Chapter 4

Genetic Linkage

4.1 Linkage and recombination: genetic distance

Contrary to Mendel's second law (Mendel, 1866), there is dependence in the inheritance of genes at syntenic loci (that is, loci on the same chromosome pair). Such loci are said to be *linked*. Where the data are affected by the alleles at more than one locus on a chromosome pair, it is no longer sufficient to consider the inheritance of genes at each locus separately.

Recall the meiosis indicators of (equation (1.2)):

$$\begin{aligned} S_{i,j} &= 0 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's maternal gene} \\ &= 1 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's paternal gene.} \end{aligned}$$

Here $i = 1, ..., m$ indexes the meioses of the pedigree, and $j = 1, ..., L$ indexes the genetic loci. The marginal distribution of each $S_{i,j}$ is as before (section 1.2):

$$\Pr(S_{i,j} = 0) \;=\; \Pr(S_{i,j} = 1) \;=\; \frac{1}{2}.$$

For different meioses i, the $S_{i,j}$ are independent.

We say that, in a given meiosis, *recombination* has occurred between two loci j and l, if the genes segregating to the gamete at these two loci are from different parental chromosomes. That is, they derive from different grandparents. For two loci, we do not need a full model for the vector $S_{i,\bullet}$ (equation 1.3). The pairwise distribution of $(S_{i,j}, S_{i,l})$ is determined by the *recombination frequency*, which is a measure of the dependence in inheritance between the two loci. For two given loci (l and j) the recombination frequency ρ between them is

$$\rho \;=\; \Pr(S_{i,l} \neq S_{i,j}) \quad \text{for each } i, \quad 0 \leq \rho \leq \frac{1}{2}. \tag{4.1}$$

For loci that are close together on a chromosome, ρ is close to 0. For independently segregating loci, $\rho = \frac{1}{2}$. Note that, although θ is the notation often used for the

recombination parameter in genetic analysis, we here use ρ and reserve θ for the more general set of all parameters of the genetic model.

A point on a gamete chromosome at which the DNA switches from being a copy of the parent's maternal [paternal] chromosome to being a copy of the parent's paternal [maternal] chromosome is known as a *crossover*. Haldane (1919) defined *genetic map distance* between any two loci as the expected number of crossovers occurring between them on a gamete. The unit of genetic distance is the Morgan, but it is often more convenient to use centiMorgans (cM). Since expectations are additive, regardless of dependence of random variables, genetic map distances are always additive. They also subsume any positional variation in recombination rates such as recombination hot-spots: they say nothing about the relationship between physical and genetic distances. A recombination occurs between two loci, if, in that meiosis, there are an odd number of crossovers between them.

In equation (4.1), we assume that the recombination frequency ρ does not vary with the meiosis i. In practice, recombination frequencies vary among meioses, a major factor in this variation being the sex of the parent. The expected number of crossovers between two locations can be quite different for a gamete from a male than for a gamete from a female. Thus genetic maps are sex-specific, where the sex in question is that of the parent producing the gamete. For ease of presentation, sex-differences in genetic maps will be ignored in this monograph. Computationally, such variation can be easily accommodated.

Haldane's original meiosis model, and other early models, were *two-strand* models. That is, the locations of crossovers between the two parental chromosomes were modeled. This is sufficient to determine the joint probabilities $\Pr(S_{i,1}, \ldots, S_{i,L})$, hence, in principle, probabilities of L-locus gene *ibd* patterns among a set of observed related individuals, and hence probabilities of observed data. In Haldane's model, these crossovers were assumed to occur as a Poisson process, rate 1 (per Morgan). Thus there is no *interference*. The number of crossovers in a given genetic distance has a Poisson distribution, the numbers of crossovers in disjoint intervals are independent, and, conditionally on the number occurring, their locations are uniformly and independently distributed, all measures being, of course, with respect to genetic (not physical) distance. The recombination frequency at genetic distance d Morgans, $\rho(d)$, as a function of d is known as the *map function*. Under the no-interference model, $\rho(d)$ is the probability that a Poisson random variable with mean d is odd:

$$\begin{aligned} \rho(d) &= \sum_{k \text{ odd}} e^{-d}\frac{d^k}{k!} = \frac{1}{2}e^{-d}\sum_{k=0}^{\infty}\left(\frac{d^k}{k!} - \frac{(-d)^k}{k!}\right) \\ &= \frac{1}{2}(1 - \exp(-2d)). \end{aligned} \tag{4.2}$$

Note that, under this model, $\rho(d)$ is an increasing function of d, $\rho(d) \to \frac{1}{2}$ as $d \to \infty$, and $\rho(d) \approx d$ as $d \to 0$. These are basic properties of map functions under most models for meiosis (see Chapter 5).

In modeling crossovers Fisher (1922) went to the other extreme: he assumed complete interference in the region of *Drosophila willistoni* chromosome he

considered. That is, at most one crossover in this chromosome region can occur in any meiosis. In this case, genetic distance and recombination frequency are equivalent. Although this model does not make sense over large chromosomal segments, current mouse data (King et al., 1991) suggest almost complete interference over regions of about 10cM.

4.2 Haplotypes, linkage, and association

The vector of alleles at loci on a chromosome is a *haplotype*, and a *multilocus genotype* is a pair of haplotypes. Note that the set of single-locus genotypes do not determine the multilocus genotype. The multilocus genotype includes a specification of *phase*; that is, which alleles (one at each locus) are on the same chromosome. Some modern literature does refer to the set of single-locus genotypes (without phase) as the multilocus genotype, but this terminology is confusing. For clarity, we refer to the potentially observable set of (single-locus) genotypes at any set of DNA marker loci as *marker phenotypes*, even when these loci do not correspond to functional genes.

For simplicity in this section we restrict attention to two diallelic loci, one with codominant alleles A_1 and A_2, and the other with codominant alleles B_1 and B_2. There are then four haplotypes A_1B_1, A_1B_2, A_2B_1 and A_2B_2. Suppose the haplotype frequencies are q_1, q_2, q_3 and q_4. There are 10 two-locus genotypes, but only 9 phenotypes. Genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1 both have the double-heterozygote phenotype A_1A_2, B_1B_2. The notation A_1B_1/A_2B_2 denotes that alleles A_1 and B_1 are on a single haplotype, and alleles A_2 and B_2 are on the other. Just as for the single-locus *ABO* blood type example (section 2.5), haplotype frequencies can be estimated from phenotype frequencies via the EM algorithm, under the general model of unconstrained patterns of association among the loci. Each phenotypic observation on an individual consists of a set of single-locus genotypes.

