Chapter 11

Other Monte Carlo Likelihoods in Genetics

11.1 Improving pedigree samplers

The ways in which MCMC samplers can be extended, combined, and improved, are almost limitless. One method has been discussed in section 10.6. Where the pedigree is not too complex, so that the L-sampler is feasible (and practical), combining the L-sampler and M-sampler on extended pedigrees can achieve more robust and reliable results with higher Monte Carlo precision (Heath and Thompson, 1997). The M-sampler (section 8.4) does not suffer poor mixing due to tightly linked loci, but can mix poorly where there are extended ancestral paths of descent in a pedigree. Additionally, the M-sampler may not be irreducible. Since the L-sampler is irreducible (section 8.3), combination of the L-sampler and M-sampler can ensure irreducibility, as well as improve mixing. Whereas the L-sampler is often the more computationally intensive, and seems to take longer to achieve stable probability estimates, the M-sampler may simply fail to sample the part of the space containing the majority of the probability mass (Table 11.1). The examples of section 10.6 all combined L and M steps with the same probability (20%) that any given step is an L-step. Obviously, there is scope for other patterns of systematic or random resampling.

There are ways to improve the meiosis sampling itself. Updating all indicators at a meiosis jointly shows much improved performance over single-site updating (Thompson and Heath, 1999). Moreover, updating by- meiosis avoids problems of poor mixing due to tight linkage. However, clearly there would be greater improvement if the vectors $S_{i,\bullet}$ for several meioses i were to be updated jointly. Likewise the L-sampler can be improved. For very tightly linked loci, single-locus updates are ineffective. However, where feasible, the L-sampler might update jointly $S_{\bullet,j}$ for several loci j. For the L-sampler, on a complex pedigree, usually no more than two or three loci can be updated jointly.

One case where updating several meioses jointly is effective and easily done is

		Update by locus	
		singly	jointly
Update by meiosis	singly	single-site: Update $S_{i,j}$. Performance poor	L-sampler: Update $S_{\bullet,j}$. Performance poor for tight linkage
	jointly	M-sampler: Update $S_{i,\bullet}$. Performance poor for extended pedigrees	LM-sampler. Improved mixing and more robust estimation

TABLE 11.1. *Single-site and joint updating schemes on a pedigree*

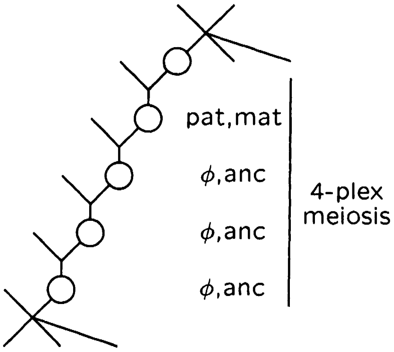

FIGURE 11.1. *A multiplex meiosis consisting of an ancestral chain of four meioses. These meioses may be jointly updated. For additional details, see text*

where there is a succession of several ancestral meioses over several generations with no phenotypic data, in each case there being one founder parent with a single offspring in the pedigree (Figure 11.1). A number of such chains may be seen in the pedigree of Figure 10.1. Recall (section 8.4) that, for such a founder parent, the meiosis to the single offspring is not scored. The relevant gene in each offspring (in this example, the paternal gene) is, in effect, a founder gene. We refer to the chain of meioses from the pedigree (non-founder) parents as a *multiplex* meiosis. For the first meiosis of the chain, we score, as usual, whether the offspring receives the parent's maternal or paternal gene. For subsequent meioses, the state is characterized by whether, at a given locus, the transmitted gene is the parent's gene from a peripheral (ϕ) or pedigree (anc) parent. The state of the multiplex meiosis is characterized by the number of meiosis indicators in the chain that currently point to a gene from one of the peripheral founders (0,1,2,3 in Figure 11.1), and the state (0=maternal, 1=paternal) of the first meiosis. With this specification, the transition probabilities remain first-order Markov along the chromosome. A pre-processing of the pedigree, assigning each multiplex meiosis its appropriate Markov transition probabilities, can greatly improve efficiency of the MCMC. There are fewer (multiplex) meioses to resample: in the example we have replaced four meioses with a total of $2^4 = 16$ states by a single multiplex meiosis with $2 \times 4 = 8$ states. For example, the state $(1, 2)$ would denote that the first individual of the chain receives her mother's paternal gene, and that for 2 of the subsequent meioses the offspring receive their mother's founder gene. (In Figure 11.1, the founder parents are male, and the pedigree parents are female.) Although this single factor of two in the number of states is not large, repeated over a large pedigree this can lead to a significant reduction. More important than the number of states is the mixing of the MCMC. Even when transition probabilities for descent down the chain are small, with joint updating alternative descent paths for an allele are more readily sampled. The ability to change the descent path down the whole chain in a single MCMC step greatly improves mixing.

