential power of a linkage analysis using *elods* (see 12.3) by Thompson et al. (1978). Since an analysis conditional on partial data was required, rejection sampling was used to discard realisations not consistent with these partial data. However, except on very small pedigrees where simulation is often unnecessary in any case, rejection sampling is not an effective method of conditional simulation, since almost all realisations are rejected. We return in section 5.4 to simulation conditional on data.

5.3 Monte Carlo likelihood and importance sampling.

One of the functions we most often need to compute on a pedigree is the likelihood — the probability of the observed data under some genetic model, and often this computation is infeasible. The likelihood on a pedigree can be written

$$L(\theta) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X})\; P_\theta(\mathbf{X})$$

where $\mathbf{X}$ are latent variables (normally the genotypes $\mathbf{G}$). Thus

$$L(\theta) \;=\; \mathrm{E}_\theta(P_\theta(\mathbf{Y} \mid \mathbf{X}))$$

This is the form given by Ott (1979), and in principle we could estimate $L(\theta)$ by simulating $\mathbf{X}$ from the prior genotype distribution under model θ and averaging the value of the penetrance probabilities $P_\theta(\mathbf{Y} \mid \mathbf{X})$ for the realised values of $\mathbf{X}$. This does not work well, except on very small pedigrees, since the realised $\mathbf{X}$ are almost certain to be inconsistent with data $\mathbf{Y}$, or at best to make infinitesimal contribution to the likelihood.

Better ideas normally involve some form of importance sampling. Note that

$$\begin{aligned}
\mathrm{E}_h(g(\mathbf{X})) &= \sum_{\mathbf{X}} g(\mathbf{X})\; h(\mathbf{X}) \\
&= \sum_{\mathbf{X}} g(\mathbf{X}) \frac{h(\mathbf{X})}{h^*(\mathbf{X})}\; h^*(\mathbf{X}) \\
&= \mathrm{E}_{h^*}\left(g(\mathbf{X}) \frac{h(\mathbf{X})}{h^*(\mathbf{X})} \right)
\end{aligned}$$

So we can simulate from $h^*(\cdot)$ and reweight the realisations; this is importance sampling. The effectiveness of this approach depends on the choice of $h^*(\cdot)$. It works best when $g(\mathbf{X})\; h(\mathbf{X})$ is the "same shape" as $h^*(\cdot)$. In fact, it would work best of all if

$h^*(\mathbf{X}) \propto g(\mathbf{X})h(\mathbf{X})$, but this would mean $h^*(\mathbf{X}) = g(\mathbf{X})h(\mathbf{X})/\sum_{\mathbf{X}} g(\mathbf{X})\ h(\mathbf{X})$, and if the denominator were known the Monte Carlo would be pointless! (Hammersley & Handscomb, 1964). The "same shape" is most important in the tails of the distribution; it is a problem if $h^*(\mathbf{X})$ is very small when $g(\mathbf{X})h(\mathbf{X})$ is not. In order to be able to use a given $h^*(\cdot)$ we need first to be able to simulate from it, and second to compute $g(\mathbf{X})h(\mathbf{X})/h^*(\mathbf{X})$ at the realised values of $\mathbf{X}$; this is sometimes far from straightforward!

Now, using the importance sampling idea, we can modify the above likelihood equation to

$$L(\theta) = \mathrm{E}_{\theta_0}(P_\theta(\mathbf{Y} \mid \mathbf{X})\frac{P_\theta(\mathbf{X})}{P_{\theta_0}(\mathbf{X})})$$

This is the version given by K. Lange in Ott (1979); rather than simulating at each θ we simulate at some canonical model θ_0, and average the reweighted penetrance probabilities $P_\theta(\mathbf{Y} \mid \mathbf{X})P_\theta(\mathbf{X})/P_{\theta_0}(\mathbf{X}))$. This estimator introduces one of the key ideas of Monte Carlo likelihood; we estimate $L(\theta)$ as a function of θ from a single set of realisations at one $P_{\theta_0}(\cdot)$. However, except on very small pedigrees, the method does not work, for the same reason as before; the chance that a realised $\mathbf{X}$ is even consistent with the data is infinitesimal.

5.4 Monte Carlo conditional on phenotypes.

Simulation on pedigrees pre-dates digital computers, but the increasing power of computing makes it a more useful proposition. Simulation of the genes descending a pedigree has been used in many studies, and is easily done. However, simulation in the presence of genetic data on current individuals is far harder. The number of possible genotypic configurations on a pedigree is immense, and the proportion that are compatible with data observed on individuals is minute. Normally data are observed on the final members of a pedigree, the current individuals, and the ancestors are unobserved. Simulating the descent of genes from these ancestors to the current individuals will very seldom produce a genotypic configuration compatible with current data — simple-minded rejection sampling is useless.

Since the data are normally at the bottom of a pedigree, it seems natural to attempt to simulate up a pedigree from the current individuals to the founders. In fact, one of the very earliest Monte Carlo pedigree analyses was done this way (Wright and McPhee, 1925). The objective was the estimation of inbreeding coefficients of the extant members of dairy herds, using the (non-computerised!) herd books. This was accomplished by selecting a random one of the two parents at successive generations up the maternal and paternal line, and scoring whether these met. Quite accurate estimates of inbreeding were obtained.

However, in the presence of data, simulation of genes up the pedigree, from the correct distribution given the observed data, is normally infeasible, since all the extant data can affect each step in the realisation, not merely the data on descendants of the individual whose genotype is to be realised. An alternative is to attempt "gene lift", simulating backwards from current data, according to some probabilities for parents conditional upon offspring, and then reweight to the true probabilities under the genetic model, using importance sampling. In principle, this is possible provided we can define sets of parents that are disjoint from their offspring. For human pedigrees this is often possible, since while multiple marriages can occur, a spouse is not normally an offspring of any other member of the spouse set. In animal pedigrees, with cross-generational mating of close relatives, defining the disjoint sets can be more problematic. In any case, *gene-lift* is seldom successful on large pedigrees. If there are multiple alleles, incompatibilities will arise and the simulation cannot proceed. Even for the simulation of a rare recessive allele, for example, where there will be no incompatibility problems, gene lift does not work. The failure of the algorithm to be able to look at distant ancestry, and so bring together the multiple descendant copies of a rare allele, means that realisations of "gene lift" are very far removed from those having non-negligible probability under the genetic model. (However, *gene-lift* may be of use in conjunction with other methods.)

Another approach to simulation conditional on (partial) data is that of Ploughman & Boehnke (1989); it is limited to cases where exact computation is feasible, at least with regard to the genetic loci for which there are data. Assume there are data $\mathbf{Y}$ on some given pedigree members, for a trait determined by a single genetic locus. Then, provided the pedigree is not too complex

$$P(\mathbf{Y}) \;=\; \sum_{\mathbf{G}} P(\mathbf{Y} \mid \mathbf{G})\, P(\mathbf{G})$$

can be computed by the methods to be described in session 8. As a by-product of this computation it will often be possible to obtain the posterior genotype probabilities $P(G_i \mid \mathbf{Y})$ for founders i, and also posterior segregation probabilities $P(G_i \mid G_{M_i}, G_{F_i}, \mathbf{Y}(i))$ where M_i and F_i are the parents of non-founder i, and $\mathbf{Y}(i)$ are the data on that part of the pedigree related to i through his offspring. Given these probabilities, conditional *gene-drop* is possible, assigning genes to founders, conditionally on the data, and then simulating their segregation down the pedigree in accordance with the posterior conditional segregation probabilities. In this way, realisations are obtained of the genotypes at the locus underlying trait data $\mathbf{Y}$. Genes at a linked locus can then easily be simulated, and hence data at that potential marker locus, and the power of a linkage study for the trait on the pedigree, conditional on the extant trait data, can be assessed. This method has become of immense practical importance in linkage studies, but we introduce it here mainly for its similarity to a Monte Carlo method for obtaining linkage likelihoods to be described in section 16.2. Note that its use is restricted to problems for which single-locus likelihoods can be computed on the pedigree. A rather similar method of simulating data at linked loci was provided by Ott (1989). In principle, Ott's method is not so restrictive as that of Ploughman and Boehnke (1989), but similar exact computations must be feasible.

5.5 Markov chain Monte Carlo.

In 5.3 we saw formulae for the likelihood as an expectation. However, those expressions were normally useless in practice because genotypes simulated from the prior will bear almost no relation to the observed data. We need to estimate

$$L(\theta) \;=\; P_\theta(\mathbf{Y}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y},\mathbf{X})$$

Importance sampling considerations dictate that the sampling distribution should be close to proportional to $P(\mathbf{Y},\mathbf{X})$, or as a function of latent variables $\mathbf{X}$ to $P(\mathbf{X} \mid \mathbf{Y})$. Intuitively also, to obtain realisations that have better than infinitesimal probability of giving a non-negligible contribution to the likelihood we must simulate conditional on the data. However

$$P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \;=\; \frac{P_{\theta_0}(\mathbf{Y},\mathbf{X})}{P_{\theta_0}(\mathbf{Y})},$$

and the denominator is unknown: if we could compute $L(\theta_0) = P_{\theta_0}(\mathbf{Y})$ we wouldn't need to be doing Monte Carlo in the first place.

Enter Markov chain Monte Carlo — a method for obtaining realisations from a distribution know only up to a normalising constant. We review briefly the Metropolis-Hastings class of algorithms (Hastings, 1970) for generating dependent realisations from a probability distribution $P_\theta(\mathbf{X})$ on a space $\mathcal{X}$, where $P_\theta(\cdot)$ may be known only up to a normalising constant. For each $\mathbf{X}$ in $\mathcal{X}$ a "proposal distribution" $q(\cdot,\mathbf{X})$ is defined. Then, if the process is now at $\mathbf{X}$ the next value is generated as follows:

1. Generate $\mathbf{X}^*$ from the proposal distribution $q(\cdot,\mathbf{X})$

2. Compute the Hastings ratio

$$h = \frac{q(\mathbf{X}, \mathbf{X}^*) P_\theta(\mathbf{X}^*)}{q(\mathbf{X}^*, \mathbf{X}) P_\theta(\mathbf{X})}$$

Note that h depends only on the ratio of densities $P_\theta(\cdot)$, so that any normalising constant need not be computed.
3. With probability $a = min(1, h)$ the process moves to $\mathbf{X}^*$ and with probability $(1 - a)$ it remains at $\mathbf{X}$.

The algorithm of Metropolis et al. (1953) is a special case; if $q(\mathbf{X}^*, \mathbf{X}) = q(\mathbf{X}, \mathbf{X}^*)$ the Hastings ratio reduces to the odds ratio of the proposal state $\mathbf{X}^*$ versus the current state $\mathbf{X}$. The Gibbs sampler (Geman and Geman, 1984) is also a special case, in which $\mathbf{X}^*$ differs from $\mathbf{X}$ in only one component X_i say, and X_i has the distribution $P_\theta(X_i|\mathbf{X}_{-i})$ where $\mathbf{X}_{-i}$ denotes the components of $\mathbf{X}$ other than X_i. In this case $a = h = 1$; in the Gibbs sampler there is no rejection step, but $\mathbf{X}^* = \mathbf{X}$ is possible, and steps are necessarily small, with at most one component of $\mathbf{X}$ being changed at each step.

In order for the time-average over the chain to converge to the expectation under the equilibrium distribution, the ergodic theorem must apply. For discrete Markov chains, we need irreducibility. However, a lot too much attention is paid to irreducibility in practice. Any chain can be made irreducible, using Metropolis rejection. For example, one might decide that once in a million trials one will propose a new realisation from the prior. Once in a million million realisations one might get something compatible with the data. Once in a million, million, million trials one might get an accepted realisation. Obviously one has done nothing at all in practice, but the sampler is irreducible. Metropolis rejected restarts are often a good idea — one of several key ideas in getting better samplers, and in assessing how good they are. But it has to be done with the practical considerations in mind. Irreducibility *per se* is useless.

There are two (related) sorts of convergence which often get confused. One is convergence of the marginal distribution of each $X^{(t)}$ to the equilibrium distribution of the Markov chain as t becomes large. The other relates to the convergence of a time-average over the chain to the expectation of the function under the equilibrium distribution. Both depend on the mixing properties of the Markov chain, the largest non-unit eigenvalues etc., but whereas the first can (in principle) be addressed by burn-in (discarding enough realisations) it is not normally a problem, in practice. However, for a probability distribution whose modes are unknown, "phase change" problems are always possible: for a genetics example see Thompson (1994d). In this case no amount of burn-in is likely to be sufficient. The second class of questions remain even if we could start in the equilibrium. This is a much bigger problem; we have to sample all the relevant parts of the space. This has to be done with a single sampler (although it could have many component samplers; Geyer, 1991b) Starting runs in different parts of the space is no help if we have no idea how to weight the realisations from the different starts. (See Geyer, 1992).

Estimation of the standard deviations of Monte Carlo estimates of expectations is essential, but hard – not hard to do, but hard to know whether one has a good estimate. Again, Geyer (1992) is a good reference. One of the simplest methods of estimating Monte Carlo variances is by using batch means (Hastings, 1970). One divides the realisations into sufficiently large batches so that the batch means are "almost independent", and relates the variance of independent batch means to the variance of the overall mean (the estimator of the expectation). One can test for autocorrelation between the batches. This is quite effective if the sampler is doing well, but can severely underestimate variance if the sampler is not getting around the space – however the same is true of other variance estimators also, and this one is easy.