For the case of two loci, haplotypes are unobservable only for the double-heterozygote phenotype A_1A_2, B_1B_2. Each individual who is A_1A_2, B_1B_2 is of genotype A_1B_1/A_2B_2 with probability $q_1q_4/(q_1q_4 + q_2q_3)$ and of genotype A_1B_2/A_2B_1 with probability $q_2q_3/(q_1q_4 + q_2q_3)$. Thus, given a set of current haplotype frequency estimates q_i, $i = 1, \ldots, 4$ and the phenotypic counts, the conditional expected genotypic counts are easily obtained. New haplotype estimates then are the expected multinomial proportions of each haplotype.

Clearly, this method can be extended to any number of loci. Thus, for example, population data can be used to estimate haplotype frequencies at a set of tightly linked SNP markers (section 1.1). However, an individual heterozygous at l loci can have any of 2^{l-1} multilocus genotypes (pairs of haplotypes). The observation is partitioned among the 2^{l-1} possible pairs, in accordance with current haplotype frequency estimates. Performance of the EM algorithm can be poor when there are many linked polymorphic marker loci, particularly when many haplotypes may not occur in the sample. Thus, for microsatellite markers with many alleles or for many tightly linked SNP markers (section 1.1), population marker phenotype data alone

may not serve to provide accurate haplotype frequencies. Better performance of the EM algorithm is obtained by constraining some haplotype frequencies to zero, when the estimates of their frequencies appear to be approaching zero.

An individual who is homozygous at both loci can pass on only one haplotype to an offspring; for example an A_1A_1, B_2B_2 individual must pass on an A_1B_2 haplotype. An individual who is homozygous at one locus can pass either of two haplotypes. Each possibility has probability 1/2 regardless of the recombination frequency ρ between the two loci; for example, an A_1A_1, B_1B_2 individual passes on A_1B_1 or A_1B_2 each with probability 1/2. Only the double heterozygote A_1A_2, B_1B_2 provides meioses which are *informative for linkage*. That is, this individual passes each of the four haplotypes A_1B_1, A_1B_2, A_2B_1 and A_2B_2, with probabilities $(1-\rho)/2$, $\rho/2$, $\rho/2$ and $(1-\rho)/2$ if his genotype is A_1B_1/A_2B_2, and with probabilities $\rho/2$, $(1-\rho)/2$, $(1-\rho)/2$, and $\rho/2$ if his genotype is A_1B_2/A_2B_1.

A measure of allelic association between the two loci is

$$\begin{aligned}\Delta &= \Pr(A_1B_1) - \Pr(A_1)\Pr(B_1) \\ &= q_1 - (q_1+q_2)(q_1+q_3) \\ &= (q_1q_4 - q_2q_3)\end{aligned}$$

since $q_1+q_2+q_3+q_4 = 1$. This measure is due to Robbins (1918) and is known as the coefficient of *linkage disequilibrium*. This name is confusing, but the term is too well established to change. In the absence of selection, allelic associations between loci arise from population structure, admixture and history. They are, however, maintained by tight linkage. Suppose the current haplotype frequencies are q_1, q_2, q_3 and q_4, as above. In expectation, in the absence of selection, allele frequencies are unchanged at the next generation. Suppose the haplotype frequencies are q_1^*, q_2^*, q_3^* and q_4^*. Now, for example, an A_1B_1 haplotype in an offspring can arise in three ways. It can be transmitted from a parental A_1B_1 without recombination. It can also be transmitted from a parental A_1B_1 with recombination, if the second parental haplotype is A_1B_1, A_1B_2, or A_2B_1. Finally, with recombination, an A_1B_2/A_2B_1 parent may transmit an A_1B_1 haplotype. Thus

$$\begin{aligned}q_1^* &= (1-\rho)q_1 + \rho q_1(q_1+q_2+q_3) + \rho q_2q_3 \\ &= q_1 - \rho(q_1q_4 - q_2q_3) = q_1 - \rho\Delta.\end{aligned}$$

Analogously, $q_2^* = q_2 + \rho\Delta$, $q_3^* = q_3 + \rho\Delta$ and $q_4^* = q_4 - \rho\Delta$. Thus

$$\begin{aligned}\Delta^* &= q_1^*q_4^* - q_2^*q_3^* \\ &= (q_1-\rho\Delta)(q_4-\rho\Delta) - (q_2+\rho\Delta)(q_3+\rho\Delta) \\ &= \Delta - \rho\Delta(q_1+q_2+q_3+q_4) + \rho^2(\Delta-\Delta) \\ &= (1-\rho)\Delta.\end{aligned}$$

In the absence of any maintaining force, such as selection, or continuing population subdivision and admixture, allelic associations decay in expectation over the generations, by a factor $(1-\rho)$. For unlinked loci ($\rho = \frac{1}{2}$) this decay is rapid, but for tightly linked loci ($\rho \approx 0$) allelic associations may be maintained over

hundreds of generations. Actual population are finite, and mating is non-random; allelic associations are often seen in small natural populations. For a more detailed discussion, see Weir (1996).

4.3 Lod scores for two-locus linkage analysis

In the absence of genetic interference (equation (4.2)), and in fact under most models for meiosis (Chapter 5), the recombination frequency, ρ, is an increasing function of genetic distance. Genetic mapping involves the ordering of loci on a chromosome, the detection of linkage, and the estimation of recombination frequencies. Some loci determine traits: others are DNA markers. Typically, a map constructed of DNA markers is then used to map the loci controlling a trait of interest. For unlinked loci, $\rho = \frac{1}{2}$. For loci that are genetically *linked*, $\rho < \frac{1}{2}$. *Linkage analysis* is concerned with estimating ρ and with testing the null hypothesis $H_0 : \rho = \frac{1}{2}$ against the alternative $H_1 : \rho < \frac{1}{2}$. Estimates and tests are based on likelihoods and likelihood ratios (Chapter 2).

If the genes (one at each of two loci) descending from given parent to a given offspring derive from different parental chromosomes, and hence from different grandparents, the offspring is said to be *recombinant* with respect to these two loci. In the simplest cases, whether an offspring i is a recombinant ($X_i = 1$) or not ($X_i = 0$) is observable. Then $P(X_i = 1) = \rho$ and the number of recombinants T in n independent meioses has the binomial $B(n, \rho)$ distribution.