The joint updating of meioses may be carried further. The Lander-Green algorithm for exact computation can be readily performed on up to 15 meioses. While one might not want to incorporate such an intensive computation into an MCMC, computation is quite feasible for, say, a subset of $m^* = 10$ of the total set of m meioses in the pedigree. The procedure is exactly as in equations (8.8) and (8.9), except now that, instead of the two values of s, $Q_\ell(\underline{s})$ must be evaluated and normalized for each of the 2^{m^*} vectors of the indicators for these m^* meioses, say 1024 values to be stored for each locus along the chromosome. Additionally the penetrance probabilities $P(Y^{(j)} \mid S_{\bullet,j})$ would be needed for each of the 1024 values. The extent to which improved mixing compensates for the increased computation remains to be investigated, but there is no doubt that joint updating will help in some cases. When the L-sampler in infeasible due to extreme complexity of the pedigree, joint updating of several meioses could ensure irreducibility of the meiosis sampler. However, this is an area where many open questions remain. In particular, on an extended pedigree, appropriate choice of the meioses to be updated jointly is far from obvious.

11.2 Interference by Metropolis-Hastings

In the absence of interference, but where different meioses exhibit different recombination probabilities, the procedure of resampling a whole meiosis jointly over loci (section 8.4) is more convenient than other forms of MCMC. Sex-specific maps can be routinely incorporated, provided they are known, and no assumptions regarding the relationship between male and female recombination frequencies are necessary. Each meiosis is resampled, and the relevant computations made, under the map appropriate to that meiosis. For multiplex meioses (section 11.1), which may contain individuals of different sexes, male and female meioses must be accounted separately, and the transition probabilities must be pre-computed, but there is no intrinsic computational difficulty.

Genetic interference in meiosis (Chapter 5) is a more complex issue, since it destroys the first-order Markov conditional-independence structure of the meiosis indicators along a chromosome. The assumption of first-order dependence in the $S_{\bullet,j}$ is crucial to the computations of sections 6.1 and 7.1 and to the M-sampler as developed in section 8.4. There is no general computational algorithm for exact computation of multilocus linkage likelihoods under interference, although Weeks et al. (1993) and Lin and Speed (1996) have shown how to incorporate interference in some cases. However, as for any multilocus problem, exact likelihood computation on an extended or complex pedigree remains computationally infeasible. In fact, for exact computations under interference, the numbers of markers, and/or pedigree structures, are severely limited, and computation is cumbersome. The erstwhile practice of transforming recombination frequencies between markers using a genetic map function, and then performing a no-interference computation becomes increasingly futile as maps become denser, and marker data on observed individuals more complete.

Although genetic interference is very seldom incorporated into linkage computations, it exists in human meiosis (Broman and Weber, 2000). Failure to incorporate it can reduce the power to detect linkage (Goldstein et al., 1995). In an analysis of data at multiple tightly linked markers from actual meioses, Thompson and Meagher (1998) have shown that interference can have a significant impact on patterns of joint segregation of genes at distances of 20cM to 30cM. Using our whole-meiosis M-sampler (section 8.4), since all the meiosis indicators for all the linked loci in an entire meiosis are resampled jointly, incorporation of an interference model is feasible.