Variance estimation also relates to the choice of spacing in sampling realisations from an McMC; the optimal spacing is the one that achieves minimum computational cost for given variance of the resulting estimator. This optial value depends on the relative costs of generating the samples and of evaluating the contribution to the estimator at the realised values, but is seldom large (Geyer 1992; see also Exercises 2, # 2, P.54).

There are the usual sorts of Central Limit Theorems for these estimators.

5.6 Likelihood ratio function estimation

From 5.3 we have,

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

where $\mathbf{X}$ are latent variables (normally the genotypes $\mathbf{G}$). Hence in genetic analysis, or in any missing-data context, we have the key formula

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

Here, $\mathbf{Y}$ are the phenotypes, and $\mathbf{X}$ are latent variables. They can be anything, but it will be useful to think of them as genotypes and/or polygenotypes of individuals.

In the expectation, $\mathbf{X}$ are the random variables, $\mathbf{Y}$ is fixed. The distribution of $\mathbf{X}$ is $P_{\theta_0}(\cdot|\mathbf{Y})$. If we could simulate $\mathbf{X}^{(t)}$, $t = 1, ..., N$, from this distribution then we could estimate the likelihood ratio by

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

Thus, we require realisations from $P_{\theta_0}(\mathbf{X}|\mathbf{Y})$, a distribution known up to the normalising constant $P_{\theta_0}(\mathbf{Y})$; McMC is a possibility. Unfortunately, in genetic examples the constraints on $\mathbf{X}$ imposed by Mendelian segregation mean that any attempted proposal distribution that makes multiple changes to the current value of $\mathbf{X}$ is likely to result in a configuration with zero probability. However, one-variable-at-a-time updating methods, such as the Gibbs sampler, are well suited to this sampling problem, at least in proposing feasible configurations. The genes and heritable effects in an individual are determined by those in his parents, and jointly with those in his spouse, influence those in his offspring. This neighbourhood structure means that the Gibbs sampler is easy to implement; each genetic effect in each individual is successively updated, conditional upon the remainder.

Note that simulation at a single model θ_0 provides an estimate of the relative likelihood $L(\theta)/L(\theta_0)$ as a function of θ, and that this will be a good estimator for those θ close to θ_0. If primary interest is in the shape of the likelihood surface in the neighbourhood of some point (such as the MLE) then preliminary McMC runs and likelihood ratio function estimates can be used to obtain a ballpark estimate. Alternatively, MCEM can be used (see 8.4). Once a ballpark estimate is found, one very large McMC run can provide an accurate estimate of the MLE and of the likelihood in the region. However, this approach has some disadvantages. One may be interested in the likelihood surface, or in log-likelihood differences, over larger regions in the parameter space. Or, the large McMC run may reveal that one's initial estimate was not sufficiently close to the MLE, and additional large runs may be necessary. It is desirable to find a method that combines realisations from all the runs, and provides an estimate of the likelihood surface over a range of parameter values.

One way of doing this was provided by Geyer (1991a); the use of Monte Carlo mixtures. McMC samplers are run at many models, covering the "range" of interest, say at $\theta_0, ..., \theta_K$. The N_j observations from $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$, gives combined set of realisations from

$$P^*(\mathbf{X}) \;=\; \frac{1}{\sum_j N_j} \sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{X}|\mathbf{Y}) = \frac{1}{\sum_j N_j} \sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{Y},\mathbf{X})/L(\theta_j)$$

and writing the likelihood estimation formula as an expectation w.r.t. this $P^*()$

$$L(\theta_j) \;=\; \mathrm{E}_{P^*}\left(\frac{P_{\theta_j}(\mathbf{Y},\mathbf{X})}{P^*(\mathbf{X})}\right)$$

Now, although we have a sample from P^*, the denominator $P^*(\mathbf{X})$ cannot be explicitly computed, since it depends on the unknown $L(\theta_j)$, but we have the implicit Monte Carlo estimating equations

$$L(\theta_j) \;=\; \sum_{\mathbf{X}^*}\left(\frac{P_{\theta_j}(\mathbf{Y},\mathbf{X}^*)}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{Y},\mathbf{X}^*)/L(\theta_l)}\right) \quad \text{for } j = 0, ..., K.$$

where the sum is over the total set of realisations $\mathbf{X}^*$. These equations can be solved iteratively for the relative values of $L(\theta_j)$. For example, keeping one value say $L(\theta_0)$ fixed (=1), and iterating the above equations directly provides global convergence. For any other value of θ in the range spanned by the set of θ_j then $L(\theta)$ (relative to $L(\theta_0) = 1$) can be estimated by

$$L(\theta) \;=\; \sum_{\mathbf{X}^*}\left(\frac{P_{\theta}(\mathbf{Y},\mathbf{X}^*)}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{Y},\mathbf{X}^*)/L(\theta_l)}\right)$$

where the sum is over the same total set of realisations as before. This solves the likelihood estimation problem, provided each sampler ($j = 0, ..., K$) is covering that part of the configuration space that has high total probability mass under the model θ_j ($j = 0, ..., K$).

Even if the separate samplers are behaving “well”, we need good “overlap” between adjacent models, to get the mixtures estimates to work. This means choosing parameter values for adjacent chains not too far apart —but far apart is measured in terms of the probabilities of the genotypic configurations generated. Endpoint distributions can thus be a problem; latent random effects with 0 variance, alleles with 0 frequency etc. Endpoint distributions give 0 probability to (almost all) realisations from other distributions. One solution is exact integration of some latent variables (see 8.6), but this is not always possible. Another possible solution involves redefining the latent variables.

As an example, consider an alternative penetrance version of a sequence of mixed models (see 8.3, 8.4). In this model, a quantitative trait $\mathbf{Y}$ is determined by

$$\mathbf{Y} \;=\; \gamma(\mathbf{f}) \;+\; \boldsymbol{\mu}(\mathbf{G}) \;+\; \mathbf{Z} \;+\; \mathbf{e}$$

incorporating additive effects of covariates $\mathbf{f}$, major genes $\mathbf{G}$, additional random effects $\mathbf{Z}$ (including heritable variation or "polygenic" effects) and independent residuals $\mathbf{e}$. $\mathrm{Var}(\mathbf{Z}) \;=\; \sigma_a^2\mathbf{A}$, where $\mathbf{A}$ is a known matrix determined by the pedigree structure. Assume a fixed mixed model is to be compared with a smaller basepoint model (for example, $\sigma_a^2 = 0$, a major gene model with no polygenic component). We then define a sequence of "non-genetic" models to span the range between the two models. Note we are not interested in the likelihoods of these models; they just form a bridge between two genetic models of interest. We call these non-genetic models, "fractional penetrance" models, as $\mathbf{Y}$ is assumed to result from a fraction λ of the major gene or polygenic effects:

$$\mathbf{Y} \;=\; \gamma(\mathbf{f}) \;+\; \lambda_1\boldsymbol{\mu}(\mathbf{G}) \;+\; \lambda_2\mathbf{Z} \;+\; \mathbf{e}$$

Prior distributions of $\mathbf{G}$ and $\mathbf{Z}$ remain fixed over the sequence of models, the distribution of $\mathbf{e}$ is modified as function of λ_1 and λ_2 to keep variance of Y-components constant, and the collection of models is defined by varying λ_1 and/or λ_2 between 0 and 1. Note that in one sense this is only a reparametrisation of the original model (σ_a^2 becomes $\lambda_2^2\sigma_a^2$), but this rescaling prevents the polygenic effects $\mathbf{Z}$ becoming degenerate at the boundary major gene model. More details are given by Thompson (1994d).

References for this section

Edwards (1967)
Geman & Geman (1984)
Geyer (1991a,b, 1992)
Hammersley & Handscomb (1964)
Hastings (1970)
MacCluer et al. (1986)
Ploughman & Boehnke (1989)
Metropolis et al. (1953)
Ott (1979,1989)
Thompson et al. (1978)
Thompson (1994d)
Wright & McPhee (1925)

Session 8: Likelihood computation on pedigrees

8.1 Peeling as sequential summation.

In writing down the probability of data on a pedigree, it is usually simpler to consider the genotypic configuration determined by Mendelian segregation probabilities, rather than the underlying *gibd* states.

$$P(\mathbf{G}) = \prod_{founders\ i} P(G_i) \prod_{nonfounders\ i} P(G_i|G_{M_i}, G_{F_i})$$

The data on individuals is determined by their underlying genotypes.

$$P(\mathbf{Y}) = \sum_{\mathbf{G}} \left(\prod_{observed\ i} P(Y_i \mid G_i) \right) P(\mathbf{G})$$

Considered as a function of the parameters θ of the genetic model this is the likelihood function $L(\theta)$. There are three classes of parameters. Those indexing the relationship between genotype and phenotype enter into $P(\mathbf{Y} \mid \mathbf{G})$; such are "penetrance parameters" the probability of a specified phenotype given a specified genotype. Parameters such as allele frequencies enter into founder genotype probabilities; these are "population parameters". The third sets of parameters are "segregation parameters" entering into the probability of offspring genotypes given the parental genotypes. These include recombination parameters; the genotypes $\mathbf{G}$ may in principle be multilocus genotypes – a specification of the pair of haplotypes of each individual.

How is the likelihood to be computed? While each term of the product can be easily evaluated, the difficulty is in the sum over $\mathbf{G}$. On a very small pedigree it may be possible to enumerate all possible genotypic configurations $\mathbf{G}$, and to compute the sum directly. In other special cases it may be possible to use a recursive algorithm to compute the gene identity pattern probabilities in the observed individuals, and hence to compute the marginal probability $P(\mathbf{G})$ for these individuals alone. However, in general this is impractical. Independently, Hilden (1970), Elston & Stewart (1971), and Heuch & Li (1972) laid the foundations of the approach that has been widely used over the last 20 years, and has made it possible to compute likelihoods of genetic models given data on large pedigrees. The approach was generalised to complex pedigrees and more complex genetic models by Cannings et al. (1978, 1980). The development presented (but not written out!) here, follows Thompson (1986).

The basic idea is simply one of efficient sequential summation. The number of terms in which a specific G_i, the genotype of individual i, appears is limited to the penetrance term for that individual, and to segregation terms from the parents and to the offspring of individual i. Thus performing a summation over the possible values of G_i results in a function of (at worst) the genotypes of i's parents, spouses and offspring. Of course, this is only useful if implemented sensibly. By starting at the edges (top/bottom/side) of the pedigree, one limits the number of individuals whose genotypes must be considered jointly. For a pedigree without loops, there are (many) sequences of nuclear families such that each is connected to the as yet unprocessed part of the pedigree via a single individual, the *pivot*. In this case, summation over the non-pivot members of each family leads to a function of only the pivot genotypes, which may be incorporated into the summation for that individual in due course. This sequential summation process has come to be known as "peeling", and the specification of the order of individuals (normally of nuclear families) in which summation will be carried out as the "peeling sequence". The summation proposed by Elston & Stewart (1971) was sequential; that of Heuch & Li (1972) was recursive (cf. section 4.2). The methods of Hilden (1970) and Cannings et al. (1978) allowed more general summation operations.

8.2 More alleles, more loci, or more complex pedigrees.

Consider a simple Mendelian model; 1 locus, 2 alleles, 3 genotypes. Due to the structure of $P(\mathbf{G})$ it is possible to compute $P(\mathbf{Y})$ proceeding sequentially through the pedigree using the Elston-Stewart or "peeling" algorithm, provided the pedigree is not too complex. If it has loops, the method can be extended, but if there are multiple interconnecting loops computation becomes infeasible. For k alleles there are $K = k(k+1)/2$ genotypes. The larger k, the sooner peeling becomes computationally infeasible on a complex pedigree.

The procedure is just the same for linked loci. The (multilocus) genotype of an individual is an unordered pair of multilocus haplotypes. That is, it is a specification of not only the single-locus genotypes, but also phase information. The segregation probabilities $P(G_i|G_{M_i}, G_{F_i})$ are functions of the recombination fractions. If there are two diallelic loci, there are 4 haplotypes, and hence 10 genotypes; computation is quite possible for a pedigree without loops, or maybe one loop. With more loci, more alleles, or more interconnecting loops, computation rapidly becomes infeasible.

For a pedigree with loops, functions on the genotypes of a "cutset" of individuals may have to be considered. This is a set of individuals who divide a processed segment of pedigree, from the unprocessed part. The processing therefore results in a function over the set of all possible genotype combinations for the individuals in the cutset. Even for a single autosomal diallelic locus, with 3 possible genotypes for each individuals, there are 3^n potential genotype combinations for n individuals. (In general, K^n, for K genotypes.) In this case, the objective of a good peeling sequence is to limit the cutset sizes as much as possible. Even so, on very complex pedigrees, with multiple intersecting loops, peeling becomes infeasible, particularly if there are more alleles, or more loci. Even for simple pedigrees, peeling becomes infeasible if there are multiple loci with multiple alleles; the number of individual genotypes (pairs of multilocus haplotypes) to be considered can be too large. This relates more to linkage analysis – see later.