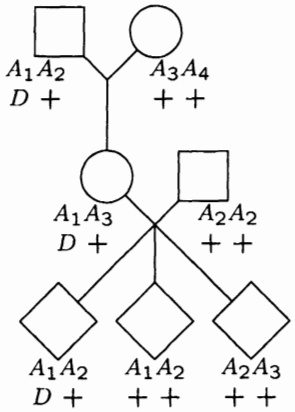

FIGURE 4.1. *Example of recombination in a three-generation family*

For example, at a DNA marker locus, suppose two grandparents have types A_1A_2 and A_3A_4, and their daughter has type A_1A_3. Suppose she marries someone of type A_2A_2 and their three children are of types A_1A_2, A_1A_2 and A_2A_3. Suppose also the grandparent of type A_1A_2, the daughter, and the first of the three children

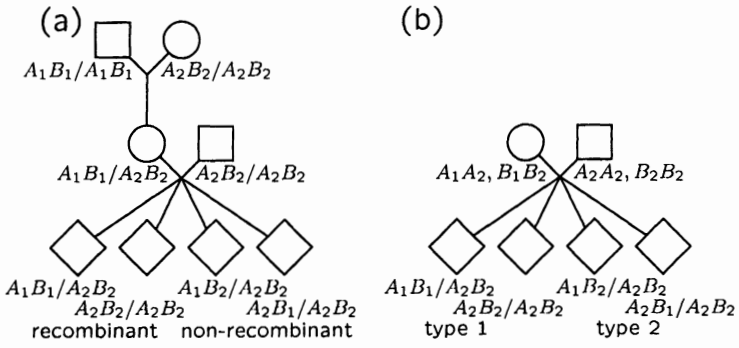

FIGURE 4.2. *Examples of (a) phase-known and (b) phase-unknown backcross linkage designs*

all carry an allele D causing some trait of interest, and the other individuals carry only normal alleles, denoted + (Figure 4.1). Then we know the trait allele D segregates with the A_1 marker allele from the grandparent to his daughter, and that the normal allele + segregates with A_3 from her other parent. To the three children from their mother, we have segregation of A_1 with D, of A_1 with +, and of A_3 with +. Thus children 1 and 3 are non-recombinant ($X_1 = X_3 = 0$) and child 2 is recombinant ($X_2 = 1$). So $n = 3$, the number of recombinants $T \sim B(3, \rho)$, and in this example T takes the value $t = 1$.

In the case where we can classify each offspring as recombinant or non-recombinant, as above, the number of recombinants in n observed offspring is $T \sim B(n, \rho)$. This type of data arises in a *backcross experiment*, where two inbred lines are crossed, and the hybrid is crossed back to either of the two lines. An example of this linkage design is shown in Figure 4.2(a). Suppose one line has only alleles A_1 at one locus and B_1 at the other (genotype A_1B_1/A_1B_1), while the other line has only A_2 and B_2 (genotype A_2B_2/A_2B_2). Then the cross will produce hybrid individuals who have genotype A_1B_1/A_2B_2. If we then cross back to the A_2B_2/A_2B_2 line, all the offspring will get A_2B_2 from that parent, and we can tell which combination A_1B_1, A_2B_2, A_1B_2 or A_2B_1 they get from their hybrid parent, and so whether or not they are recombinant.

Suppose n offspring of such matings are scored, and t are recombinant. To test for linkage, we compare the likelihood to its value in the absence of linkage ($\rho = \frac{1}{2}$). The log-likelihood difference is

$$\text{(4.3)} \quad \text{lod}(\rho) \;=\; \ell(\rho) - \ell(\frac{1}{2}) \;=\; t\log(\rho) \;+\; (n-t)\log(1-\rho) \;+\; n\log(2).$$

In linkage analysis it is traditional to use logs to base 10, and to refer to (4.3) as the *lod score* (Morton, 1955). In our numerical examples we shall use natural logarithms except where specified, for easier comparison with standard statistical results.

The maximum likelihood estimate of ρ is $\widehat{\rho} = t/n$, provided $2t \leq n$: note only values of $\rho \leq \frac{1}{2}$ have meaning under the model (4.2). Then to test $\rho = \frac{1}{2}$ against $\rho < \frac{1}{2}$, we may consider the maximized value of the lod score:

$$\text{lod}(\widehat{\rho}) \quad = \quad t \log t \; + \; (n-t)\log(n-t) - n\log(n/2) \tag{4.4}$$

provided $2t \leq n$, and 0 otherwise. This maximized lod score is a decreasing function of t, and we reject the null hypothesis $\rho = \frac{1}{2}$ if $t < t_0$. The critical value t_0 may be chosen to give a specified size of the test (type I error).

In many linkage experiments, however, or in human genetics where we do not have designed crosses, we often cannot classify all individuals as recombinant and non-recombinant. There are many possibilities, but a typical one is the *phase-unknown backcross*. This arises if one parent is A_1A_2, B_1B_2 and the other is A_2A_2, B_2B_2 as above, but now we do not know whether the first parent received A_1B_1 and A_2B_2 (type 1 combinations) from her parents, or A_1B_2 and A_2B_1 (type 2 combinations). This design is shown in Figure 4.2(b). Suppose we have families of this kind, and in each family we type just two offspring. Since each offspring gets A_2B_2 from the father, we can, as before, determine what each got from the mother. Either both offspring get the same "type" of combination (type 1 or type 2), or there is one of each. If there is one of each, then one offspring must be a recombinant and the other not; so this event has probability $\rho^* = 2\rho(1-\rho)$. If they get the same "type" of combination, then either both are recombinant, or neither is, so this event has probability $1 - \rho^* = \rho^2 + (1-\rho)^2$. So instead of a $T \sim B(n, \rho)$ count of recombinants, we have a $W \sim B(n, \rho^*)$ count of families.

Note however, that for $0 \leq \rho \leq \frac{1}{2}$, ρ^* is a 1-1 monotone increasing function of ρ, and when $\rho = \frac{1}{2}$ $\rho^* = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$. So testing $H_0 : \rho = \frac{1}{2}$ against the one-sided alternative $H_1 : \rho < \frac{1}{2}$, is exactly equivalent to testing $H_0^* : \rho^* = \frac{1}{2}$ against the one-sided alternative $H_1^* : \rho^* < \frac{1}{2}$. Thus the test follows exactly as before; we reject $\rho^* = \frac{1}{2}$ and conclude there is linkage if $W < w_0$, where again the critical value w_0 is determined by the desired size of the test.

4.4 Power, information and *Elod*s

For simplicity, consider the case of the phase-known backcross, where $T \sim B(n, \rho)$. Now when n is large, T is approximately $N(n\rho, n\rho(1-\rho))$, and under $H_0 : \rho = \frac{1}{2}$ it is a *good* approximation to take $T \sim N(\frac{n}{2}, \frac{n}{4})$. So $\frac{2}{\sqrt{n}}(T - \frac{n}{2}) \sim N(0,1)$ and for a test size α we reject H_0 in favor of $H_1 : \rho < \frac{1}{2}$ if $\frac{2}{\sqrt{n}}(T - \frac{n}{2}) \leq \Phi^{-1}(\alpha)$ where Φ is the standard Normal cumulative distribution function. For example, for $\alpha = 0.025$, $\Phi^{-1}(\alpha) = -1.96 \approx -2$, so H_0 is rejected if $T \leq \frac{n}{2} - \sqrt{n} = k^*$ (Table 4.1).