In the M-sampler, given marker data $\mathbf{Y}$ at loci $j = 1, \ldots, L$, meiosis indicators at meiosis i, $S_{i,\bullet} = (S_{i,1}, \ldots, S_{i,L})$ are realized from

$$(11.1) \qquad P(S_{i,\bullet} \mid S_{k,\bullet}, k \neq i, \mathbf{Y}) \quad \propto \quad P(\mathbf{Y} \mid \mathbf{S})\, P^{(H)}(\mathbf{S})$$

where $\mathbf{S}$ is the total set of meiosis indicators for all loci at all meioses of the pedigree, and the super-script (H) denotes the Haldane (no-interference) model. We now continue to use equation (11.1) as our proposal distribution, and add a Metropolis acceptance step (Metropolis et al., 1953), to provide the correct conditional distribution of $S_{i,\bullet}$ under interference (denoted $P^{(I)}(\cdot)$). The required Hastings-ratio $h(\mathbf{S}^{\dagger}; \mathbf{S})$ (equation 8.2) for current $\mathbf{S}$ and proposed $\mathbf{S}^{\dagger}$ is

$$\begin{aligned}
h(\mathbf{S}^{\dagger}; \mathbf{S}) &= \frac{P^{(I)}(\mathbf{S}^{\dagger}, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y})} \, \frac{P^{(H)}(S_{i,\bullet} \mid S_{k,\bullet}, k \neq i, \mathbf{Y})}{P^{(H)}(S^{\dagger}_{i,\bullet} \mid S_{k,\bullet}, k \neq i, \mathbf{Y})} \\
&= \frac{P^{(I)}(\mathbf{S}^{\dagger}, \mathbf{Y}) P^{(H)}(\mathbf{S}, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y}) P^{(H)}(\mathbf{S}^{\dagger}, \mathbf{Y})} \\
&= \frac{P(\mathbf{Y}|\mathbf{S}^{\dagger}) P^{(I)}(\mathbf{S}^{\dagger}) P(\mathbf{Y}|\mathbf{S}) P^{(H)}(\mathbf{S})}{P(\mathbf{Y}|\mathbf{S}) P^{(I)}(\mathbf{S}) P(\mathbf{Y}|\mathbf{S}^{\dagger}) P^{(H)}(\mathbf{S}^{\dagger})} \\
&= \prod_{k=1}^{m} \frac{P^{(I)}(S^{\dagger}_{k,\bullet})}{P^{(I)}(S_{k,\bullet})} \, \frac{P^{(H)}(S_{k,\bullet})}{P^{(H)}(S^{\dagger}_{k,\bullet})} \\
&= \frac{P^{(I)}(S^{\dagger}_{i,\bullet})}{P^{(I)}(S_{i,\bullet})} \, \frac{P^{(H)}(S_{i,\bullet})}{P^{(H)}(S^{\dagger}_{i,\bullet})}.
\end{aligned}$$

recombination	prob under	prob under	prob under	prob	prob
patterns	model I	model II	model 0	ratio	ratio
	$d = 25.54\text{cM}$	$d = 0.2\text{cM}$	$d = 25.54\text{cM}$	(I)	(II)
	$\rho = 0.2554$	$\rho = 0.2$	$\rho = 0.2$		
r r r r	0.0	0.0	0.0016	0.0	0.0
r r r n, n r r r	0.0	0.0	0.0064	0.0	0.0
r r n r, r n r r	0.0027	0.0	0.0064	0.422	0.0
r r n n, n n r r	0.0027	0.0	0.0256	0.106	0.0
r n r n, n r n r	0.1196	0.05	0.0256	4.672	1.953
r n n r	0.0054	0.05	0.0256	0.212	1.953
n r r n	0.0054	0.0	0.0256	0.212	0.0
n n n r, r n n n	0.1223	0.1	0.1024	1.194	0.977
n n r n, n r n n	0.125	0.15	0.1024	1.221	1.465
n n n n	0.2446	0.35	0.4096	0.597	0.854

TABLE 11.2. *Probabilities of recombination (r) and non-recombination (n) in four equal marker intervals, under interference models I and II and under the Haldane model of no interference (model 0)*