8.3: Models with Gaussian random effects

Polygenic models are those in which the discrete genotypes are replaced by Gaussian random effects; the genotype configuration $\mathbf{G}$ becomes a vector of polygenic values $\mathbf{z}$, and sums become integrals. The founder probabilities $P(G_i)$ become the $N(0, \sigma_a^2)$ population densities for z_i, the segregation probabilities $P(G_i | G_{M_i}, G_{F_i})$ become the segregation $N((z_{M_i} + z_{F_i})/2, \sigma_a^2/2)$ densities for z_i given z_{M_i} and z_{F_i}, and the penetrances $P(Y_i \mid G_i)$ (ignoring all other possible fixed/random effects) become the $N(z_i, \sigma_e^2)$ density for Y_i given z_i. Elston and Stewart (1970) noted that, since for Gaussian distribution all marginals and conditional are also Gaussian, and since a Gaussian form is specified by its mean and variance, the peeling process can also be used to compute the likelihood for σ_a^2, and for other parameters, such as an environmental variance σ_e^2. In this case, the sequential summation is just successive integration of latent additive genetic effects. Additional Gaussian latent effects can be incorporated, for example effects of shared environment (Cannings et al., 1980). Also complex pedigrees are no problem, in principle. In fact, the computational process is simpler than for discrete genotypes; for a cutset of n individuals we have a n-variate normal distribution, specified by n means, and $n(n+1)/2$ covariance terms, in place of the K^n discrete genotype combinations. The sequential Elston-Stewart summation method becomes a sequential integration of Normal densities, which is easy.

8.4 The mixed model and MCEM

A model that has attracted much interest is the mixed model (Morton & MacLean, 1974) which has both discrete genotypes and latent Gaussian effects. The mixed model is

$$\mathbf{Y} \ = \ \boldsymbol{\gamma}(\mathbf{f}) \ + \ \boldsymbol{\mu}(\mathbf{G}) \ + \ \mathbf{Z} \ + \ \mathbf{e}$$

and incorporates additive effects of covariates $\mathbf{f}$, major genes $\mathbf{G}$ (including linked markers), additional random effects $\mathbf{Z}$ (including heritable variation (polygenic effects) and non-heritable effects (shared environment)), and independent residuals $\mathbf{e}$. $\text{Var}(\mathbf{Z}) = \sigma_a^2\mathbf{A}$ where $\mathbf{A}$ is a known matrix determined by the pedigree structure.

The likelihood for such a model is a discrete mixture of Gaussians, the number of components being the number of feasible configurations of discrete genotypes on the pedigree. Alternatively, it is a Gaussian mixture of components resulting from peeling the discrete genotypes. There is no way to peel both types of latent genotypes. Various methods of approximating the likelihood have been proposed: the most widely used is that of Hasstedt (1982). For a general discussion of propagation of probabilities on graphs, for both continuous and discrete latent variables, see Lauritzen (1992).

An alternative appaoch is Monte Carlo integration, using Markov chain Monte Carlo to realise $(\mathbf{G}, \mathbf{Z})$. The conditional distribution of latent variables $(\mathbf{G}, \mathbf{Z})$ given data $\mathbf{Y} = \mathbf{y}$ can be simulated from, for given parameter values θ, using the pedigree Gibbs sampler. We have the MCLR formula from the general case (see 5.6):

$$\frac{L(\theta)}{L(\theta_0)} = \int_{\mathbf{Z}} \sum_{\mathbf{G}} \frac{P_\theta(\mathbf{Z}, \mathbf{G}) f_\theta(\mathbf{y}|\mathbf{Z}, \mathbf{G})}{P_{\theta_0}(\mathbf{Z}, \mathbf{G}) f_{\theta_0}(\mathbf{y}|\mathbf{Z}, \mathbf{G})} dP_{\theta_0}(\mathbf{Z}, \mathbf{G}|\mathbf{y})$$

with MC estimate

$$\frac{1}{N} \sum_{t=1}^{N} \frac{P_\theta(\mathbf{Z}^{(t)}, \mathbf{G}^{(t)}) f_\theta(\mathbf{y}|\mathbf{Z}^{(t)}, \mathbf{G}^{(t)})}{P_{\theta_0}(\mathbf{Z}^{(t)}, \mathbf{G}^{(t)}) f_{\theta_0}(\mathbf{y}|\mathbf{Z}^{(t)}, \mathbf{G}^{(t)})}$$

Thompson & Guo (1991) gave the likelihood ratio estimator in this form; see also 8.6 for alternative Monte Carlo estimators of the likelihood ratio for the mixed model.

Note that the likelihood ratio estimator only works well when θ is close to θ_0; only a local estimate of the surface is obtained by simulation at θ_0. However, we can estimate not only the likelihood ratio, but any other expectations conditional upon the observed data $\mathbf{Y} = \mathbf{y}$. In particular we can estimate the expected complete-data log-likelihood, and hence implement a Monte Carlo EM algorithm to find ball-park estimates of the parameters. The likelihood ratio Monte Carlo estimator then enables the likelihood surface in the neighbourhood of this point to be investigated. We present Monte Carlo EM here for the mixed model where it has proven most useful; note however that is can be used analogously for any genetic model where the conditional independence structure is as for this model.

Estimation of θ is a "missing data" problem with unobserved major genotypes $\mathbf{G}$ and polygenotypes (additive genetic effects or breeding values) $\mathbf{Z}$. The likelihood based on the "complete data" $(\mathbf{y}, \mathbf{G}, \mathbf{Z})$ is:

$$
\begin{aligned}
f_\theta\quad(\quad & \mathbf{y}|\mathbf{G},\mathbf{Z})P_\theta(\mathbf{G})f_\theta(\mathbf{Z}) = \\
& c(\theta)h(\mathbf{G})\exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mu_1\mathbf{1}_1-\mu_2\mathbf{1}_2-\mu_3\mathbf{1}_3-\mathbf{Z})'(\mathbf{y}-\mu_1\mathbf{1}_1-\mu_2\mathbf{1}_2-\mu_3\mathbf{1}_3-\mathbf{Z})\right] \\
& \times\exp\left[(2\mathbf{1}_F'\mathbf{1}_1+\mathbf{1}_F'\mathbf{1}_2)\log\frac{p}{1-p}-\frac{1}{2\sigma_a^2}\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}\right]
\end{aligned}
$$

where $c(\theta)$ is a function of θ but not of $\mathbf{G}$ or $\mathbf{Z}$.

Thus the natural sufficient statistics for θ in the complete-data model are $2\mathbf{1}_F'\mathbf{1}_1 + \mathbf{1}_F'\mathbf{1}_2$, $\mathbf{1}_i'(\mathbf{y}-\mathbf{Z})$ $(i=1,2,3)$, $\mathbf{1}_i'\mathbf{1}$ $(i=1,2,3)$, $\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}$ and $\mathbf{e}'\mathbf{e}$, whose unconditional expectations are $2pn_F$, $n_i\mu_i$, n_i, $n\sigma_a^2$ and $n\sigma_e^2$, respectively, where $n_F = \mathbf{1}_F'\mathbf{1}$ is the total number of founders in the pedigree, n_i is the expected number of individuals with genotype i, and n is the total number of individuals in the pedigree. Hence, from the exponential family form of the complete-data likelihood, if we denote new values of parameters by $*$, we obtain the following EM equations for estimating θ:

$$
\begin{aligned}
p^* &= \frac{1}{2n_F}\mathrm{E}_\theta(2\mathbf{1}_F'\mathbf{1}_1+\mathbf{1}_F'\mathbf{1}_2|\mathbf{y}) \\
\mu_i^* &= \frac{\mathrm{E}_\theta\left[\mathbf{1}_i'(\mathbf{y}-\mathbf{Z})|\mathbf{y}\right]}{\mathrm{E}_\theta(\mathbf{1}_i'\mathbf{1}|\mathbf{y})} \qquad (i=1,2,3) \\
\sigma_a^{2*} &= \frac{1}{n}\mathrm{E}_\theta(\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}|\mathbf{y}) \\
\sigma_e^{*2} &= \frac{1}{n}\mathrm{E}_\theta(\mathbf{e}'\mathbf{e}|\mathbf{y})
\end{aligned}
$$

Note that the estimate of allele frequency is simply the expected proportion of genes of given allelic type in founders. Also, if there are missing observations and only k individuals are observed, then n in the equation for σ_e^{*2} should be replaced by k. In this case, $\mathbf{e}$ involves only with those whose trait value is observed.

The expectations in the above equations cannot be computed analytically, but can be estimated by Monte Carlo, running a sampler at one set of parameter values to obtain a large set of realisations, and then updating them with the above equations. This is MCEM: see Guo and Thompson (1992, 1994) for further details.

8.5 Combining peeling and Monte Carlo

There are two main ways in which it may be useful to combine peeling computations with Monte Carlo integration. In the first, the pedigree is divided into parts on which exact computations are feasible, and those for which Monte Carlo may be necessary. In the second approach, the latent variables on each member of the pedigree are divided; exact integration is used for some variables, and Monte Carlo for others.

In the first approach, exact likelihood contributions conditional on genotypes of some individuals can be accumulated from peripheral sections of the pedigree, by peeling. These potentials are then incorporated into the McMC sampling on the pedigree core (Thompson, 1991). The simplest example is where data on a set of final offspring is used to compute a pair potential on the parent couple. In the context of peeling alone, this was used by Thompson (1978b) to simplify probability computations on the ancestral Tristan da Cunha pedigree, by preprocessing data on the extant sibships. A similar technique, in the context of McMC estimation on a very large and complex pedigree, was used by Geyer and Thompson (1995).

A case of particular interest in integrating analytically over some of the latent variables arises for the mixed model:

$$\mathbf{Y} = \boldsymbol{\mu}(\mathbf{G}) \; + \; \mathbf{Z} \; + \; \mathbf{e}$$

(see 8.4; for notational simplicity we have omitted covariate effects $\gamma(\mathbf{f})$). Denoting the combined set of genetic parameters by θ, the likelihood on the basis of the mixed model, given data $\mathbf{Y} = \mathbf{y}$ is

$$L(\theta) = f_\theta(\mathbf{y}) = \sum_{\mathbf{G}} \int_{\mathbf{Z}} f_\theta(\mathbf{y}|\mathbf{G},\mathbf{Z}) P_\theta(\mathbf{G}) f_\theta(\mathbf{Z})\, d\mathbf{Z}$$

The likelihood ratio formula is

$$\frac{L(\theta)}{L(\theta_0)} = \mathrm{E}_{\theta_0}\left(\frac{f_\theta(\mathbf{y}|\mathbf{G},\mathbf{Z}) P_\theta(\mathbf{G}) f_\theta(\mathbf{Z})}{f_{\theta_0}(\mathbf{y}|\mathbf{G},\mathbf{Z}) P_{\theta_0}(\mathbf{G}) f_{\theta_0}(\mathbf{Z})} \;\middle|\; \mathbf{y} \right)$$

with a corresponding Monte Carlo estimator

$$\frac{1}{N} \sum_{l=1}^{N} \left(\frac{f_\theta(\mathbf{y}|\mathbf{G}^{(l)},\mathbf{Z}^{(l)}) P_\theta(\mathbf{G}^{(l)}) f_\theta(\mathbf{Z}^{(l)})}{f_{\theta_0}(\mathbf{y}|\mathbf{G}^{(l)},\mathbf{Z}^{(l)}) P_{\theta_0}(\mathbf{G}^{(l)}) f_{\theta_0}(\mathbf{Z}^{(l)})} \right)$$

Provided the pedigree is not too complex, for given $\mathbf{G}$

$$f_\theta(\mathbf{y}|\mathbf{G}) = \int_{\mathbf{Z}} f_\theta(\mathbf{y}|\mathbf{G},\mathbf{Z}) f_\theta(\mathbf{Z})\, d\mathbf{Z}$$

can be computed since this is just integration of a highly structured multivariate Normal distribution (Elston and Stewart, 1971; section 8.3). Moreover, for given $\mathbf{Z}$

$$f_\theta(\mathbf{y}|\mathbf{Z}) = \sum_{\mathbf{G}} f_\theta(\mathbf{y}|\mathbf{G},\mathbf{Z}) P_\theta(\mathbf{G})$$

can be computed by an analogous procedure for discrete summations (Elston and Stewart, 1971; Cannings et al., 1978; section 8.1). These are the *peeling* computations described previously, and we refer to exact integration of polygenic values as *polygenic peeling*, and to exact summation over major genotype configurations as *major-gene peeling.* The combined integral and sum cannot be performed exactly; the net result would be a mixture of multivariate normal integrals, with as many components to the mixture as there are major-genotypic configurations on the pedigree. (It is of interest that the dependence structure of discrete and continuous variables falls within the framework of Lauritzen (1992) for full exact computation of a likelihood. However, the pattern of variable dependence means that all components of the mixture distribution could have to be separately computed.) The exact computation of either $f_\theta(\mathbf{y}|\mathbf{G})$ or $f_\theta(\mathbf{y}|\mathbf{Z})$ provides two alternative expressions for the likelihood ratio, namely

$$\frac{L(\theta)}{L(\theta_0)} = \mathrm{E}_{\theta_0}\left(\frac{f_\theta(\mathbf{y}|\mathbf{G})P_\theta(\mathbf{G})}{f_{\theta_0}(\mathbf{y}|\mathbf{G})P_{\theta_0}(\mathbf{G})} \;\Big|\; \mathbf{y}\right)$$

and

$$\frac{L(\theta)}{L(\theta_0)} = \mathrm{E}_{\theta_0}\left(\frac{f_\theta(\mathbf{y}|\mathbf{Z})f_\theta(\mathbf{Z})}{f_{\theta_0}(\mathbf{y}|\mathbf{Z})f_{\theta_0}(\mathbf{Z})} \;\Big|\; \mathbf{y}\right)$$

respectively, and hence two additional Monte Carlo estimators:

$$\frac{1}{N}\sum_{l=1}^{N}\left(\frac{f_\theta(\mathbf{y}|\mathbf{G}^{(l)})P_\theta(\mathbf{G}^{(l)})}{f_{\theta_0}(\mathbf{y}|\mathbf{G}^{(l)})P_{\theta_0}(\mathbf{G}^{(l)})}\right)$$

and

$$\frac{1}{N}\sum_{l=1}^{N}\left(\frac{f_\theta(\mathbf{y}|\mathbf{Z}^{(l)})f_\theta(\mathbf{Z}^{(l)})}{f_{\theta_0}(\mathbf{y}|\mathbf{Z}^{(l)})f_{\theta_0}(\mathbf{Z}^{(l)})}\right)$$

To implement these estimates, only one of the sets of realisations $\mathbf{G}$ or $\mathbf{Z}$ would be used, although the Gibbs sampler implementation requires that both are generated exactly as before (unless the pedigree is sufficiently small and simple that peeling is an option in every McMC updating step). The hope is that the reduction in Monte Carlo variance due to the partial exact computation will compensate for increased computation in using each realisation. Details are given by Thompson (1994c).