Using equation (4.4), we find that the (base 10) lod score is around 1 for a number of recombinants at the critical value for a test of size $\alpha = 0.025$ of $H_0 : \rho = \frac{1}{2}$ (Table 4.1). Traditionally, a base-10 lod score of 3 is required to infer linkage (Morton, 1955). This is a more stringent test, the idea being that if two arbitrary locations in the genome are chosen the prior probability of linkage is small. Also given in the table is the upper bound on the number of recombinants that will provide a lod score of 3.

offspring sampled n	critical value k^*	recombinant proportion k^*/n	lod score $\text{lod}_{10}(k^*/n)$	recombinants for lod score 3
25	≈ 7	≈ 0.3	1.088	≤ 3
100	≈ 40	≈ 0.4	0.874	≤ 31
625	≈ 287	≈ 0.46	0.905	≤ 267
1024	≈ 480	≈ 0.48	0.869	≤ 452

TABLE 4.1. *Critical values for a test size $\alpha = 0.025$ and base-10 lod scores for binomial samples*

Type	genotypes	number	each prob
I	$A_1A_1, B_2B_2,\ A_2A_2, B_1B_1$	2	$\rho^2/4$
II	A_1A_1, B_1B_2	1	$\frac{1}{2}(\rho^2 + (1-\rho)^2)$
III	A_1A_1, B_1B_2 etc.	4	$\frac{1}{2}\rho(1-\rho)$
IV	$A_1A_1, B_1B_1,\ A_2A_2, B_2B_2$	2	$(1-\rho)^2/4$

TABLE 4.2. *The groups of offspring genotypes in an intercross design. Note the A_1A_1, B_1B_2 type includes both double-heterozygote two-locus genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1. The third group includes the four types heterozygous at one of the two loci: A_1A_1, B_1B_2, A_1A_2, B_1B_1, A_2A_2, B_1B_2 and A_1A_2, B_2B_2*

Now if ρ is the true value, the probability H_0 is rejected is

$$\begin{aligned} \Pr(T < k^*; \rho) &= \Pr\left(\frac{T - n\rho}{\sqrt{n\rho(1-\rho)}} < \frac{k^* - n\rho}{\sqrt{n\rho(1-\rho)}}\right) \\ &\approx \Phi\left(\frac{k^* - n\rho}{\sqrt{n\rho(1-\rho)}}\right) = \Phi\left(\frac{\Phi^{-1}(\alpha) + \sqrt{n}(1-2\rho)}{2\sqrt{\rho(1-\rho)}}\right) \end{aligned} \tag{4.5}$$

again using the Normal approximation to the Binomial distribution. This is the power function of the test, and decreases over $0 \leq \rho \leq \frac{1}{2}$. Clearly, for a given sample size, linkage is more easily detected when ρ is small. Conversely, for given ρ, one may use (4.5) to determine the sample size n required for given power. The case of the phase-unknown backcross is analogous, with ρ being replaced by ρ^*, and n now denoting the number of two-child families.

In order to get more information, an *intercross* experiment may be performed, instead of a *backcross*. In this case two phase-known hybrid parents, each of type A_1B_1/A_2B_2 are mated. There are nine types of offspring, but these fall into four groups, shown in Table 4.2. Each type within a group has the same probability, as a function of ρ, and hence the total count of offspring in each group contains all the available information for linkage. (These total counts are the sufficient statistics for ρ.)

Consider a sample of size n, with n_j in class j, $j = 1, 2, 3, 4$. As in equation (2.5),

Types	H_2 : general	H_1:total prob	H_0 : $\rho = \frac{1}{2}$
I	q_1	$\frac{1}{2}\rho^2$	0.125
II	q_2	$\frac{1}{2}(\rho^2 + (1-\rho)^2)$	0.25
III	q_3	$2\rho(1-\rho)$	0.5
IV	q_4	$\frac{1}{2}(1-\rho)^2$	0.125

TABLE 4 3. *Probabilities of data observations in an intercross design. Given are the total probabilities of each group of types shown in Table 4.2, under the three alternative hypotheses*

the log-likelihood for these multinomial data is, up to an additive constant,

$$\ell_n(\mathbf{q}) = \sum_{j=1}^{4} n_j \log_e q_j(\rho).$$

The probabilities of each phenotype group are shown in Table 4.3, under the general multinomial model H_2, the general linkage model H_1, and in the absence of linkage H_0.

For example, suppose $\mathbf{n} = (1, 72, 42, 85)$.

Under H_2 : *general* q_j, $\sum_{j=1}^{4} q_j = 1$, $\hat{q}_j = n_j/n$,

or $\hat{\mathbf{q}} = (0.005, 0.36, 0.21, 0.425)$. The dimension of H_2 is 3.

Under H_1 : *general* ρ, for these data we find, by evaluating the log-likelihood, that $\hat{\rho} = 0.12$ giving $\mathbf{q}(\hat{\rho}) = (0.007, 0.394, 0.211, 0.387)$. The dimension of H_1 is 1.

The null hypothesis is of no linkage; H_0 : $\rho = \frac{1}{2}$. This has dimension 0, and the fixed probabilities $\mathbf{q}(\frac{1}{2}) = (0.125, 0.25, 0.5, 0.125)$ of the four classes of types.

We see that the estimated cell probabilities under H_1 and H_2 are in good agreement, but quite different from those under H_0. Computing the maximized log-likelihoods for H_i, $i = 0, 1, 2$, we find that they are -307.76, -217.87, and -217.14 respectively. For testing null H_0 against alternative H_1, the (base e) lod score is 89.9. Twice this value (179.8) has approximately a χ^2_1 if H_0 is true. So clearly H_0 is rejected.

For testing null H_1 against alternative H_2, the lod score is 0.73, and twice this value (1.46) is χ^2_2 if H_1 is true. So H_1 is not rejected.

For multinomial data in general, we can find the form of the Kullback-Leibler information (section 2.2). Suppose $\mathbf{q}$ is the true value of $\mathbf{q}$, and $\mathbf{q_0}$ is some hypothesized value.

$$\ell_n(\mathbf{q}) = \sum_{j=1}^{4} n_j \log_e q_j.$$

So for a sample size n

$$\begin{aligned} K_n(\mathbf{q_0}; \mathbf{q}) &= \mathrm{E}_{\mathbf{q}}(\ell_n(\mathbf{q}) - \ell_n(\mathbf{q_0})) \\ &= n \sum_{j=1}^{4} q_j \log_e q_j - n \sum_{j=1}^{4} q_j \log_e q_{0j} \end{aligned}$$

True ρ	0.0	0.1	0.2	0.3	0.4	0.5
Intercross data	1.04	0.479	0.226	0.089	0.021	0.0
Backcross (phase known)	0.69	0.368	0.193	0.082	0.021	0.0
Backcross (phase unknown)	0.35	0.111	0.033	0.006	0.0004	0.0

TABLE 4 4. *Comparison of the information in linkage designs per offspring individual sampled: Kullback Leibler information for testing $\rho = 1/2$ as a function of the true value of ρ*

or, for a single observation,

$$K_1(\mathbf{q_0}; \mathbf{q}) = \sum_{j=1}^{4} q_j \log_e \left(\frac{q_j}{q_{0j}} \right).$$

(Note the notation is reversed from section 2.2. Here $\mathbf{q}$ is the true parameter value, and $\mathbf{q_0}$ is the hypothesized value.) In the case of linkage analysis data, $q_j = q_j(\rho)$ and the null hypothesis is $H_0 : \rho = \frac{1}{2} : \ q_{0j} = q_j(\frac{1}{2})$. Evaluating K_1 for the above *phase-known intercross* experiment, and for the previous binomial *phase-known* and *phase unknown* backcross experiments, we obtain the measures of information per offspring individual shown in Table 4.4.