The acceptance probability is then $a = \min(1, h(\mathbf{S}^\dagger; \mathbf{S}))$. This considerable reduction in the expression for $h(\mathbf{S}^\dagger; \mathbf{S})$, and consequent ease of computation of the acceptance probability relies on three facts:

(1) the probability of data $\mathbf{Y}$ given meiosis pattern $\mathbf{S}$ or $\mathbf{S}^\dagger$ does not depend on the interference process (I) or (H), giving rise to $\mathbf{S}$,

(2) the independence of meiosis patterns $S_{k,\bullet}$ at different meioses k (when not conditioned on data $\mathbf{Y}$), and

(3) $S^\dagger_{k,\bullet} = S_{k,\bullet}$ for $k \neq i$.

As an example, consider again the standard test pedigree (Figure 1.1): this example was also given in Thompson (2000*a*). As in section 10.4, consider five equispaced marker loci, 25.54cM apart (recombination frequency 20% under the Haldane no-interference model). We consider the case of extreme position interference, but no chromatid interference, in which chiasmata on the underlying tetrad are equispaced at 50cM spacing. Then using the notation of section 5.3 for the indicator vectors $\mathbf{C}$ of presence (1) or absence (0) of chiasmata in the four intervals there are only 5 possible values: $\mathbf{C} = (1,0,1,1)$, $(1,1,0,1)$, $(0,1,1,0)$, $(1,0,1,0)$ or $(0,1,0,1)$. Under a model which places the first marker uniformly in an interval between two chiasmata, these five possible chiasmata indicator vectors have probabilities 0.0216, 0.0216, 0.0216, 0.4676 and 0.4676 respectively. Using equation (5.2), these translate to the probabilities of patterns of recombination (r) or non-recombination (n) given in under model (I) in Table 11.2. In this table, pairs of vectors having the same probability under any model are listed together. For example, for equispaced markers, patterns *rrrn* and *nrrr* have the same probability, by symmetry. The tabulated probability refers to the probability of each of the two patterns. We see there are substantial differences in the probabilities under this interference model (I) and under no interference (model 0; Haldane), However, the ratios are not so extreme as to make the MCMC ineffective. All probabilities are strictly positive under the proposal (Haldane) distribution, and non-zero ratios differ by a factor of at most 22 (0.212 to 4.672).

It is not clear that the correct assessment of interference effects should be through imposing equal genetic distance; that is, total expected numbers of crossovers. If instead we constrain the recombination frequency between adjacent markers to be 20%, the distance under our model of complete position interference is 20cM. Again there are five possible indicator vectors **C** of chiasmata presence/absence, but this time these are $(0,0,1,0), (0,1,0,0), (0,1,0,1), (1,0,0,1)$, and $(1,0,1,0)$, each having probability 0.2. Since chiasmata have an exact 50cM spacing and the marker intervals are 20cM, it is not longer possible for there to be chiasmata in two adjacent intervals. Again assuming the first marker is randomly and uniformly placed relative to the chiasmata, and using equation (5.2), the corresponding probabilities of patterns of recombination/non-recombination are as given for model (II) in Table 11.2. Again the ratios, for model (II) relative to the proposal (Haldane) model are not extreme; this time the non-zero ratios range only from 0.854 t0 1.953. Although both models (I) and (II) have some recombination vector events of probability 0, this does not lead to invalid estimates. If proposed, these vectors will not be accepted. The total probability under the Haldane model of recombination vectors that cannot be accepted under the interference models is not large (0.014 under model I, 0.104 under model II).