References for this section

Cannings et al. (1978, 1980) | Elston & Stewart (1971) | Geyer & Thompson (1995)
Guo & Thompson (1992, 1994) | Hasstedt (1982) | Heuch & Li (1972)
Hilden (1970) | Lauritzen (1992) | Morton & MacLean (1974)
Thompson (1978b) | Thompson (1986) | Thompson (1991)
Thompson & Guo (1991) | Thompson (1994c)

Session 10: Ascertainment and the sequential sampling of pedigrees

> The notes for this section are all in the technical report, Thompson (1993), and references therein. A PostScript file of this technical report is available by anonymous ftp. For an overview of the classical sibship ascertainment problem see also George & Elston (1991), while for a discussion of recent papers see Elston (1995) and references therein.

In the course, this session follows an introduction to segregation analysis in independently ascertained sibships, given by Dr. Elston. It was followed by a session including some comments by Dr. Elston on the general ascertainment problem (including a discussion of Vieland & Hodge, 1995), and by Dr. Thomas on conditional Gibbs sampling.

Outline:

- Classical ascertainment in sibships; brief review. The denominator is $(1-(1-p\pi)^s)$ for a segregation probability p, ascertainment probability π, and family size s. The likelihood is conditional both on the sizes and numbers of ascertained sibships. If there is information on the size distribution of segregating families, conditioning on s loses information. If there is information on the number of segregating families (and if π represents a true probability of ascertainment, not merely a representation of relative probabilities), conditioning on the number of ascertained families loses information. *BUT NOT MUCH.*
- Just as all computation is a trade-off of time against store, all statistics is a trade-off of robustness against information. The point is not whether to condition; everything is conditional on the sampling frame of reference. The point is to recognise the level at which we are conditioning, and to assess the consequences in terms of the trade-off.

- The classical framework extended to pedigrees. Computation of the probability that a pedigree contains at least one proband (see section 8.3 of Thompson (1993)); ascertainment probabilities conditional on the individual characteristics of the individuals (age, sex, geographic location). The single ascertainment case, either as infinitesimal probabilities of individual ascertainment, the denominator becoming the expected number of probands in the pedigree, or via age/sex/etc restrictions limiting the possibilities to at most one potential proband per pedigree.
- The sequential framework; the general formulation, as per Cannings & Thompson (1977). Coin tossing examples: 1) The binomial vs negative binomial; 2) retossing the ones that show heads in a homogeneous population; 3) allowing for heterogeneity of the population (of coins).
- Sequential schemes for sibships; the example of Table 3 of the Thompson (1993).
- Single (multiplex) event ascertainment in an infinite population; correction via a sequential scheme in which initial events not meeting the criterion are discarded.
- Robustness and information. The Ewens framework, as presented in Thompson (1993). Examples: 1) Threshold for a quantitative trait, COT vs COP; 2) Linkage analysis, when ascertainment is only through trait values; 3) the example of ascertainment through the two eldest sibs in a sibship of size 3, from Thompson (1993).

References for this section

Cannings & Thompson (1977)   Elston (1995)   George & Elston (1991)
Thompson (1993) and refs therein   Vieland & Hodge (1995)

Session 12; Lod scores and likelihoods for linkage

The basics of linkage and recombination were introduced by Dr. Elston in session 1 of the course. The basic terminology is repeated here. Genes are located on *chromosomes*. Every individual has two of each of the 22 different types of chromosome (excluding the sex chromosomes). In each pair, one is maternal (derining from genetic material in the mother), and the other is paternal. The pair of genes at a particular locus (position on the relevant pair of chromosomes) determine the genotype at that locus. If two loci are close together, the genes will tend to be inherited together. That is, if the offspring receives a copy of the parent's maternal chromosome at one locus is will do so also at the other. Likewise the paternal. The *recombination frequency* between the two loci measures the frequency with which the offspring's genes from the same parent at the two loci are copied from different parental chromosomes, and hence derive from different grandparents (see the segregation indicators introduced in section 4.4).

12.1 Map distance and map functions

We consider first the problem of map distance; note we are saying nothing whatever about physical distance. Haldane (1919) defined map distance as the expected number of crossovers (in a 2-strand model). Map distance is thus additive along the chromosome, and an odd number of crossovers results in a recombination. In the absence of interference, crossovers occur as a Poisson process (rate 1/Morgan) along the chromosome. The relationship between recombination frequency (r) and map distance (x) is then easily derived:

$$
\begin{aligned}
r(x + \delta x) &= r(x)(1 - \delta x) + (1 - r(x))\delta x \\
\frac{\mathrm{d}r}{\mathrm{d}x} &= (1 - 2r(x)) \\
\log_e(1-2r) &= -2x \quad (r(0) = 0) \\
r(x) &= \frac{1}{2}(1-\exp(-2x)) \\
r(0) = 0, \qquad & r(\infty) = \frac{1}{2} \text{ and } r(x) \approx x \text{ for small } x
\end{aligned}
$$

This is Haldane's mapping function. For a 4-strand model with no interference, only one half of the gametic products of meiosis are involved in any given crossover between non-sister strands, and $r(x) = \frac{1}{2}(1-\exp(-x))$.

In practice interference occurs. Consider loci A, B and C. Map distances must satisfy

$$x_{AC} = x_{AB} + x_{BC}$$

and in the absence of interference

$$r_{AC} \;=\; r_{AB}(1-r_{BC}) \;+\; r_{BC}(1-rAB) \;=\; r_{AB} \;+\; r_{AC} \;-\; 2r_{AB}r_{BC}$$

Muller (1916) defined the *coefficient of coincidence* as

$$c \;=\; \frac{P(recombination\ in\ both\ AB\ and\ BC)}{r_{AB}r_{BC}}$$

in which case

$$r_{AC} \;=\; r_{AB} \;+\; r_{AC} \;-\; 2c\, r_{AB}r_{BC}$$

The Kosambi (1944) map function takes $c = r_{AC}$ which results in the relationship

$$r_{AC} \;=\; \frac{r_{AB}+r_{BC}}{1+4r_{AB}r_{BC}} \quad \text{and} \quad r(x) \;=\; \frac{1}{2}\left(\frac{e^{4x}-1}{e^{4x}+1}\right)$$

This map function has been much used, but in common with many of the early map functions is not multilocus feasible; that is, it can result in negative probabilities of gametic outcomes when more than 3 loci are involved. In order to obtain a multilocus feasible map, it is important to assume some model for the crossover process. The simplest such model is Haldane's model of the absence of interference.

12.2 Linkage analysis; the basic counting problem.

In linkage analysis, the aim is to locate genes contributing to a trait, by analysing the evidence for cosegregation of the trait with known genetic markers within a pedigree. A linkage likelihood is no different from any other likelihood for a genetic model, except in that the parameters to be estimated are those of segregation of genes from parents to offspring.

Figure 2 shows a simple example in which complete information is available. There are three genetic loci, labelled A, B, and C, each with three corresponding alleles a_i, b_i and c_i, $i = 1, 2, 3$. 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 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 $a_1b_2c_2$

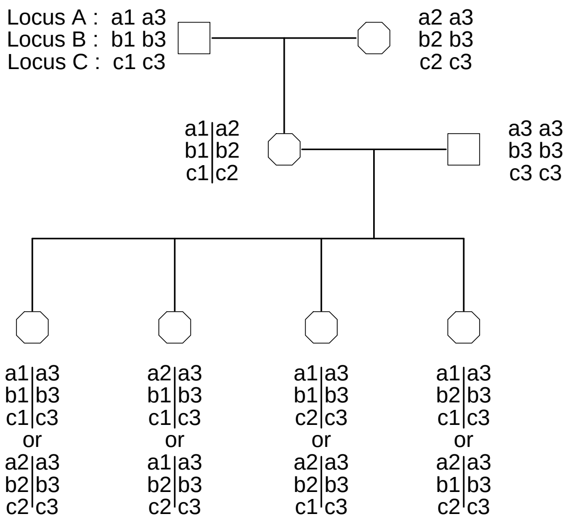

Figure 2: Segregation of genes at three linked loci.

(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 ($a_1b_2c_1$ for example) will have much smaller frequency, since these can arise only as a result of recombination in both intervals (AB and BC).

In such an example, linkage estimation and testing is simply a matter of counting. Indeed, 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 made the first *gene ordering* inference, by methods analogous to those still used today. 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. Fisher (1922) 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.

12.3 Lods and Elods

Linkage analysis in human genetics did not start until the 1930s, 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 2). Haldane (1934) and Fisher (1934) furthered the likelihood approach — the approach of computing the probability of the observed data, under the probability model. Haldane (1934) derived maximum likelihood estimates of recombination fractions, in the context of linkage detection. Fisher (1934) distinguished clearly between the problem of linkage detection, for which he cited Haldane (1934) as providing the likelihood approach, and estimation, in which the precision of estimation is the issue. Thus he used Fisher Information as his measure of informativeness of family data.

For linkage detection, the relevant statistic is the *lod* score, which is just the log-likelihood ratio, for linkage at a specified set of recombination fractions relative to the alternative of no linkage (free recombination; $r = 0.5$. For historical reasons, the logs in lod scores are taken to base 10 rather than base e (Morton, 1955). If there are other model parameters, the lod score is also a function of these. The *lod* score is well defined for $0 \leq r \leq 1$, but is often only plotted for $0 \leq r \leq \frac{1}{2}$, since these are the biologically relevant values. If there is no (implicit) information on parental phase, then the *lod* score is symmetric about $r = 0.5$; it can sometimes be useful to plot the *lod* score over the entire range in order to assess the information on phase.

Recombination affects the probabilities of offspring genotypes conditional upon their parents', but observations are made at the phenotypic level. For traits of incomplete penetrance, or other more complex traits, there may be other parameters in the likelihood or *lod* score. Ott (1991) defines

$$Z_2 = \log_{10}\left(\frac{L(\hat{\theta},\hat{r})}{L(\hat{\theta}, r=0.5)}\right)$$

and

$$Z_1 = \log_{10}\left(\frac{L(\hat{\theta},\hat{r})}{L(\theta^*, r=0.5)}\right)$$

where here θ denotes the parameters other than the recombination frequency (penetrances etc.), and $\hat{\theta}$ is the maximum likelihood estimate, while θ^* is the MLE in the absence of linkage. Note that $Z_2 \geq Z_1$. Z_1 is the log-likelihood test statistic that has (perhaps) some theoretically justifiable properties. Usually Z_2 is easier to compute, since terms involving the trait probability distribution may cancel from the ratio. Usually also however, it is easy to adjust Z_2 to Z_1 since

$$Z_2 - Z_1 = \log_{10}\left(\frac{L(\theta^*, r = 0.5)}{L(\hat{\theta}, r = 0.5)}\right)$$

which does not involve marker data or recombination frequencies.

For details on practical linkage analysis, and on the currently available programs, see Terwilliger & Ott (1994). Haldane & Smith (1947) and Smith (1953) also provide excellent accounts of the classical development of linkage analysis approaches and procedures, and studies of their properties.

For linkage detection, a useful measure of information is given by an expected log-likelihood difference. In the case of linkage analysis, the expected lod score is precisely this; an *elod*. Who originated *elods* as a measure of information for linkage analysis?— C.A.B. Smith or E. Sturt, maybe, but I am still seeking an exact reference. Thompson et al.(1978) used *elods* in this way; interestingly they estimated them by Monte Carlo — *gene-drop* Monte Carlo with rejection.