This is a measure of information, per offspring sampled, for detecting linkage when ρ is the true value. We see that the more ρ differs from $\frac{1}{2}$ the more information there is, as expected. Also each phase-known offspring contributes at least twice as much as each of the two offspring in the phase-unknown case. Particularly when ρ is close to $1/2$, the phase-unknown two-offspring design has low power. We see that each *intercross* offspring contains more information than a *backcross* offspring, also as expected. However, note that there is *not* twice as much information in the intercross offspring, as there would be if we could tell the difference between the A_1B_1/A_2B_2 and A_1B_2/A_2B_1 offspring (see Table 4.3). As ρ gets closer to $\frac{1}{2}$ there is almost no additional information in doing an intercross design rather than a backcross.

Note that for $\rho = \frac{1}{2}$, the Kulback-Leibler information is the expected base-e lod score at the true value ρ_T of the recombination frequency. This measure of information is very widely used in linkage analysis, and is known as the *Elod* (Thompson et al., 1978). Note that we expect the base-e lod score to be approximately nK_1 when n is large. For our previous data with $n = 200$, we had $\hat{\rho} = 0.12$; in fact, the data were simulated at $\rho = 0.1$. Then 200×0.479 is about 95, in good agreement with the lod score value of 90 which we obtained. This also tells us that if we had realized that ρ might be around 0.1, it was very wasteful to breed 200 mice. When $\rho = 0.1$, about 20 mice are expected to give a lod score (base e) of more than 9; this is plenty to detect that $\rho \neq \frac{1}{2}$. (Note again that we have used natural logarithms in these examples, contrary to standard practice in genetics.)

The material of sections 4.3 and 4.4 extends readily to the estimation and testing of two recombination frequencies ρ_m in males, and ρ_f in females. Similar likelihood ratio tests may be used to test equality of male and female recombination frequencies. For a much more detailed account of classical linkage analysis and more

modern developments, the reader may consult the excellent text of Ott (1999).

4.5 Two-locus kinship and gene identity

The recursive equations for multiple kinship coefficients of section 3.4 (equations (3.6) and (3.7)) extend to multiple loci, conditioning on the meiosis indicators in a given meiosis, over the loci in question. Consider, for example, the case of $\psi_2(L_1(B^{(1)}, C), L_2(B^{(1)}, E))$. This expression denotes the two locus kinship probability, that, in a single gamete segregating from B, the gene at locus L_1 is *ibd* to that on a gamete segregating from individual C, while the gene at locus L_2 is *ibd* to that on a gamete segregating from individual E. The identical superscript "(1)" on the individual B indicates that we are considering here a single meiosis i from B, rather than two separate meioses to different offspring. Now if B is not an ancestor of C or E, we may condition on the four events $(S_{i,1},\ S_{i,2}) = (0,0),\ (0,1),\ (1,1),\ (1,0)$ with probabilities $\frac{1}{2}(1-\rho),\ \frac{1}{2}\rho,\ \frac{1}{2}(1-\rho),\ \frac{1}{2}\rho$ respectively, where ρ is the recombination frequency between locus 1 and locus 2. Thus we obtain

$$\begin{aligned}\psi_2(L_1(B^{(1)},\ C), L_2(B^{(1)},\ E)) \ &= \ \frac{1}{2}(1-\rho)\psi_2(L_1(M_B^{(B)}, C), L_2(M_B^{(B)}, E)) \ + \\ &\frac{1}{2}\rho\psi_2(L_1(M_B, C), L_2(F_B, E)) + \frac{1}{2}(1-\rho)\psi_2(L_1(F_B^{(B)}, C), L_2(F_B^{(B)}, E)) \\ &+ \ \frac{1}{2}\rho\psi_2(L_1(F_B, C), L_2(M_B, E)).\end{aligned}$$

Again, the superscript specifies which meiosis from an individual is considered—here the ones from M_B and F_B to B. In the case of two loci it is necessary to distinguish the meioses from a given parent. The full set of equations for determining two-locus gene identity probabilities between genes segregating from up to four individuals are given by (Thompson, 1988). These equations can be used to determine two-locus *ibd* state probabilities, even on a large and complex pedigree.

At two linked loci, there are also many more possible gene identity patterns (Denniston, 1975). Some relationships which have identical gene *ibd* probabilities at a single locus can, in principle, be distinguished by data at linked loci. The simplest example is for the three unilineal ($\kappa_2 = 0$) pairwise relationships of grandmother-granddaughter (G), half-sisters (H), and aunt-niece (N). Each of these relationships has $\boldsymbol{\kappa} = (\frac{1}{2}, \frac{1}{2}, 0)$, and hence they are indistinguishable on the basis of data at independently segregating loci. For such relationships, gene identity at two linked loci is summarized by

$$\kappa_{1,1}(\rho) \ = \ P(\textit{share 1 gene ibd at each of 2 loci at recombination } \rho).$$

For the three relationships above, we have

$$G \ : \ \kappa_{1,1}(\rho) \ = \ \frac{1}{2}(1-\rho)$$

$$H \ : \ \kappa_{1,1}(\rho) = \frac{1}{2}(\rho^2 + (1-\rho)^2) = \frac{1}{2}R \text{ say}$$
$$N \ : \ \kappa_{1,1}(\rho) = \frac{1}{2}((1-\rho)R + \rho/2).$$

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

Note that, although $\kappa_{1,1}(\rho)$ is sufficient to specify pairwise genotype and phenotype distributions, it does not determine the two-locus kinship of the individuals, unlike at a single locus where $\psi = (\kappa_1 + 2\kappa_2)/4$. The shared genes at the two loci may be on the same haplotype in the individual, or on different ones. In fact, in H they are necessarily on the same (maternal) haplotype in the two half-sibs, while in G they may be on either haplotype of the grandmother. For N, for the first term they are on the same haplotype in the aunt, while the last term corresponds to the case where the genes at the two loci are on two different haplotypes in the aunt. In fact, G and H have the same two-locus kinship, $(1/8)(1-\rho)^2R$.