Gene *ibd* pattern	single-locus prior	single-locus conditional	no interference marker M5	marker M3
All 4 genes *ibd* genes	29	133	127	180
3 of 4 genes *ibd*	156	286	356	381
2 pairs of *ibd* genes	84	154	118	130
2 of 4 genes *ibd*	484	354	303	251
all 4 non-*ibd*	247	73	96	58
mean log-probability $\log P_\theta(\mathbf{S}))$			-44.69	
mean log-probability $\log P_\theta(\mathbf{Y}\|\mathbf{S})$			-33.23	
MCMC steps (accepted %)			10^7(100%)	

TABLE 11.3. *Gene ibd probabilities* (×1000) *for single loci, and under no interference (Haldane model)*

The marker data at each locus assumed are as in sections 10.4 and 10.6 for the five individuals of the pedigree with marker phenotypes observed (Figure 10.3). The allele frequencies are again assumed to be 0.2, 0.2, 0.4, and 0.2 for the four alleles at each locus. Again, sampling latent meiosis indicators **S** conditional on the marker data, we score gene *ibd* probabilities among the four potentially distinct C alleles. In Table 11.3 are shown the gene *ibd* probabilities for single loci, and for linked markers under the Haldane model of no interference for the central marker M3 and an end marker M5. These are the same values seen for marker loci in

Gene *ibd* pattern	Model I		Model II	
	marker M5	marker M3	marker M5	marker M3
All 4 genes *ibd*	74	104	101	152
3 of 4 genes *ibd*	305	339	334	370
2 pairs of *ibd* genes	94	114	106	126
2 of 4 genes *ibd*	349	321	327	276
all 4 non-*ibd*	178	122	132	76
mean log-probability $\log_e P_\theta(\mathbf{S})$	-50.74		-45.23	
mean log-probability $\log_e P_\theta(\mathbf{Y}\mid\mathbf{S})$	-34.12		-33.58	
MCMC steps (accepted %)	10^7(68.6%)		10^7(80.4%)	

TABLE 11.4. *Gene ibd probabilities (×1000) under the recombination pattern probabilities given for interference models (I) and (II) in Table 11.2. Each run consisted of 10,000,000 whole-meiosis Gibbs/Metropolis updates, and took about 1 hour CPU on a DEC Alpha 400-233 work-station with 256MB memory*

Table 10.3 in the case of a trait locus unlinked to the markers: they are shown again here for easier reference in the context of interference effects. The *prior* is the probability given by the pedigree alone, without marker data. The *conditional* is the probability when the marker phenotypes are assumed for a single locus. The table shows that the data increase probability of gene *ibd*—not surprisingly since the four genes scored are of the same allelic type. Having data at five linked markers reinforces the inference of gene *ibd*, particularly for the marker $M3$ in the center of the map. Note that the marker spacing is 25.54cM, so that the five loci extend over 1 Morgan. In every meiosis of the pedigree there is a probability 0.5904 of at least one recombination among these five markers. Even so, the concordant data at these linked markers reinforces probabilities of gene *ibd*.

The results of 10^7 MCMC meiosis resamples are given in Table 11.4. We see a substantial effect of interference on the conditional probabilities of gene *ibd*. In particular, probabilities that all four C alleles are *ibd* are reduced, and that all are distinct are increased. The percentage of MCMC proposals accepted and the expected base-e complete-data log-likelihoods both provide an indication of the effect of interference. In comparison to the non-interference case, matching recombination frequencies (model (II)) provides closer results than does matching genetic distances (model (I)).

The interference models considered in this section are extreme, assuming complete position interference, although no chromatid interference. Other less extreme examples still show substantial impact on genome sharing among relatives at distances of 20 to 30cM. For example, Browning and Thompson (1999) considered the aunt-niece-sibs example of section 4.5, using a chi-square model with parameter $m = 2$ for the interference process (example (4) of section 5.7). Although the impact of interference on genetic inferences remains a little studied area, the results here suggest that further study is warranted. Although interference will have little impact on mapping Mendelian traits when markers are highly informative, it will affect the resolution of genes contributing to quantitative traits or to disease liability. Its impact will also be greater in using tightly linked but less informative markers, such as Single Nucleotide Polymorphisms (SNPs): see section 1.1. For such markers, haplotypes cannot be readily inferred, even with data on pedigrees, and interference will affect the imputation probabilities for such haplotypes.