In designing a linkage study, normally phenotypic data on a trait are already available, and the objective is to consider the power to detect linkage, conditional on these data, were data at a linked marker to be obtained. A method of obtaining such a conditional *elod* was developed by Ploughman & Boehnke (1989) (see section 5.4). This method is immensely useful and has become very widely used in applied linkage studies, but it does not provide a general method for simulation conditional upon data. Note only must the single-locus computations be feasible (see 5.4), but it is restricted to situations where all the data relating to one locus are already available (to be conditioned upon) and none of the data at the second (potential marker) locus.

References for this section

Fisher (1922)	Fisher (1934)	Haldane (1919)
Haldane (1934)	Haldane & Smith (1947)	Morton (1955)
Ott (1991) and see also Ott (1991) for other refs; Kosambi (1944), Muller (1916) etc.		
Ploughman & Boehnke (1989)	Smith (1953)	Sturtevant (1913)
Terwilliger & Ott (1994)	Thompson et al. (1978)	

Session 13; Multipoint mapping, & other complications

13.1 Location scores and multipoint computations

Although linkage is normally detected through pairwise analysis of a trait with each marker, once linkage is detected, estimation of the trait locus position using multipoint analysis is often also desirable. Different marker loci are informative in different families, or in different segregations within families, and estimation within the constraints of a known marker map allows the location of the trait locus to be pinpointed more precisely (provided the map is correct and appropriate marker allele frequencies are used). With the introduction of the earliest RFLP DNA markers (Botstein et al., 1980), and the new forms of DNA markers developed since then, the potential for obtaining multiple marker loci in most regions of the genome has become almost unlimited.

With the increasing availability of marker loci, 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") computed for each hypothesised location of the trait locus relative to the hypothesis that no trait locus is linked to this segment of the the marker map. Specifically, let $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_T)$ denote the data for markers (M) and trait (T), and $\mathbf{G} = (\mathbf{G}_M, \mathbf{G}_T)$ the underlying genotypes. Further let θ_T denote the parameters of the trait model, $\mathbf{x}_M$ the map positions of the markers and x_T the map position of the hypothesized trait locus. The map positions $\mathbf{x} = (\mathbf{x}_M, x_T)$ determine the recombination fractions, for example by using Haldane's map function (section 12.1). The total set of parameters is $\theta = (\theta_T, \mathbf{x})$, but if θ_T and the marker map $\mathbf{x}_M$ are held fixed the location score is

$$V(x_T) = 2\log_e\left[\frac{L(\theta_T,\ x_T,\ \mathbf{x}_M)}{L(\theta_T,\ -,\ \mathbf{x}_M)}\right]$$

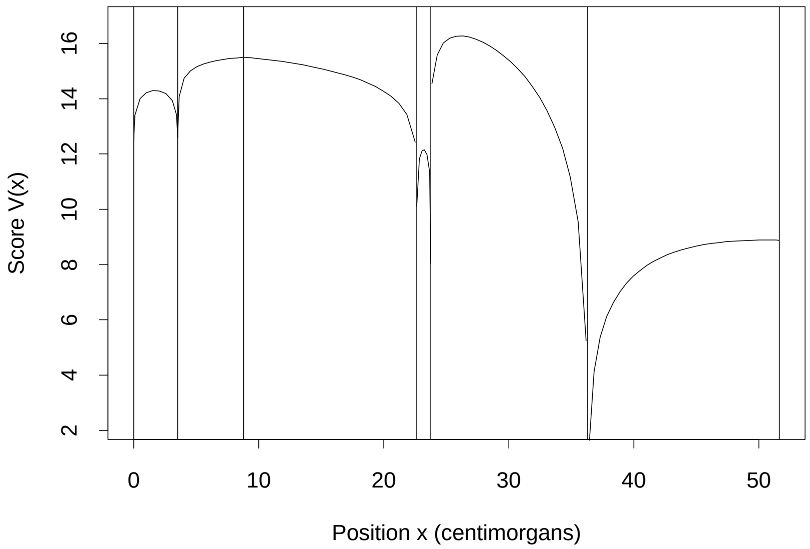

Figure 3: Multipoint linkage curve, based on a graph of Irwin et al. (1994). Vertical lines designate positions of marker loci.

Note also that the likelihood is

$$\begin{aligned} P_\theta(\mathbf{Y}) &= \sum_{\mathbf{G}} P_\theta(\mathbf{Y} \mid \mathbf{G}) P_{\mathbf{x}}(\mathbf{G}) \\ &= \sum_{(\mathbf{G}_M, \mathbf{G}_T)} P_{\theta_T}(\mathbf{Y}_T \mid \mathbf{G}_Y) P(\mathbf{Y}_M \mid \mathbf{G}_M) P_{x_T}(\mathbf{G}_T \mid \mathbf{G}_M) P_{\mathbf{x}_M}(\mathbf{G}_M) \end{aligned}$$

The varying parameter x_T enters only into the third component of each term in the sum. This will be important when we consider Monte Carlo estimation of this sum in section 16.2. As in section 12.3, it may be desirable to use different estimates of the trait model in the numerator and denominator, if the linkage information has a substantial impact on unknown trait model parameters. As in section 12.3, adjustment to a different trait model under absence of linkage in straightforward, involving only the trait models and trait data. If a different trait model is estimated for each trait locus position, evaluation of location scores can become very computationally intensive.

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 3). 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.

There have been a number of programs written to compute multipoint linkage likelihoods. The most widely used are the LINKAGE programs (Lathrop et al., 1984) which use a form of peeling to compute exact likelihoods. However, with multiallelic markers and multiple loci, there are large numbers of possible multilocus genotypes for each individual, and computation becomes infeasible. There have been recent improvements to the programs in the form of a new package FASTLINK (Cottingham et al., 1993), but for complex pedigrees exact computation remains infeasible for more than a very few loci. Lander & Green (1987) adopted a different approach to the exact computation of multilocus linkage likelihoods. Instead of using multilocus genotypes as the latent variables of a peeling algorithm, they used segregation indicators (see section 4.4), and their sequential summation was over loci rather than over individuals. However, the approach was designed for simple pedigrees with no missing data, such as are used in map construction or for mapping in experimental populations. The method could not deal with extended pedigrees or extensive missing data.

13.2 Heterogeneity

There are many types of heterogeneity in genetic linkage. The first is that of inhomogeneity of crossover rates along the physical chromosome. This type of inhomogeneity, and the possibility of recombination hotspots, does not affect linkage analysis *per se* since there is no assumed relayionship between map distance (which determines recombination frequencies) and physical distance. However, inhomogeneities of the physical map can greatly complicate the discovery of particular genes even when the linkage map is well established.

The second type of heterogeneity is that between males and females. Overall the females human genetic map is about 30% longer than the male map. That is, crossovers occur at roughly 30% greater rate in the formation of gametes segregating from females. However, this relationship is not uniform over the genome, but itself displays heterogeneity. In many practical linkage analyses, male and female recombination rates are separately estimated, or some relationship between them is assumed. This issue is discussed fully by Ott (1991). From experimental populations, it is well known that many environmental factors also influence recombination rates.

Of much greater practical importance in modern linkage analysis is locus heterogeneity of the genes underlying a trait. That is, the same phenotype may be caused, in different families (or even within a single extended family), by alleles at different loci. This was first addressed in human genetics by Fisher (1936). He considered linkage of Friedriech's Ataxia with the ABO blood groups. Only one family showed any evidence of recombination, but in this family there were 4 obligatory recombinants among six offspring, negating tight linkage. Fisher developed a test for heterogeneity, but cautioned against overinterpretation of results on this small data set. Smith (1953) developed a model framework for locus heterogeneity, which has been further developed by Ott (1991). In this model, there is a prior probability α that a family will be of the "linked type" and probability $(1-\alpha)$ that the locus causing the trait in that family is unlinked to the markers. The overall likelihood, based on the data on all the available families, is then computed both as a function of α and the linked trait locus recombination frequency (or position) relative to the marker (or marker map). The heterogeneity parameter may be estimated, or a test of $\alpha < 1$ may be performed. The method provides also a posterior probability, for each family, of whether it is of the "linked type". Further details are given by Ott (1991).

13.3 Other complications; some brief notes

For linkage analysis with a single marker, model misspecification or other aberrations (sporadic cases, pedigree errors, etc.) will often decrease the power to detect linkage, and increase the estimated recombination frequency, by creating apparent recombinants. However, with enough data, the linkage will still be detected provided the correct general mechanism is assumed (Clerget-Darpoux et al., 1986; Williamson and Amos, 1990)

However, if multipoint linkage analysis is attempted, with a fixed map distance between the markers, evidence for the locus at any point in an interval may disappear; one distance from marker to trait locus cannot be increased, without decreasing the distance to the other. Erroneous marker allele frequencies can also cause wither false positives or false negatives in linkage analysis. This is also an increasing problem with new DNA markers for which the available population samples may be small, or derived from individuals from a different population.

The lod score, as above, is

$$\log_{10}\left(\frac{L(r,\theta_0)}{L(\frac{1}{2},\theta_0)}\right)$$

for a known trait model θ_0. The wrod score is

$$\log_{10}\left(\frac{L(r,\theta)}{L(\frac{1}{2},\theta)}\right)$$

where the model θ may be wrong. The mod score is the maximised wrod score

$$max_\theta\left[\log_{10}\left(\frac{L(r,\theta)}{L(\frac{1}{2},\theta)}\right)\right]$$

Of course, maximising a lod score, or log-likelihood-ratio, does not provide equivalent properties as maximising a log-likelihood. Maximising a lod score to assess the trait model θ is also a possible procedure. For further discussion, see Elston (1989, 1994) and references therein.

Although interference exists, it is normally ignored in linkage analyses; its effect on linkage likelihoods is small, and the power to detect interference from data on human pedigrees is slight. The effect of interference is to decrease the number of multiply recombinant gametes. In practice, it is often found that apparent recombinants do cluster in certain gametes or in certain families. This is often the result of marker typing errors, or linkage heterogeneity among families. Restricting the number of implied recombination events within a gamete can be a useful way to detect such typing errors or heterogeneity, which have a far greater impact than interference phenomena at the genetic distances relevant to linkage mapping.

13.4 Multiple tests and prior probability of linkage

Given a disease or other trait of interest, the purpose of linkage analysis is to locate the genes contributing to the trait. This may be done by testing for linkage with known markers for candidate genes; for example, in a family in which combined familial hyperlipidemia is apparently segregating, one will test for linkage to DNA markers for the Low Density Liproprotein (LDL) receptor locus. More often, however, a genome scan is undertaken; the trait is tested for linkage with each of a large number of DNA marker loci, spanning the genome. The problem of multiple tests then arises. Further, since the DNA markers are themselves linked, the tests are dependent, and assessment of results is therefore not straightforward.

In this assessment, the prior probability of linkage is also relevant. Haldane & Smith (1947) first considered this aspect of mapping a simple Mendelian trait, adopting a Bayesian approach to determine the posterior probability of linkage. Where such a simple trait is at issue, then the locus controlling it is somewhere in the genome, and the prior probability of linkage to some marker locus increases as the number of loci tested increases, perhaps compensating to some degree for the multiple testing. This argument presupposes however that there is a locus for the trait; for complex traits this may not be clear. For complex traits, apparent cosegregation of the trait with the alleles of some marker somewhere in the genome should be treated with caution. Validation of the result on an independent set of families is highly desirable.

However, for complex traits, validation of a linkage is also problematic. Where alleles at several loci contribute to the trait, either in different populations, or in different families within a population, or in different parts of the same extended pedigree, analysis of new data often may not confirm the putative linkage result. Often genome scans may result in false positive indications of linkage, but equally it may never be possible to validate some true positives by additional linkage analyses. Eventually, discovery of the gene is the validation.

See Elston (1994) and Ott (1991) for further discussion.

References for this section

Botstein et al. (1980)	Clerget-Darpoux et al. (1986)	Cottingham et al. (1993)
Elston(1989, 1994)	Fisher (1936)	Haldane (1919)
Haldane & Smith (1947)	Lander & Green (1987)	Lathrop et al. (1984)
Ott (1991)	Williamson & Amos (1990)	

Session 16: Modern Monte Carlo methods of linkage analysis

16.1 Speeding up McMC samplers for multiallelic markers

Since computation of linkage likelihoods on large or complex pedigrees, for highly polymorphic markers, with lots of missing data, is computationally demanding, or sometimes quite infeasible, Monte Carlo provides an obvious alternative. The essence of the approach is as previously; we consider the problem as a missing data problem, and use Monte Carlo integration over the latent variables to obtain estimates of a likelihood ratio, lod score, or location score curve.

The first Monte Carlo approaches to linkage likelihood estimation used the multilocus genotypes as the latent variables. For certain data configurations, the single-site Gibbs sampler is not irreducible when a locus is multiallelic. However, theoretical irreducibility can always be easily achieved; the problem is practical irreducibility which can be a problem even for diallelic loci on large and complex pedigrees, particularly if underlying genotypes are highly constrained (but not determined) by the data. Several methods for more efficient sampling of the space of feasible underlying genotype configurations have been developed. Some of these are due to Shili Lin (Lin et al. 1994 a,b). Others are due to Eric Sobel (Sobel & Lange, 1993), and to Charles Geyer (Geyer & Thompson, 1995). We briefly outline here only the first methods, directed specifically towards sampling of genotypes at polymorphic marker loci where there are many unsampled individuals in the pedigree.