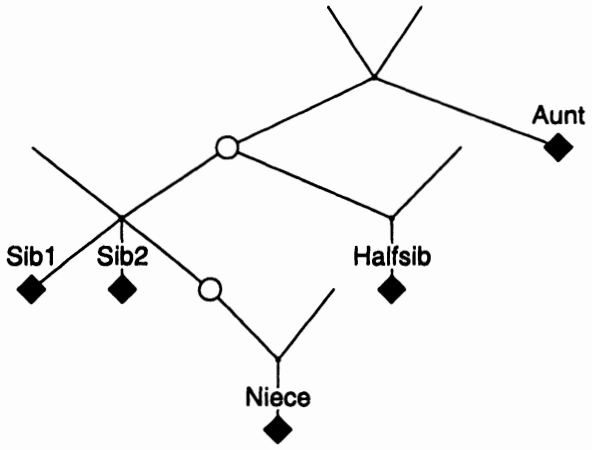

FIGURE 4.3. *Multi-locus genetic marker data are available on a pair of sibs, and on a third related individual, who may be an aunt, niece, or half-sister of the pair*

Returning again to the example of section 3.6, consider three individuals consisting of a pair of individuals who are putative full sibs, and a third who may be the aunt, niece, or half-sib of the sib pair (Figure 4.3). This example arose in a real example of inference of relationships considered by Browning (1999). Only with joint analysis of the data at linked loci on all three individuals are the alternative three relationships identifiable (Table 4.5). In the real-data example, the most likely

	Individuals	
	Pairwise	Joint
Loci unlinked	$H \equiv N \equiv A$	$H \equiv N$
Loci linked	$N \equiv A$	H, N, A identifiable

TABLE 4.5. *Distinguishing relationships among three individuals who are putatively a pair of sisters with an aunt, niece, or half-sib*

relationship is that the third individual is a niece of the sib pair (Browning and Thompson, 1999). Due to one member of the sib pair having data at relatively few markers, the inference is not conclusive. However, with data on a 10cM genome scan, for example, there would be no difficulty in distinguishing the relationships provided analysis is performed jointly both over individuals and over loci.

Some pairwise relationships which provide identical two-locus kinship coefficients have different three-locus kinship coefficients (Thompson, 1988). Thus, there are relationships that are non-identifiable on the basis of gametes observed at pairs of loci (whatever the values of the recombination frequencies between them), but that are identifiable on the basis of gametes observed at trios of loci. One may conjecture that there are relationships non-identifiable on the basis of L-locus kinship, but identifiable on the basis of $L+1$-locus kinship.

4.6 Homozygosity mapping with a single marker

We introduce the ideas both of linkage analysis for linkage detection and of association analysis for the fine-scale localization of trait genes via *homozygosity mapping* using the ideas of two-locus gene *ibd* already encountered. Homozygosity mapping was developed by Lander and Botstein (1987) to detect linkage for the loci determining rare recessive disease traits, but as noted by Smith (1953) the principle is the same as in any linkage analysis: a likelihood for the recombination frequency ρ, or more generally for the trait locus location, is computed. With a single marker locus, the maximized likelihood under the hypothesis of linkage $\rho < \frac{1}{2}$ is compared with the likelihood under the hypothesis that the trait locus is not linked to the marker locus or loci $\rho = \frac{1}{2}$. In the case of homozygosity mapping, the linkage inference is based on data on unrelated affected inbred individuals. It relies on the fact that an inbred affected individual has high probability of carrying two *ibd* genes at the the trait (disease) locus (section 3.3), and hence also at any closely linked marker locus. Since *ibd* genes are necessarily of the same allelic type, such individuals will show a patch of homozygosity in the neighborhood of the trait locus; where the same markers are homozygous across multiple inbred affected individuals, the evidence for linkage accumulates.

Suppose the frequency of the recessive disease allele is q, and at the marker locus alleles A_i have frequencies p_i. Suppose that the affected individual has inbreeding coefficient f, and probability $f_2(\rho)$ of carrying genes *ibd* at both of two loci between which the recombination frequency is ρ. Then the probability the individual is

autozygous at a specific one of the two loci but not the other is $f - f_2(\rho)$, and the probability he is autozygous at neither is $(1 - 2f + f_2(\rho))$. If the individual has marker phenotype A_jA_l, he cannot be autozygous at the marker locus, and we have likelihood ratio

$$\begin{aligned}\frac{L(\rho)}{L(\rho=\frac{1}{2})} &= \frac{\Pr(\text{data};\ \rho)}{\Pr(\text{data};\ \rho=\frac{1}{2})} \\ &= \frac{2p_jp_l(q(f-f_2(\rho))+q^2(1-2f+f_2(\rho)))}{2p_jp_l(q(f-f^2)+q^2(1-f)^2)} \\ &= \frac{(f-f_2(\rho))+q(1-2f+f_2(\rho)))}{(1-f)(f+q(1-f))}. \qquad (4.6)\end{aligned}$$

Since sampling is through an affected individual, the data probabilities required here are those of the marker phenotypes, conditional on the affected trait phenotype. However, since the marginal probability of an affected individual, $qf + q^2(1-f)$, does not depend on ρ, the likelihood ratio is also the ratio of the joint probabilities of marker and trait phenotypes. The joint probabilities are slightly more easily considered.

Since $f_2(\rho)$ is a decreasing function of ρ, with value f^2 at $\rho = \frac{1}{2}$, the likelihood ratio (4.6) is always less than one. A heterozygous marker phenotype provides evidence against linkage. However, even at $\rho = 0$, where the value is $q(1-f)/(f + q(1-f))$ the evidence against linkage is not strong unless q is very small. Affected individuals may not carry *ibd* genes at the trait locus.

If the individual has homozygous marker phenotype A_jA_j the likelihood ratio is

$$\begin{aligned}\frac{L(\rho)}{L(\rho=\frac{1}{2})} &= \frac{\Pr(\text{data};\ \rho)}{\Pr(\text{data};\ \rho=\frac{1}{2})} \\ &= \frac{qp_jf_2(\rho) + q^2p_j(f-f_2(\rho)) + qp_j^2(f-f_2(\rho)) + q^2p_j^2(1-2f+f_2(\rho))}{qp_jf^2 + q^2p_jf(1-f) + qp_j^2f(1-f) + q^2p_j^2(1-f)^2} \\ &= \frac{f_2(\rho) + q(f-f_2(\rho)) + p_j(f-f_2(\rho)) + qp_j(1-2f+f_2(\rho))}{f^2 + qf(1-f) + p_jf(1-f) + qp_j(1-f)^2}. \qquad (4.7)\end{aligned}$$

The coefficient of the decreasing function of ρ, $f_2(\rho)$, is $(1-q)(1-p_j)$, and thus this likelihood ratio is maximized at $\rho = 0$. At this value, $f_2(\rho) = f$, so the likelihood ratio is

$$\frac{f + (1-f)qp_j}{(f+(1-f)q)(f+(1-f)p_j)}.$$

This is always greater than one, and is larger if q or p_j is small.