11.3 Inference of typing or pedigree error

Throughout this monograph, we have focused on the case where the pedigree relationship among individuals is known, and often where the marker map and other parameters of the model for the marker data are assumed known without error. We have also not explicitly considered the possibility of errors in marker genotypes. However, as noted in section 1.4, the probability of data under a known genetic model is a likelihood for the pedigree relationships among the individuals. Also, on an assumed pedigree it is clearly possible to address other aspects of the model for the data, including possible typing errors. In analyses of real data, errors, uncertainty, or heterogeneity in the marker model often arise and may have an impact on inference. Traditionally, the approach has been to correct for errors in advance of other analyses, usually on a marker-by-marker basis. This can be unsatisfactory (Broman, 1999), and with greater automation of marker genotyping it becomes important to have methods of analyzing multilocus marker data, and allowing within the analysis for possible errors in typing or specification of individual relationships.

For inference of possible data error, the general method is simply one of generalizing the model for the relationship between underlying latent variables $S_{\bullet,j}$ or genotypes $G_{\bullet,j}$ at locus j, and the observable data $Y_{\bullet,j}$ on the individuals. The likelihood is most easily considered as in equation (3.9) or (6.1):

$$\begin{aligned}\Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{S},\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{L} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \prod_{j=1}^{L} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right). \qquad (11.2)\end{aligned}$$

The dependence structure (Figure 6.1), and hence general Baum-algorithm computational approach (section 6.1) remain unchanged. The generalization is only in $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ for each locus j. It may be easier to consider likelihood computation with an additional layer of latent variable—the true genotypes determined by the underlying pattern of gene *ibd* (Kumm et al., 1999):

$$(11.3) \qquad \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) = \sum_{G_{\bullet,j}} \Pr(Y_{\bullet,j} \mid G_{\bullet,j}) \Pr(G_{\bullet,j} \mid S_{\bullet,j}).$$

Assuming typing errors are individual-specific

$$\Pr(Y_{\bullet,j} \mid G_{\bullet,j}) = \prod_i \Pr(Y_{i,j} | \, G_{i,j})$$

a product over individuals i. Nonetheless, computation of (11.3) and hence the likelihood (11.2) can become computationally intensive for general error models and more than a very few individuals. In principle, the likelihood (11.2) can be used to estimate the form and the parameters of the error model. More practically, the reverse Baum algorithm (section 7.1) can be used to to determine the loci at which there is a high probability of error given all the observed data: that is, $\Pr(G_{i,j} \neq Y_{i,j} \mid \mathbf{Y})$ is large.

Under a given penetrance model, the likelihood of alternative relationships can be compared. Boehnke and Cox (1997) used the Baum algorithm to compute likelihoods for alternative sib and half-sib relationships from multilocus marker data. Browning (1999) extended this to a variety of extended-family relationships, up to second cousins. On larger pedigrees, in principle at least, MCMC may be used to obtain a Monte Carlo likelihood ratio. Since the likelihood is given by equation (11.2), we have the likelihood ratio equation (9.1) which, in the present context becomes

$$\frac{P_1(\mathbf{Y})}{P_2(\mathbf{Y})} = \mathrm{E}_1\left(\frac{P_2(\mathbf{Y},\mathbf{S})}{P_1(\mathbf{Y},\mathbf{S})} \;\Big|\; \mathbf{Y}\right)$$

where the subscripts on probabilities and expectations designate two alternative relationship hypotheses. Any of the MCMC samplers of earlier sections can be used to sample from

$$P_1\,(\mathbf{S} \mid \mathbf{Y}) = \frac{P_1(\mathbf{Y},\mathbf{S})}{P_1(\mathbf{Y})} \propto P_1(\mathbf{Y} \mid \mathbf{S})P_1(\mathbf{S}).$$

Care is needed in implementing these likelihood ratio estimators, since different relationships may imply a different number of relevant meioses. Unlike in the comparison of different genetic models, the penetrance probabilities $P_1(\mathbf{Y} \mid \mathbf{S})$ may depend on the hypothesized relationship. Nonetheless, we must consider MCMC sampling of $\mathbf{S}$ not of *ibd* patterns $J(\mathbf{S})$, although the latter are more readily compared for alternative relationships. In the assumed absence of interference, the segregation process $\mathbf{S}$ is Markov along the chromosome, but the agglomerated process $J(\mathbf{S})$ is not (section 4.8).