Although there may be multiple possible alleles at a particular DNA marker locus, only the ones actually segregating in the observed pedigrees are of interest. Further, for linkage analysis it is the pattern of segregation, not the specific marker genotypes, that are relevant. It is often therefore possible to recode the marker alleles, and to reduce the number that must be considered, in such a way that no linkage information is lost. Care must be taken in the analysis of such recoded data, especially where different recodings are used in different pedigrees of the data set, specifically with regard to population allele frequencies, which relate of course to the true allele designations, not to the recoded form. Errors in population alleles frequencies can lead to spurious linkage results. Although highly polymorphic markers complicate computation, they provide more information for linkage because the segregation paths of the alleles can be more readily inferred. Such markers are also very useful in the detection of pedigree errors; pedigree errors can lead to failure to detect true linkage.

The reducibility of the Gibbs sampler for genetic loci with more than two alleles was first addressed by Sheehan & Thomas (1993). Their method used modification of either the segregation probabilities or the penetrance probabilities, so that the sampler was no longer irreducible. Because of the form of the modification used, no reweighting was required in order for the realisations to represent the distribution of genotypes under the true genetic model. All realisations consistent with the true model required equal weight; those inconsistent with it are just dropped from the output sample. Lin et al. (1993) used similar penetrance modifications to achieve irreducibility, but used Metropolis-coupled samplers (Geyer, 1991b), coupling a sampler under the true model to samplers which were not only irreducible, but also moved more quickly around the space. Rather than a uniform penetrance modification, only the changes necessary to achieve irreducibility are made. The expansion of the space that is sampled is therefore limited; Lin refers to these limited additional configurations, feasible under the modified genetic model but not under the true one, as providing "bridges" between the state classes of the Gibbs sampler under the true model. Lin (1994a) showed how characterisations of the irreducible state classes under the true sampler can often be found. In this case much smaller and more effective changes to the true model can be made to achieve irreducibility. Only the penetrances for a few key individuals must be modified; the "bridges" are much narrower..

The methods of Lin et al. (1993, 1994 a,b) also use a form of "heated proposals", resulting in samplers that move around the space of genotypic configurations far more effectively. Recall that for the Gibbs sampler the proposal distribution is

$$\begin{aligned} q_i(\mathbf{X}^*;\mathbf{X}) &= P_\theta(\mathbf{X}_i \mid \mathbf{X}_{-i}, \mathbf{Y}) \text{ for component } i \\ X_j^* &= X_j \text{ for } j \neq i \quad \text{or } \mathbf{X}^*_{-i} \; = \; \mathbf{X}_{-i} \end{aligned}$$

and this sampler is widely used in genetic problems since this conditional distribution is easily computed, whereas probabilities relating to larger proposed changes to the genotypic configuration on a pedigree are normally infeasible. Thus, one possibility is to based a Metropolis-Hastings sampler on this local conditional, but in a way that makes movement from higher-probability states more probable. The method of Lin et al. (1994b) proposes

$$\begin{aligned} q_i(\mathbf{X}^*;\mathbf{X}) &\propto (P_\theta(X_i^* \mid \mathbf{X}_{-i}, \mathbf{Y}))^{1/T} \text{ for component } i \\ X_j^* &= X_j \text{ for } j \neq i \quad \text{or } \mathbf{X}^*_{-i} \; = \; \mathbf{X}_{-i} \end{aligned}$$

The Hastings ratio is then

$$\begin{aligned} h &= \frac{q(\mathbf{X};\mathbf{X}^*)P_\theta(\mathbf{X}^* \mid \mathbf{Y})}{q(\mathbf{X}^*;\mathbf{X})P_\theta(\mathbf{X} \mid \mathbf{Y})} = \frac{(P_\theta(X_i \mid \mathbf{X}^*_{-i}, \mathbf{Y}))^{1/T} P_\theta(\mathbf{X}^* \mid \mathbf{Y})}{(P_\theta(X_i^* \mid \mathbf{X}_{-i}, \mathbf{Y}))^{1/T} P_\theta(\mathbf{X} \mid \mathbf{Y})} \\ &= \frac{(P_\theta(\mathbf{X} \mid \mathbf{Y}))^{1/T} P_\theta(\mathbf{X}^* \mid \mathbf{Y})(P_\theta(\mathbf{X}_{-i} \mid Y))^{1/T}}{(P_\theta(\mathbf{X}^* \mid \mathbf{Y}))^{1/T} P_\theta(\mathbf{X} \mid \mathbf{Y})(P_\theta(\mathbf{X}^*_{-i} \mid Y))^{1/T}} \\ &= \frac{(P_\theta(\mathbf{X}^* \mid \mathbf{Y}))^{1-1/T}}{(P_\theta(\mathbf{X} \mid , \mathbf{Y}))^{1-1/T}} = \frac{(P_\theta(X_i^* \mid \mathbf{X}_{-i}, \mathbf{Y}))^{1-1/T}}{(P_\theta(X_i \mid \mathbf{X}_{-i}, \mathbf{Y}))^{1-1/T}} \end{aligned}$$

Heating increases the probability of moving from high probability states, but it increases the probability of staying in low-probability states. So how can heating work? The sampler needs to stay in relatively low-probability states long enough for genotypic changes to pervade the pedigree. Without heating, the sampler will often move onto a "bridge", but immediately move back off.

16.2: Three approaches to Monte Carlo location score curves

The method of Monte Carlo estimation of linkage likelihoods or location scores that follows most directly from session 5 is that of Thompson & Guo (1991). The underlying genotypes for both trait and marker are sampled from their conditional distribution given the trait and marker data, using the Gibbs sampler, or some improved sampler based on the Gibbs sampler, updating the genotypes individual by individual. An estimate of a ratio of location scores is provided by the likelihood ratio formula of section 5.6, which in the notation of section 13.1 becomes

$$\frac{L(\theta_T,\ x_{T_1},\ \mathbf{x}_M)}{L(\theta_T,\ x_{T_0},\ \mathbf{x}_M)} = \mathrm{E}_{\theta_0}\left(\frac{P_{\theta_1}(\mathbf{Y}_T,\mathbf{Y}_M \mid \mathbf{G}_T,\mathbf{G}_M)P_{\theta_1}(\mathbf{G}_T,\mathbf{G}_M)}{P_{\theta_0}(\mathbf{Y}_T,\mathbf{Y}_M \mid \mathbf{G}_T,\mathbf{G}_M)P_{\theta_0}(\mathbf{G}_T,\mathbf{G}_M)} \;\middle|\; \mathbf{Y}_T,\mathbf{Y}_M\right)$$

for two trait locus positions x_{T_1} and x_{T_0} and $\theta_i = (\theta_T, x_{T_i}, \mathbf{x}_M)$. Noting the partition also given in session 13.1, and the fact that only the position of the trait locus differs between numerator and denominator, the above equation reduces to

$$\frac{L(\theta_T,\ x_{T_1},\ \mathbf{x}_M)}{L(\theta_T,\ x_{T_0},\ \mathbf{x}_M)} = \mathrm{E}_{\theta_0}\left(\frac{P_{x_{T_1}}(\mathbf{G}_T \mid \mathbf{G}_M)}{P_{x_{T_0}}(\mathbf{G}_T \mid \mathbf{G}_M)} \;\middle|\; \mathbf{Y}_T,\mathbf{Y}_M\right)$$

Thus only the conditional probability of trait genotype given marker genotypes apears explicitly in the estimator; computation of this term is generally very straightforward.

This procedure provides an estimator of the local relative location score; simulation at a single x_{T_0} provides the curve as a function of neighbouring x_{T_1}. To obtain estimates of the relative values for trait locations in distinct marker intervals, or an estimate of the absolute level of the location score curve (relative to the basepoint where there is no trait locus in this region of the marker map) the method of Monte Carlo mixtures of section 5.6 can be used. We note that the location score as given here, and in section 13.1, is for a fixed trait model θ_T. Adjustment to alternative trait models, dependent on trait locus position, is possible. The simplified likelihood ratio estimator involving only $P_{x_T}(\mathbf{G}_T \mid \mathbf{G}_M)$ requires the same θ_T in the numerator and denominator, but the general formula above does not.

The methods of Thompson & Guo (1991) developed further by Lin et al. (1994 a,b) focus on estimation of likelihood ratios, sampling multilocus genotypes. The approach of Lange and Sobel (1991) differs in two ways. First the underlying state space is augmented by variables that specify grandparental origins of genes. This enables samplers that make larger incremental changes in the underlying state to be implemented, but it also makes the space to be sampled even larger. The other key difference is in the formulation of their location score Monte Carlo estimate.

Lange & Sobel (1991) formulation is

$$\begin{aligned} P(\mathbf{Y}_T \mid \mathbf{Y}_M) &= \sum_{\mathbf{S}} P(\mathbf{Y}_T \mid \mathbf{G}_M(\mathbf{S}))P(\mathbf{S} \mid \mathbf{Y}_M) \\ &= \mathrm{E}(\ P(\mathbf{Y}_T \mid \mathbf{G}_M(\mathbf{S})) \mid \mathbf{Y}_M) \end{aligned}$$

Here $\mathbf{G}_M$ is marker genotype configuration, and $\mathbf{S}$ includes in addition the parental origin indicators of genes. Hence $P(\mathbf{Y}_T \mid \mathbf{Y}_M)$ is estimated by Monte Carlo, using samples $\mathbf{S}$ from $P(\mathbf{S} \mid \mathbf{Y}_M)$. The location score is precisely $2\ \log_e\ (P(\mathbf{Y}_T|\mathbf{Y}_M)/P(\mathbf{Y}_T))$ computed for various possible locations of the trait locus. This approach has a possible advantage in that the conditional probability $P(\mathbf{Y}_T \mid \mathbf{Y}_M)$ is estimated directly, rather than a likelihood ratio. Also sampling is conditional only on marker data, and only marker genotypes (and parental origins) are realised. Trait genotypes $\mathbf{G}_T$ are implicit in the evaluation of $P(\mathbf{Y}_T \mid \mathbf{G}_M(\mathbf{S}))$ but do not have to be considered explicitly.

While McMC seems an obvious approach, it is not the only Monte Carlo possibility, and in some cases an alternative method may be much more efficient. One approach is that of sequential imputation (Kong et al., 1994). The development for the case of multilocus linkage likelihoods is given by Irwin et al. (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 X_l the underlying genotypes at that locus. A realisation X_l^* is obtained for each locus in turn from the distribution

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

and predictive weights w_i are accumulated:

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

It is then readily shown that the joint simulation distribution for $\mathbf{X}^* = (X_1^*, ..., X_m^*)$ is

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

and thus

$$\mathrm{E}_{P^*}(w_m(\mathbf{X}^*)) \ = \ \sum_{\mathbf{X}^*} w_m(\mathbf{X}^*)P^*(\mathbf{X}^*) \ = \ P_{\theta_0}(\mathbf{Y})$$

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.

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:

$$\begin{aligned} \mathrm{E}_{\theta_0}(g(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y}) &= \sum_{X} g(\mathbf{X},\mathbf{Y}) P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \\ &= \sum_{\mathbf{X}} g(\mathbf{X},\mathbf{Y}) P^*(\mathbf{X}) w_m(\mathbf{X}) / P_{\theta_0}(\mathbf{Y}) \\ &= \mathrm{E}_{P^*}(g(\mathbf{X},\mathbf{Y}) w_m(\mathbf{X})) / P_{\theta_0}(\mathbf{Y}) \end{aligned}$$

However, the normalising factor $P_{\theta_0}(\mathbf{Y})$ is the unknown likelihood. Although (4) provides a Monte Carlo estimate of $P_{\theta_0}(\mathbf{Y})$ so that

$$\mathrm{E}_{\theta_0}(g(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y}) = \mathrm{E}_{P^*}(g(\mathbf{X},\mathbf{Y}) w_m(\mathbf{X})) / \mathrm{E}_{P^*}(w_m(\mathbf{X})),$$

we then a have ratio estimate, with its attendant problems.

McMC avoids this problem since realisations are directly (in the limit) from $P_{\theta_0}(\cdot \mid \mathbf{Y})$, but it has other disadvantages. In summary:

Sequential imputation

samples from approximately the right distribution

McMC

samples approximately from the right distribution

Provided the single locus computations can be performed, and provided there are not too many missing data, sequential imputation may often give superior performance to McMC. However, in some cases even the single-locus computations may be demanding (for highly polymorphic markers) or infeasible (on complex pedigrees). Moreover, if there is a high proportion of missing marker phenotypes, the imputation distribution may diverge from the true conditional distribution of genotypes given phenotypes, and the importance weights may have high variance. The problem of the large space of latent variables remains.

Table summary of the 3 Monte Carlo methods for multipoint linkage.

	Thomson & Guo 1991	Kong & Irwin 1991**	Lange & Sobel 1991
pedigree	*arbitrary*	*simple**	*simple**
simulation	McMC	i.i.d.	McMC
on space	$(\mathbf{G}_M, \mathbf{G}_T)$	$(\mathbf{G}_M, \mathbf{G}_T)$	$(\mathbf{G}_M +)$
use of	$P(\mathbf{G}_T \mid \mathbf{G}_M)$ ratio	predictive weights	$P(T \mid \mathbf{G}_M^+)$
realisations	easy	w_m, intensive	intermediate
function estimator	Yes (local)	No	Yes (global)

Note *: simple here means a pedigree on which single-locus likelihoods can be computed very rapidly. In practice, this normally means any zero-loop pedigree.