Likelihood ratios are multiplicative over unrelated pedigrees i, or log-likelihoods are additive. The base-10 log-likelihood ratio, or lod score is

$$\text{lod}(\rho) = \sum_i \log_{10}\left(\frac{L_i(\rho)}{L_i(\rho=\frac{1}{2})}\right)$$

where $L_i(\cdot)$ is the likelihood contributed by pedigree i. The maximized lod score is

$$\max_{0\le\rho\le\frac{1}{2}} (\mathrm{lod}(\rho)).$$

Of course, in combining over pedigrees, the maximizing ρ may be neither 0 nor $\frac{1}{2}$. In this case, the form of $f_2(\rho)$ is also relevant, not merely the value of f. Again, a useful measure of information for linkage analysis is the expected lod score or *Elod* (section 4.4):

$$\textit{Elod}(\rho) \quad = \quad \mathrm{E}_\rho(\mathrm{lod}(\rho)). \tag{4.8}$$

The *Elod* is additive over independent pedigrees. Each affected individual with inbreeding coefficient f has probability $f/(f+(1-f)q)$ of having two *ibd* genes at the disease locus. Hence, at $\rho = 0$, the contribution of each such affected individual to the *Elod* is

$$\frac{f}{f+(1-f)q}\sum_j p_j \log\left(\frac{f+(1-f)qp_j}{(f+(1-f)q)(f+(1-f)p_j)}\right) + \frac{(1-f)q}{f+(1-f)q}\log\left(\frac{(1-f)q}{f+(1-f)q}\right).$$

As $q \to 0$, this has limiting value

$$-\sum_j p_j \log(f+(1-f)p_j).$$

For example, for the affected offspring of first-cousin marriages ($f = 1/16$), and a polymorphic marker locus (for example, $p_j = 0.1$ for each of 10 alleles) the value is $\log(6.4)$. A small number of unrelated affected individuals all homozygous at the same polymorphic marker locus provides strong evidence for linkage.

Homozygosity mapping, and linkage analysis in general, can provide good evidence for linkage. With sufficient data, the loci determining simple Mendelian traits can be localized down to 1 cM ($\rho = 0.01$) (Boehnke, 1994). However, even with data at multiple linked loci, finer localization is normally impossible; there are insufficient informative meioses in the set of pedigrees to resolve loci that are too tightly linked. The above development of homozygosity mapping assumed, as do most linkage analyses, absence of allelic association between the trait and marker loci. However, most current copies of a rare recessive disease allele may trace to a single mutation, say on a haplotype carrying marker allele A_j. Then, as seen in section 4.2, at tight linkage, the allelic association between the loci decays slowly. In this case, not only will the majority of affected inbred individuals be homozygous at the marker locus, but most "unrelated" affected inbred individuals will be homozygous A_jA_j, due to remote coancestry of the disease alleles, not modeled by the analysis of the separate pedigrees. In effect, the analysis makes use of the absence of recombination at a large number of ancestral meioses from the original disease mutation to the current affected individuals. Such allelic associations have been used to assist in the fine-scale mapping of many rare recessive diseases including cystic fibrosis (Cox et al., 1989) and Werner's syndrome (Goddard et al., 1996).

4.7 Meiosis at multiple linked loci

We now introduce notation for a chromosome with L ordered loci, $1, \ldots, L$. For ease of notation, we assume that the loci are ordered $1, \ldots, L$ along the chromosome. We consider again the meiosis indicators of equation (1.2), and the vector notation of equation (1.3). Different meioses are independent, but the components of the meiosis indicator vector for meiosis i, $S_{i,\bullet} = (S_{i,1}, \ldots, S_{i,L})$, are dependent. Recall also the notation $S_{\bullet,j} = (S_{1,j}, \ldots, S_{M,j})$ for the set of all meiosis indicators on the pedigree, at locus j (equation (1.3)). Let the intervals between successive loci be $I_1, \ldots, I_{L-1}$. Let $R_j = 1$ if a gamete is recombinant on interval I_j, and $R_j = 0$ otherwise ($j = 1, \ldots, L-1$). Then, in a given meiosis i,

$$\begin{aligned} R_j &= 1 \text{ if } S_{i,j} \neq S_{i,j+1}, \text{ and} \\ R_j &= 0 \text{ if } S_{i,j} = S_{i,j+1}, \quad j = 1, \ldots, L-1. \end{aligned} \tag{4.9}$$

Each vector $(R_1, \ldots, R_{L-1})$ determines two equiprobable vectors $S_{i,\bullet} = (S_{i,1}, \ldots, S_{i,L})$. A model for $S_{i,\bullet}$ is equivalent to a model for $(R_1, \ldots, R_{L-1})$. One simple model for the distributions of $S_{i,\bullet}$ over more than two loci is considered in this section. More general models for $(R_1, \ldots, R_{L-1})$ will be considered in Chapter 5.

In considering the probability of data on related individuals in a pedigree (equation (3.9)):

$$L = \Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}). \tag{4.10}$$

Often (although not always), data observations will be specific to a given locus. For example, for DNA marker loci we observe phenotypes of individuals at given loci. Let $Y_{\bullet,j}$ denote the all data pertaining to locus j, so the full data pertaining to this chromosomal region is $\mathbf{Y} = (Y_{\bullet,1}, \ldots, Y_{\bullet,L})$, and

$$\Pr(\mathbf{Y} \mid \mathbf{S}) = \prod_j \Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j}))$$

where $\mathbf{J}(S_{\bullet,j})$ is the pattern of gene identity by descent among observed individuals, at locus j, which is determined by $S_{\bullet,j}$. Since meioses i are independent, equation (4.10) becomes

$$L = \Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \left(\prod_j \Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j})) \right) \left(\prod_i \Pr(S_{i,\bullet}) \right). \tag{4.11}$$

To proceed further, we need a model for the vector $S_{i,\bullet}$. Such models may derive from our model for the process of meiosis (Chapter 5) or may be based on computationally convenient assumptions. In either case, it is the binary meiosis indicators (1.2) which provide a means to trace the descent and ancestry of genes, at multiple linked loci. Just as for a single locus (section 3.6), they determine patterns of gene-identity-by-descent (gene *ibd*), which in turn determine patterns of phenotypic similarity among relatives.