In any give meiosis, there are relatively few changes in $S_{i,j}$ as j changes. As the number of linked marker loci becomes very large and they are thus tightly linked, it becomes inefficient to use the complete set of components of $\mathbf{S}$ as the latent variables, and also difficult to get effective samplers on this space. Instead, one may consider a set of latent processes $S_i(z)$ where z is the position on the chromosome measured in terms of genetic distance. This framework was first developed by Donnelly (1983), and used by Bickeboller and Thompson (1996*a*; 1996*b*) to study the descent of genome in small pedigrees. Browning (1998) used the same underlying model to develop importance-sampling methods of estimating Monte Carlo likelihoods for alternative pedigree relationships. Browning (1999) extended the approach to the development of Monte Carlo likelihood methods to distinguish between alternative models of meiosis and genetic interference, including the models discussed in sections 5.6 and 5.7.

11.4 Other Monte-Carlo procedures for linkage analysis

Another broad area of linkage analysis not addressed in this monograph is the mapping of loci contributing to quantitative traits, or quantitative trait loci (QTL). For linkage designs in experimental organisms there are well developed methods for detecting, mapping, and resolving the QTL contributing to increasingly complex traits (Knott and Haley, 1992; Zeng, 1994; Long et al., 1995). Increasingly, on larger or more complex problems MCMC is used (Hoeschele, 1994; Sorensen et al., 1995; Satagopan et al., 1996). Heath (1997) developed methods of segregation and linkage analysis on extended pedigrees, for models involving multiple QTL contributing additively to a complex quantitative trait.

There are two main differences between MCMC methods for QTL analysis and the methods developed in this monograph. First, a Bayesian approach (section 2.4) is normally taken. For complex models, with many nuisance parameters, a likelihood approach has limitations. The traditional likelihood approach has been to maximize over these parameters, obtaining a profile likelihood for the parameters of interest. However, a Bayesian approach which integrates or samples (in the case of Monte Carlo) over nuisance parameters may provide a better reflection of the true information regarding parameters of interest. Using MCMC, samples are realized from the posterior probability distributions of parameters. A disadvantage of a Bayesian approach is that there is no exact computational approach against which MCMC results can be compared. As seen in section 10.6, even our best MCMC samplers need tuning to produce accurate likelihood estimates. For a Bayesian posterior probability distribution for parameters of a complex model, there is no way to assess the accuracy of Monte Carlo results. There is also no standard interpretation of findings. Whereas there may not be unanimity regarding the exact meaning of a base-10 lod score of 3.5, say, there is no collective experience at all regarding, say, a finding of 97% probability that at least two QTL contribute to a trait.

A second major difference between the methods of this monograph and MCMC methods for QTL analysis also relates to the model complexity, but to its effect on the MCMC methods used. For a model such as that of Heath (1997) in which the number of QTL contributing to a trait can vary, the dimension of the model is not fixed. In sampling over the parameters of a varying number of QTL, the number of parameters sampled changes. Thus methods of reversible-jump MCMC (Green, 1995) must be used, to sample between models of varying dimension.

11.5 Monte-Carlo likelihoods in population genetics

One of the earliest uses of MCMC in genetic analysis was not on pedigrees, but on the evolutionary history of populations and species. Since data are normally observed in the present, forwards simulation of the evolutionary process is of limited use in developing Monte Carlo inference procedures. Just as on a pedigree, effective realizations must be conditioned on the data. Kingman (1982) developed the theory of the coalescent, which allows for study of the ancestry of a current sample from a population. Kuhner et al. (1995) developed Monte Carlo likelihood methods for estimating evolutionary parameters, based on MCMC resampling of coalescent ancestries of the current population sample. Griffiths and Tavaré (1994*b*; 1994*a*) also developed a Monte Carlo likelihood approach to similar problems. Their approach uses importance sampling (section 7.3) rather than MCMC, and they realize successive events in the ancestry of a current sample. Stephens and Donnelly (2000) have given a recent synthesis, discussion, and extension of these two approaches.