**: U. Chicago 1991 technical report; now published as Irwin et al. (1994).

Thompson (1995) provides a review of these and other aspects of modern linkage analysis; sections 12.2, 12.3, 13.1 and 16.2 are based on Thompson (1995). Stephens & Smith (1993) provide another McMC version of multipoint linkage analysis, using a Bayesian Gibbs sampler approach.

16.3 Use of segregation indicators as latent variables.

In some cases, it may be possible to reduce the space of latent variables significantly, by using the segregation indicators rather than genotypes as the latent variables. In the Monte Carlo likelihood formulation of section 5.3, the latent variables $\mathbf{X}$ may be chosen in any way, provided only $P_\theta(\mathbf{Y} \mid \mathbf{X})\, P_\theta(\mathbf{X})$ can be easily computed both at the simulation model, in order to implement the chosen Metropolis-Hastings algorithm, and at any other model whose relative likelihood in required, in order to compute the contribution to the Monte Carlo likelihood ratio estimator. As seen in section 4.4, segregation indicators provide an efficient determination of *gibd* patterns among a set of individuals. Provided the probability of phenotypic data, for a given *gibd* pattern can be very rapidly determined, there is no need to consider latent genotypes. For data on a large number of individuals, this is infeasible; there are too many *gibd* patterns, and, more importantly, too many possible ways to assign allelic types to genes in a manner consistent with the phenotypic data (Thompson, 1974).

The approach used by Thompson (1994b) is to introduce a segregation indicator, as in section 4.4, for each locus and for each segregation of the pedigree, indicating whether the gene in that segregation at that locus is of grandpaternal (1) or of grandmaternal (0) origin. A Metropolis sampler is used; a segregation indicator is chosen at random, and it is proposed to change it from 0 to 1, or from 1 to 0. The probability of this segregation indicator given all others is straightforward. Different segregations are independent, and the indicator at a given locus depends only on those at the neighboring ones. Moreover, for data on a very few individuals (≤ 6 say), the probability of observed phenotypes, given underlying segregation indicators is straightforward, and hence sampling of the segregation indicators conditional on the observed data is also straightforward (Thompson, 1994b). The sampler is irreducible; the reducibility problems caused by multiallelic markers are avoided.

This method was used by Thompson (1994a) to obtain linkage likelihoods for homozygosity mapping of a rare recessive on small but complex pedigrees. The pedigree of figure 1 (section 4.2) is one of these; although this pedigree is small it defeats exact computation for more than a very few loci. Homozygosity mapping (Smith, 1953) rests on the fact that an inbred individual affected by a rare recessive trait is likely to be so due to receiving two genes *ibd* from a recent common ancestor of his two parents. Then also, genes at tightly linked marker loci will also be *ibd* following the same segregation paths, and hence the affected individual will be homozygous, since *ibd* genes must be of the same allelic type. Hence patches of homozygosity in the genome, shared by unrelated individuals having the same rare recessive trait, provide evidence for linkage. In effect, this is simply linkage analysis, where the only data available are on the affected inbred individuals. Because only one individual is observed, it is very much more effective to use as the latent variables the segregation indicators from common ancestors of the parents of the individual, rather than the multilocus genotypes of the many unobserved ancestors.

On a large pedigree, the space of segregation indicators may be large – there is one indicator for every segregation and for every locus – but it is very much smaller than the space of all possible multilocus genotypes. The space may also be constrained by the data, and singe-site Metropolis updating may not provide an efficient sampling of the space, but the constraints on movement are less than for genotypes. For genotypes, any change must maintain consistency with Mendelian segregation; no individual anywhere in the pedigree can have an allele that neither parent has. For the segregation indicators, the only constraint on consistency between indicators and data is that it must be possible for the genes *ibd* in the observed individuals to be of the same allelic type.

References for this section

Geyer & Thompson (1995)	Irwin et al (1994)	Kong at al. (1994)
Lange and Sobel (1991)	Lin et al. (1993, 1994a,b)	Sheehan & Thomas (1993)
Sobel & Lange (1993)	Smith (1953)	Stephens & Smith (1993)
Thompson (1974)	Thompson (1994a,b, 1995)	Thompson & Guo (1991)

17. Detection of genes for complex traits: gene identity in multiple affected individuals at multiple loci.

We have seen how patterns of gene identity by descent underlie the joint distribution of traits in sets of relatives (sections 4.1, 4.7). We have seen how gene identity by descent at linked loci (section 4.4) underlies linkage analysis (section 12.2). We have seen how inferred gene identity at disease loci and linked markers underlie linkage detection methods using affected sib pairs, affected inbred individuals, or other pairs or sets of affected relatives (Elston session #14; see also 16.3). With the increasing availability of data on a dense map of markers, genome sharing in sets of affected relatives can be detected, and may provide methods for the detection of genes contributing to susceptibility to complex traits.

In order to increase power, by avoiding genetic heterogeneity within the sample, relatives should be considered. In order to increase the significance of observed patterns of genome sharing, remote relatives and multiple relatives are preferred. To increase informativity of marker data w.r.t. underlying gene identity by descent, multiple linked markers are important. We are therefore interested in patterns of gene identity by descent in sets of several affected relatives, at several linked loci, conditional on marker data. How many is several?

Well maybe six affected relatives? This will depend on the trait, and what is available. However, note that each additional relative (provided not too closely related) provides greatly increased information about patterns of genome sharing. Note also that the number of possible patterns of genome sharing at a single locus escalate with the number of individuals (Thompson, 1974). For more than six individuals (with 194,000+ patterns, although few will have even non-zero probability), it becomes possible only to look at extreme patterns (e.g. all the individuals share a gene) which will likely cause relevant genes to be missed.

Maybe 3 or 4 loci? If we are doing a genome scan, marker loci will of necessity be linked, but there is no point having very closely linked markers. The purpose of having multiple linked loci is to distinguish more clearly between identity by state and identity by descent. Provided markers are sufficiently polymorphic and sufficiently immutable, no additional information on identity by descent in that genome region is provided by additional markers.

How can we analyze patterns of gene identity by descent conditional on marker data at several loci, in several individuals. There has been some recent theoretical work (Feingold, 1993; Guo 1994), using continuous genome models (Haldane, 1919; Fisher 1949), first considered in the pedigree analysis context by Donnelly (1983). However, analytical results are limited. An obvious alternative is McMC, used for this purpose by Thompson (1994a). The segregation (grandparental origin) indicators introduced earlier, as a way of tracing genes on pedigrees, are easily sampled by McMC even conditionally on data (Thompson, 1994b; see also 16.3.

References for this section

Donnelly (1983)	Guo (1994)	Feingold (1993)
Fisher (1949)	Haldane (1919)	Thompson (1974, 1994a,b)

20. A note on disequilibrium mapping

Genetic linkage analysis provides limited resolution of loci, due to the limited number of segregations that can be observed within a pedigree or set of pedigrees. The solution in experimental populations is the use of recombinant inbred lines, where repeated backcrossing, maintaining heterozygosity at specific loci, allows the effects of multiple segregations in separating very tightly linked loci to be observed directly. The human geneticist does not have this option, but the effects of multiple segregations can be observed through residual allelic associations in an extant population. Where a new variant allele arises in conjunction with a very closely linked marker allele, the association will be maintained for many generations. Incomplete association of the variant with closely linked markers provides evidence for locus order. The multiple opportunities for recombination over the history of the variant allow for finer resolution in detecting the location of the variant locus against a map of very closely linked markers.

Note first that random mating of individuals is equivalent to random union of gametes. Under random mating, two individuals of certain (multilocus) genotypyes come together in a mating, to produce an offspring of some genotype in accordance with Mendelian segregation probabilities (and recombination probabilities). Under random union of gametes, two gametes of specified haplotype come together in an individual, who will in turn produce gametes of some haplotype in accordance with

Mendelian segregation probabilities (and recombination probabilities). For most population genetic purposes it is sufficient (and easier) to consider the frequencies of gametes over successive generations, rather than of genotypes.

Let q_{AB}, q_{Ab}, q_{aB} and q_{ab} be the gametic frequencies at some generation, and r the recombination frequencies between the two diallelic loci, having alleles A and a, and B and b, respectively. It then follows under random union of gametes that the frequency q^*_{AB} at the next generation satisfies

$$\begin{aligned} q^*_{AB} &= (q_{AB})^2 + 2q_{AB}q_{Ab}\frac{1}{2} + 2q_{AB}q_{aB}\frac{1}{2} + 2q_{AB}q_{ab}\frac{1-r}{2} + 2q_{Ab}q_{aB}\frac{r}{2} \\ &= q_{AB} - rD \end{aligned}$$

where $D = (q_{AB}q_{ab} - q_{Ab}q_{aB})$. Similarly

$$q^*_{ab} = q_{ab} - rD \text{ while } q^*_{Ab} = q_{Ab} + rD \text{ and } q^*_{aB} = q_{aB} + rD$$

and hence

$$D^* \;=\; (q^*_{AB}q^*_{ab} \;-\; q^*_{Ab}q^*_{aB}) \;=\; (1-r)D$$

and after t generations an initial allelic association caused by a new mutation arising on a particular genetic background has decayed by a factor of $(1-r)^t$.

This equation conveys the essence of the historical information available in allelic associations. Edwards (1980 & earlier) provides discussions of the half-lives of haplotypes based on this equation, and shows that for $r \leq 0.002$ associations are expected to remain for several thousand years. More importantly, these papers also address the question of the number of segregations in the ancestral coalescent of a current sample of extant copies of a given allele (Thompson (1978a) and refs therein). This provides a measure of the likely length of chromosome seqment that will remain in association with a variant allele.

However, in the 1980s, allelic associations in the context of linkage was thought of in the context of equilibrium results, under which a standardised measure of squared disequilibrium is expected to be proportional to $1/(1+4Nr)$, where r is the recombination frequency between the loci, and N a measure of effective population size (the same for all loci). Several papers used this result in attempts to order tightly linked loci (see e.g. Chakravarti et al., 1984). This is not a useful framework for such analyses, as numerous theoretical papers have shown. Even if a population equilibrium obtains (which it never does), there may be almost no power to detect linkage disequilibrium, and what power there is is highly dependent on the allele frequencies at the loci in question.

Nonetheless, "haplotype analysis" the analysis of allelic associations between a disease allele at markers tightly linked markers has continued to be widely used, and has been quite helpful in providing more precise locations of disease loci, the best known eample being that of cystic fibrosis (Kerem et al., 1989). This success if obviously attributable to the non-equilibrium historical perspective presented by Edwards (1980).

Recently there has been much interest in this approach in the context of genetic isolates such as the Finnish population (Hästbacka, et al. (1992), Jorde, (1995)), in which a disease may be allelically homogeneous, and recent, and the population has been relatively isolated since the presumed origin of the disease allele.

Studies by Thompson & Neel (1995) in the context of Amerindian populations have shown that the vast majority of extant allelic variants postdate population expansion due to agricultural development, and thus in evolutionary terms are "young". It is thus not clear that an isolated population is necessary for such analyses, although allelic homogeneity is. Kaplan et al. (1995) provide a clear discussion of many of these issues.

References for this section

Chakravarti et al. (1984)
Edwards, (1980)
Hästbacka, et al. (1992),
Jorde (1995)
Kaplan et al., (1995)
Kerem et al. (1989)
Thompson E. A. (1978a)
Thompson & Neel (1995)

EXERCISES; from the Wageningen Course, 1994

(excluding exercises based on the compute software)

Exercises 1: Monte Carlo on pedigrees.

1) An individual in a pedigree has inbreeding f. This is to be estimated by gene drop through the pedigree. Suppose in fact f=0.109375. How many trials are needed to estimate f to within 0.001.

2) There are two individuals of interest in the pedigree, and an estimate of $(f_1 + f_2)/2$ is required. Suppose in fact $f_1 = 0.11$ and $f_2 = 0.0625$. To avoid problems of dependence you decide to use n_1 trials to estimate f_1 and n_2 trials to estimate f_2. How should you choose n_1 relative to n_2, if you know that f_1 and f_2 are small, but f_2 is probably approximately $f_1/2$.

3) You decide to use the same realisations on the pedigree to estimate both f_1 and f_2, and hence $(f_1 + f_2)/2$. Suppose each realisation now takes 1.5 times as long as when simulating for each of A and B. Suppose also that the probability A and B are both inbred at a given locus is (a) $f_2/2$ (b) $f_1 \times f_2$

4) You have a data set consisting of two separate pedigrees. You estimate likelihoods L_1 and L_2, by a Monte Carlo average, in two ways. First you do not realise that the pedigrees are separate, so you simulate genotypes simultaneously on both. Then you do realise, and use separate simulations to estimate each likelihood. Which method is more effective?