The simplest models for meiosis assume *no interference*: this implies that the R_j are independent. Under this model, the dependence structure of the $S_{i,j}$ takes a simple form, with a first-order Markov property over loci j, and with meioses i being independent. The probability of any given indicator $S_{i,j}$ conditional on all the others, $\mathbf{S}_{-(i,j)} = \{S_{k,l}; (k,l) \neq (i,j)\}$, depends only on the indicators for the same meiosis and the two neighboring loci:

$$\begin{aligned} \Pr(S_{i,j} = s \mid \mathbf{S}_{-(i,j)}) &= \Pr(S_{i,j} = s \mid S_{i,j+1}, S_{i,j-1}) \\ &= \rho_{j-1}^{|s-S_{i,j-1}|}(1-\rho_{j-1})^{1-|s-S_{i,j-1}|} \\ &\qquad \rho_j^{|s-S_{i,j+1}|}(1-\rho_j)^{1-|s-S_{i,j+1}|} \end{aligned} \tag{4.12}$$

for $s = 0, 1$, where $\rho_j = \Pr(R_j = 1) = \Pr(S_{i,j} \neq S_{i,j+1})$ is the recombination frequency between locus j and locus $j+1$. Note that equation (4.12), is just counting the recombination/non-recombination events in intervals I_{j-1} and I_j, implied by the three indicators $(S_{i,j-1}, S_{i,j} = s, S_{i,j+1})$.

4.8 Multi-locus kinship and gene identity

Under the assumptions of conditional independence or absence of genetic interference, computation and Monte Carlo are, in principle, straightforward. The meiosis indicators, $\mathbf{S} = \{S_{i,j}\}$, are independent over meioses i, and are Markov over a sequence of loci j along a chromosome. The recursive equations for two-locus kinship generalize to the multilocus case, although becoming progressively more complicated. The probability of a *recombination pattern* in the intervals between marker loci is straightforward, being the product of the probabilities of recombination or non-recombination in successive intervals (equation (4.12)).

However, it is the resulting patterns of gene identity by descent among observed individuals that determine probabilities of observed data (equation (4.11)). Although the component $S_{i,j}$ are Markov over loci j, this is not usually so for the resulting patterns of gene *ibd*, $\mathbf{J}(S_{\bullet,j})$, among observed individuals. Different values of $S_{\bullet,j}$ may give rise to the same *ibd* pattern. Along the chromosome, the *ibd* process is an agglomeration of the $S_{\bullet,j}$ process. Grouping the states of a Markov chain does not, in general, produce a Markov chain.

As a specific example, consider again the pedigree of Figure 3.1, and suppose we are interested only in autozygosity of the final individual. Marginally at each locus the autozygosity probability is 7/64 or 0.1094 (section 3.2). Consider three loci, separated by a recombination frequencies of $\rho_1 = \rho_2 = 0.1$. The two-locus inbreeding coefficient of the final individual at recombination frequency 0.1 is 0.0566. This may be computed exactly by the recursive method outlined in section 4.5. Between the outer loci, in the absence of interference, the recombination frequency is

$$\begin{aligned} \rho &= \rho_1(1-\rho_2) + \rho_2(1-\rho_1) \\ &= 0.1 \times 0.9 + 0.1 \times 0.9 = 0.18. \end{aligned}$$

ibd state			Exact	True	Markov
N	N	N	$0.8915 - \delta$	0.7901	0.7881
N	N	I	$0.0183 + \delta$	0.0478	0.0497
N	I	N	$\delta - 0.0038$	0.0257	0.0255
N	I	I	$0.0566 - \delta$	0.0271	0.0273
I	N	N	$0.0183 + \delta$	0.0478	0.0497
I	N	I	$0.0345 - \delta$	0.0050	0.0031
I	I	N	$0.0566 - \delta$	0.0271	0.0273
I	I	I	δ	0.0295	0.0293

TABLE 4.6. *Prior autozygosity probabilities over three linked loci for the final individual of the pedigree of Figure 3.1*

At recombination frequency 0.18, the two-locus inbreeding coefficient of the final individual is 0.0345. These one- and two-locus values determine the three-locus probabilities up to one degree of freedom. We have

$$\begin{aligned}
\Pr(I\ N\ N) + \Pr(I\ N\ I) + \Pr(I\ I\ N) + \Pr(I\ I\ I) &= \Pr(I) = 0.1094 \\
\Pr(N\ I\ N) + \Pr(N\ I\ I) + \Pr(I\ I\ N) + \Pr(I\ I\ I) &= \Pr(I) = 0.1094 \\
\Pr(N\ N\ I) + \Pr(I\ N\ I) + \Pr(N\ I\ I) + \Pr(I\ I\ I) &= \Pr(I) = 0.1094.
\end{aligned}$$

Also, by symmetry, since $\rho_1 = \rho_2$,

$$\Pr(I\ I\ N) = \Pr(N\ I\ I) \text{ and } \Pr(N\ N\ I) = \Pr(I\ I\ N).$$

Then also

$$\begin{aligned}
\Pr(I\ I\ N) + \Pr(I\ I\ I) &= \Pr(I\ I\ N) + \Pr(I\ I\ I) \\
&= \Pr(I\ I; \rho = 0.1) = 0.0566 \\
\Pr(I\ N\ I) + \Pr(I\ I\ I) &= \Pr(I\ I; \rho = 0.18) = 0.345.
\end{aligned}$$

Fixing $\Pr(I\ I\ I) = \delta$, these equations determine all the probabilities, as given in the first column of Table 4.6, under the heading "exact". The values in the column labeled "true" are in fact obtained by Monte Carlo (section 3.7), using 10^8 independent realizations of genes on the pedigree, and are accurate to 10^{-4}. They are fully consistent with the exact probabilities. These probabilities may also be estimated using Markov chain Monte Carlo (Chapter 8). A comparison of the alternative Monte Carlo procedures in this example is given by Thompson (1994*a*).

The final column of Table 4.6 shows the probabilities that would be obtained, using the two-locus transition probabilities, and assuming the process to be first-order Markov. For this (assumed) Markov process of identity (I) and non-identity (N) the transition probabilities, and hence the three-locus probabilities, are determined as follows:

$$\Pr(I \to I) = 0.0566/0.1094 = 0.5174,$$

$$\begin{aligned} \Pr(I \to N) &= 1 - 0.5174 = 0.4826, \\ \Pr(N \to I) &= (0.1094 - 0.0566)/(1.0 - 0.1094) = 0.0593, \\ \text{and} \quad \Pr(N \to N) &= 1 - 0.0593 = 0.9407. \end{aligned}$$

The resulting probabilities patterns of I and N over the three loci are shown in the final column of Table 4.6, labeled "Markov". None of the probabilities computed using the Markov assumption is completely accurate, but those having I at the second locus are close to Markov. The state I acts approximately (but not exactly) as a renewal state of the process. Proportionately, the probability that under the Markov assumption deviates most from the true value is that for the trio of states (I, N, I). Conditional on non-*ibd* at the center locus, the probability of I at the third locus is substantially increased by knowledge of state I at the first. The reason for this is that the states of $\mathbf{S}$ resulting in I are few and clustered in the total space of $\mathbf{S}$-values. For a fuller discussion of this see Thompson (1994*a*). The non-Markovian nature of I and N holds even for simpler pedigrees. It may seem that the differences in the probabilities are small, and substantial only for the state of very small probability. However, depending on the phenotypic data, states of low prior (pedigree) probability may have high probability conditional on the phenotypic data.