More recently, Monte Carlo likelihood approaches have been used in a wide variety of population-genetic areas. One of these is the development of Monte Carlo likelihood methods for fine-scale mapping. Due to the limited number of meioses, the resolution of loci from pedigree data is no finer than about 1 cM (Boehnke, 1994). As described briefly in section 4.6, allelic associations resulting from slow decay of initial linkage disequilibrium between a new mutation and a tightly linked marker locus can provide evidence for linkage and for precise localization of a disease locus. This has been a recent focus of several successful mappings of loci with rare recessive disease alleles (Cox et al., 1989; Hästbacka et al., 1992; Goddard et al., 1996). The current marker haplotypes of chromosomes carrying disease alleles are the outcome of their patterns of shared ancestry, and recombination events occurring in the meioses of that ancestry.

The first attempt at Monte Carlo likelihood analysis for this problem (Kaplan et al., 1995) used forwards simulation of the population, but suffered again from the disadvantage of being unable to condition effectively on current data. The methods of Rannala and Slatkin (1998) and Graham and Thompson (1998) use Monte Carlo realization of the coalescent ancestry of the disease sample as the basic tool in obtaining a Monte Carlo likelihood for fine-scale localization of a rare allele. Note that the ancestry of a sample ascertained for a rare allele is quite different from that of a random sample from the population. There is a very strong ascertainment effect: Griffiths and Tavaré (1998) provide applicable results.

In the case of Graham and Thompson (1998), recombinations relative to the putative disease locus are then realized on the ancestry, and exact computational methods used to compute the likelihood contribution of a given recombination history. For a single marker at recombination frequency ρ to the disease locus

$$L(\rho) \;=\; P_{\mathbf{q},\rho,\Pi}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}) P_{\rho,\Pi}(\mathbf{X}) \tag{11.4}$$

where $\mathbf{X}$ denotes the latent variables of coalescent ancestry at the disease locus, and recombination events between disease locus and marker occurring in the meioses of that ancestry. The nuisance parameters are marker allele frequencies $\mathbf{q}$ which enter only into the penetrance probability $P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})$, and Π the parameters of the demographic history of the population. Note the similarity of equation (11.4) to those of likelihoods on pedigrees, for example equations (1.5), (3.9), or (7.8). However, unlike the Monte Carlo likelihoods based on those equations, here

$P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})$ is computed exactly, while the latent variables $\mathbf{X}$ are realized from their distribution under the given population model and hypothesized recombination frequency ρ. Thus a direct Monte Carlo estimate of the likelihood (11.4) is obtained.

Between the time-scale of evolution and coalescent ancestry and that of meioses in a defined pedigree, are the population-genetic models that provide probability distributions for the change of allele frequencies over generations, due to migration, population admixture, and random genetic drift. Here also, Monte Carlo methods of likelihood computation may be applied, the data $\mathbf{Y}$ being allele sample counts for different alleles, at different generations, and the latent variables $\mathbf{X}$ being the underlying true allele counts. Parameters of interest are those that determine the rate of change of allele frequencies, including the effective population size. Estimation of effective population size is of interest in the assessment of endangered populations. The dependence structure is identical to that of Figure 6.1. Instead of first-order Markov dependence of meioses at loci along a chromosome, we have first-order Markov generation-to-generation transitions of allele frequencies. The samples $\mathbf{Y}_j$ taken at a given generation j depend only on the allele frequencies $\mathbf{X}_j$ at that time. Equation (6.1) gives the form of the likelihood. Anderson and Thompson (1999) have used MCMC to obtain Monte Carlo likelihoods for the problem of estimating effective population size.

At every level, genetics provides examples of clearly defined highly structured probability models. The latent variables of genetics are "real": meioses, genotypes, recombination events, allele counts, and ancestral history. Monte Carlo methods are well suited to these problems, and often exact computation of likelihoods and probability distributions is infeasible. This final chapter has described a number of areas in which these methods are being applied, beyond those of linkage analysis from pedigree data which has been the focus of earlier chapters. These are only a few current examples; doubtless others will follow.