Exercises 2

1. Increasing efficiency by exact integration?

Let (X_i, Y_i), $i = 1, ..., N$ be independently and identically distributed pairs of random variables, with probability distribution $P(\cdot)$. It is desired to use Monte Carlo to estimate $\mathrm{E}_P(g(X,Y))$. One estimator is

$$T_N = \frac{1}{N}\sum_{i=1}^{N} g(X_i, Y_i)$$

Suppose it is possible to compute $\gamma(Y) = \mathrm{E}_P(g(X,Y) \mid Y)$; another estimator is then

$$S_N = \frac{1}{N}\sum_{i=1}^{N} \gamma(Y_i)$$

Show $var(S_N) < var(T_N)$.

Suppose it costs c_0 to generate each pair (X_i, Y_i) (and the same if only Y_i is to be generated). Suppose it costs c_1 to evaluate $g(X,Y)$, and $c_2 > c_1$ to evaluate $\gamma(Y)$.

Discuss the computational efficiencies of the estimators T and S depending on the relative values of c_0, c_1, and c_2 and the variances of $g(X, Y)$ and $\gamma(Y)$.

2. Spacing of realisations in an McMC chain

It is sometimes argued that realisations from an McMC should be taken "many steps apart, to reduce dependency". Suppose it is desired to estimate $\mathrm{E}_P(g(X)$ where now the X_i, $i = 1, ..., N$ are generated from a Markov chain. Suppose it costs c_0 to generate each successive X_i, and c_1 to compute $g(X_i)$. Discuss the computational efficiency of wide spacing of realisations, assuming convergence to the equilibrium distribution has already been achieved. See Geyer (1992).
Discuss also the point of the assumption.

3. A "counter-example" to the efficiency of integrating.

If the successive (X_i, Y_i) are dependent, the result that integrating decreases the variance (see # 1 of exercises 2) may not hold. The details of the following example are due to Geyer. Let $Y_i|X_i$ be normal $N(\rho X_i, (1 - \rho^2))$, and let $X_{i+1}|Y_i$ be normal $N(\rho Y_i, (1-\rho^2))$. You may find it easier to think of an AR(1) time series $\{Z_t; t = 1, ...\}$ with AR parameter ρ, with $X_t = Z_{2t-1}$ and $Y_t = Z_{2t}$.

Let $g(X, Y) = X - bY$, so $\gamma(Y) = (\rho - b)Y$. If you are familiar with time series, or very persevering, you can show, for large N, in previous notation

$$Nvar(T_N) = (1 + b^2) - 2\rho b + 2\frac{\rho}{(1 - \rho^2)}\Big((1 + b^2)\rho - b(1 + \rho^2)\Big)$$

and

$$Nvar(S_N) = (\rho - b)^2\frac{1 + \rho^2}{1 - \rho^2}$$

Conclude that it is not always true that $var(S_N) < var(T_N)$; for example, take $\rho = 0.8$ and $b = 2.5$.

(For those who can't face the time series, but don't want to take the result for granted, the result can also be shown for $N = 2$.)

Bibliography

Botstein, D., White, R. L., Skolnick, M. H., and Davis, R. W. (1980). Construction of a linkage map in man using restriction fragment polymorphism. *Amer. J. Jum. Genet.* **32:** 314–331.

Cannings, C., & Thompson, E. A. (1977). Ascertainment in the sequential sampling of pedigrees. *Clinical Genetics* **12:** 208–212.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Adv. Appl. Prob.* **10:** 26–61.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1980). Pedigree analysis of complex models. In *Current Developments in Anthropological Genetics* (J. Mielke and M. Crawford, eds.). Plenum Press, New York (pp. 251-298).

Chakravarti, A., Buetow K.H., Antonarkis S.E., Waber P.G., Boehm, C.D., and Kazazian, H. H. (1984). Nonuniform recombination within the human β-globin gene cluster. *American Journal of Human Genetics* **36:** 139-1258.

Clerget-Darpoux, F., Bonaiti-Pellie, C., & Hochez, J. (1986). Effects of Misspecifying Genetic Parameters in Lod Score Analysis. *Biometrics* **42:** 393–399.

Cotterman C. W. (1940) *A Calculus for Statistico-Genetics.* Ph.D. Thesis, Ohio State University. Published in *Genetics and Social Structure.* P. A. Ballonoff (ed). Academic Press, 1974: New York.

Cottingham, R. W., Idury, R. M., and Schäffer, A. A. (1993) Faster sequential genetic linkage computations. *American Journal of Human Genetics* **53:** 252–263.

Denniston, C. (1975) Probability and genetic relationship: two loci. *Annals of Human genetics* **39:** 89–104.

Donnelly, K. P. (1983) The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biolology* **23:** 34–64.

Edwards, A.W.F. (1967) Automatic construction of genealogies from phenotypic information (AUTOKIN). *Bulletin of the European Society of Human Genetics* **1,** 42–43.

Edwards, J.H. (1980) Allellic association in man. Allelic association in man. In, Population Structure and Genetic Disorders, A. W. Erikson (ed.). Academic Press; New York. Pp. 239–256. (Figure 3; P. 246)

Elston, R.C. & Stewart, J. (1971) A general model for the genetic analysis of pedigree data. *Human Heredity* **21:** 523–542.

Elston, R. C. (1989) Man bites dog? The validity of maximizing lod scores to determine the mode of inheritance. *American Journal of Medical Genetics* **34:** 487–488.

Elston, R. C. (1994). P-values, power, and pitfalls in the linkage analysis of psychiatric disorders. In *Genetic Approaches to Mental Disorders.* Gerson E.S. & Cloninger C.R. (eds) American Psychiatric Press, Inc. Washington DC. Pp. 3–21.

Elston, R. C. (1995). 'Twixt cup and lip: how intractable is the ascertainment problem. *American Journal of Human Genetics* **56:** 15–17.

Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *Journal of Applied Probability* **30:** 766-779.

Fisher, R. A. (1922b). On the systematic location of genes by means of crossover observations. *Amer. Nat.* **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. *Ann. Eugen.* **6:** 66–70.

Fisher, R. A. (1936). Heterogeneity of linkage data for Friedrich's ataxia and the spontaneous antigens. *Ann. Eugen.* **7:** 17–21.

Geman, S. & Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6:** 721–741.

George, V. T., & Elston, R. C. (1991) Ascertainment: an overview of the classical segregation analysis model for independent sibships. *Biometric Journal* **6:** 741–753.

Geyer, C. J. (1991a) Reweighting Monte Carlo Mixtures. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C. J. (1991b) Markov chain Monte Carlo maximum likelihood. *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Pp 156-163. Interface Foundation of North America, Fairfax Station, VA.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7** 473–511.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. In press for 9/95; *Journal of the American Statistical Association*

Guo S.W. and Thompson, E. A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *Amer. J. Hum. Genet.* **51:** 1111-1126.

Guo, S. W. and Thompson, E. A. (1994) Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics:* **50:** 417-432.

Guo, S.W. (1994) Computation of identity by descent proportions shared by two siblings. *American Journal of Human Genetics* **54:** 1104–1109.

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 linkage 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 & Co., London.

Hasstedt, S.J. (1982) A mixed-model likelihood approximation on large pedigrees. *Computers in Biomedical Research,* **15,** 295–307.

Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., & Lander E. S. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* **2:** 204–211.

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

Heuch, I. & Li, F. M. H (1972). PEDIG—A computer program for calculation of genotype probabilities, using phenotypic information. *Clinical Genetics* **3:** 501–504.

Hilden, J. (1970). GENEX—An algebraic approach to pedigree probability calculus. *Clinical Genetics* **1:** 319–348.

Irwin, M., Cox, N., and Kong, A. (1994). Sequential imputation for multilocus linkage analysis. *Proc. Natl. Acad. Sci. USA* **91:** 11684–11688.

Jorde, L. (1995) Linkage disequilibrium as a genetic mapping tool. *Am J Hum Genet* **56:** 11–14.

Kaplan, N. L., Hill, W. G. & Weir, B. S. (1995) Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet* **56:** 18–32.

Karigl, G. (1981). A recursive algorithm for the calculation of gene identity coefficients. *Annals of Human Genetics* **45:** 299–305.

Kerem B., Rommens, J. M., Buchanana, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M. et al. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science* **245:** 1073–1080.

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. *Proceedings of the National Academy of Sciences (USA)* **84** 2363–2367.

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

Lathrop, G.M., Lalouel, J-M., Julier, C. & Ott, J. (1984) Strategies for multilocus linkage analysis in humans. *Proc.Natl.Acad.Sci. USA.* **81:** 3443–6.

Lauritzen, S. J. (1992) Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association* **87:** 1098–1108.

Lin, S., Thompson, E. A., and Wijsman, E. M. (1993) Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *I.M.A. J. Math. Appl. Med. & Biol.,* **10:** 1–17.

Lin, S., Thompson, E. A., and Wijsman, E. M. (1994) Finding non-communicating sets for Markov chain Monte Carlo estimations on pedigrees. *Amer. J. Hum. Genet.,* **54:** 695–704.

Lin, S., Thompson, E. A. and Wijsman, E. M. (1994) An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees. *Ann. Hum. Genet.* **58:** 343–357.

MacCluer, J.W., VandeBerg, J.L., Read, B. & Ryder, O.A. (1986) Pedigree Analysis by Computer Simulation. *Zoo Biology* **5:** 147–160.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. 1953 Equations of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21,** 1087–1092.

Morton, N.E. (1955) Sequential tests for the detection of linkage. *Amer. J. Hum. Genet.* **7:** 277–318.

Morton, N. E., & MacLean, C. J. (1974) Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. *American Journal of Human Genetics* **26:** 489–503.

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. (1989) Computer simulation methods in linkage analysis. *Proc. Natl. Acad. Sci. USA* **86:** 4175–4178.

Ott, J. (1991) *Analysis of Human Genetic Linkage.* (2 nd. ed.) The Johns Hopkins University Press; Baltimore, MD.

Ploughman, L. M., & Boehnke, M. (1989) Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44:** 543–551.

Sheehan, N. A. and Thomas, A. W. (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics:* **49:** 163–175.

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

Sobel, E., and Lange, K. (1993) Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* **2:** 263–282.

Stephens, D.A. & Smith, A. F. M. (1993) Bayesian inference in multipoint gene mapping. *Annals of Human Genetics* **57:** 65–82.

Sturtevant, A. H. (1913) The linear association of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool* **14:** 43–59.

Terwilliger, J. D. & Ott J. (1994). *Handbook of Human Genetic Linkage.* The Johns Hopkins University Press.

Thompson, E. A. (1974) Gene identities and multiple relationships. *Biometrics* **30:** 667–680.

Thompson E. A. (1978a) The number of ancestral haplotypes contributing to a sample of B8 alleles. *Nature* **272:** 288.

Thompson, E. A. (1978b) Ancestral inference II: The founders of Tristan da Cunha. *Annals of Human Genetics* **42:** 239–253.

Thompson, E.A., Kravitz, K., Hill, J. & Skolnick, M.H. (1978) Linkage and the power of a pedigree structure. In *Genetic Epidemiology.* N.E. Morton (ed.), pp. 247–253. Academic Press, New York.

Thompson, E.A. (1983). A recursive algorithm for inferring gene origins. *Ann. Hum. Genet.* **47:** 143–152

Thompson, E.A. (1986). *Pedigree Analysis in Human Genetics.* The Johns Hopkins University Press, Baltimore, MD.

Thompson E.A. (1988). Two-locus and three-locus gene identity by descent in pedigrees. *I.M.A. J. Math. Appl. in Med and Biol.* **5:** 261–280.

Thompson E.A. and Morgan K. (1989). Recursive descent probabilities for rare recessive lethals. *Ann. Hum. Genet.,* **53:** 357-374.

Thompson E. A. (1991) Probabilities on complex pedigrees; the Gibbs sampler approach. *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* Pp 321-328. 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 Journal of Mathematics Applied in Medicine & Biology* **8:** 149–169.

Thompson, E. A. (1993). Sampling and ascertainment in genetic epidemiology; a tutorial review. Technical report No. 243, Department of Statistics, University of Washington.

Thompson, E. A. (1994a). Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface conference* J. Sall and A. Lehman (eds.) Pp. 498-506. Interface Foundation of North America: Fairfax Station, VA.

Thompson, E. A. (1994b). Monte Carlo likelihood in genetic mapping. *Statistical Science* **9:** 355–366.

Thompson, E. A. (1994c). Monte Carlo likelihood in genetic analysis. In *Probability, Statistics, Optimization: a tribute to Peter Whittle.* F.P. Kelly (ed.). Wiley; New York. Chapter 21, Pp. 281-293.

Thompsom, E. A. (1994d). Monte Carlo likelihood in the genetic analysis of complex traits. *Phil. Trans. Roy. Soc. (Lond.) Series B* **344:** 345–351.

Thompson, E. A. (1995) Likelihood and linkage: from Fisher to the future. (1994 JSM Fisher Lecture.) *Ann. Statist.*; in press

Thompson E.A. & Neel J.V. (1995) Private polymorphisms in the Amerindian Population; How many? How old? How useful for genetic taxonomy? Symposium on Genetic Taxonomies; submitted.

Vieland, V. J., & Hodge, S. E. (1995). Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. *American Journal of Human Genetics* **56:** 33–43.

Williamson, J. A., & Amos, C. I. (1990). On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage, when the model is misspecified. *Genetic Epidemiology* **7:** 309–318.

Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist* **56:** 330–338.

Wright, S. & McPhee, H.C. (1925) An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees. *Journal of Agricultural Research* **31:** 377–383.