Chapter 10

Case studies using the M- and LM-samplers

10.1 Background to a study

Much of the material in sections 10.1 to 10.5 has recently been published (Thompson, 2000*a*). It was presented at a one-day Royal Statistical Society conference in March 1999, and was discussed again in July 1999 at the CBMS Summer Course. Section 10.6 is the result of more recent work.

First, the methods of the previous chapters are illustrated using data based on an extended Icelandic pedigree, provided by Dr. J. H. Edwards. The trait, apparent in three families, was thought to be a simple recessive, with an animal analogue suggesting a possible location on human Chromosome 1 (Remmers et al., 1996). However, findings were negative, and for purposes of illustration Heath and Thompson (1997) simulated marker data, conditional on a recessive trait locus in the chromosomal region. The resimulation of data assumed the same marker locations, population allele frequencies, and marker phenotype availability as in the original data. Marker data were simulated conditional on descent paths at the trait locus that implied that the four affected final individuals would be so. No phenotypic assumptions were imposed for other pedigree members. Using these simulated data, there was some evidence for excess gene identity by descent among the six parents of affected individuals (Heath and Thompson, 1997). However, in attempting to analyze these simulated data, under the assumption of a rare recessive trait, findings were ambiguous, primarily due to the fact that no founders were ancestral to more than three of the six parents of the affected individuals, even though the ancestry of the families was fully traced for seven generations. Accounting for the affected individuals required three separate origins of the recessive disease allele within the pedigree. For current purposes, we have therefore also modified the pedigree structure, making possible a single ancestral origin of the disease allele, and realized disease ancestry accordingly (Figure 10.1).

Conditional on the realized gene ancestry, we have resimulated marker data.

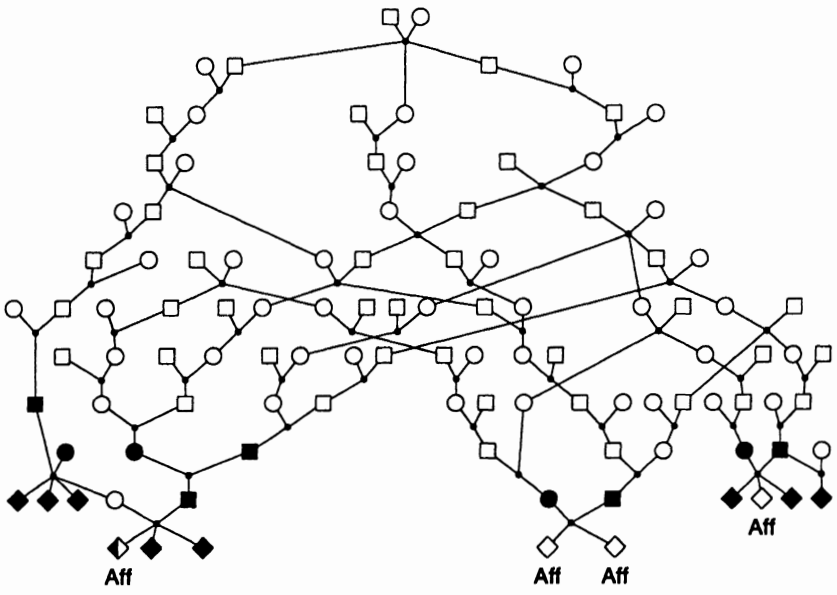

FIGURE 10.1. *The modified Icelandic pedigree. The four individuals marked "Aff" are affected. Those shaded black have marker data available at the majority of the 17 marker loci. The affected half-shaded individual is typed at only two of the marker loci*

Since the data are simulated, we avoid difficulties caused by errors in marker map or in meiosis model assumptions –for example we did not incorporate recombination heterogeneity between the sexes. The marker allele frequencies and data availability are as in the original data. Very few data are available on the affected individuals themselves (two markers on only one of the four cases), and overall the pedigree is quite sparsely observed, with the data being on the majority of the close relatives of affected individuals (Figure (10.1)).

There are data at 17 marker loci, some of which are quite polymorphic, exhibiting up to 7 alleles, even among the 18 observed individuals. Some were also tightly linked: indeed the ones adjacent to the putative trait locus were less than 2cM apart. In simulating marker data, using the original map, we obtained haplotype sharing among the three nuclear families containing affected individuals over 5 or 6 markers. To have data corresponding to modern linkage detection problems, we resimulated data using a genetic map with marker intervals at 10% recombination frequency, with the disease locus at the center of one interval (recombination 0.0528 to each flanking marker). Some realizations then gave almost no genes *ibd* among the three families at any marker locus, with obvious consequent problems for linkage detection. The necessary scale of the map is dependent on the pedigree structures

locus	number of alleles	true *ibd* state	true phenotypes
trait	2	a a a a a a	222222
M1	6	b d - m k z	226166
M2	7	b d g m k z	575373
M3	6	b d g m x z	656155
M4	6	b d g m x z	364442
M5	4	b e h m v z	333333
M6	3	b e h m v t	113311
M7	6	b e i a x t	643624
M8	7	b e i a n t	445426
M9	7	b a - a - a	576677
M10	6	a a a a a a	444444
M11	8	b a w w w y	378887
M12	4	b a v - w y	142123
M13	5	b f w - - y	312425
M14	6	b f w p x y	125323
M15	7	b e m p x y	156447
M16	7	b - n r x z	727324
M17	7	- - r w x z	136344

TABLE 10.1. *True gene identity by descent simulated on the modified Icelandic pedigree*

available for analysis (Thompson, 1997). However, our chosen data realization did exhibit a gene *ibd* in all six parents of affected individuals at one of the two markers flanking the disease locus. Since the data are simulated the "true" trait location is known; this is mid-way between markers M10 and M11.

The simulated data in three affected offspring individuals are shown in Table 10.1. For true *ibd* status, each letter indicates a different founder haplotype. A founder origin occurring once only in the set of six haplotypes is denoted "-". The disease allele at the trait locus is allele "2". Note that, apart from data at two loci for one individual, the marker types of affected individuals are not observed. Observations are available only on relatives of these individuals. Thus, at locus M4, although there are only three like alleles in the six haplotypes of affected individuals, the observed data permit the possibility that four of the six genes are *ibd*.

10.2 Conditional gene *ibd* probabilities

Given the trait and marker phenotypic data, we first analyzed conditional probabilities of gene identity by descent among haplotypes segregating from each member of each of the three parent couples with affected offspring. The marker allele frequencies and recombination probabilities used in simulating the data were assumed in the analysis. The trait allele frequency was assumed to be $q = 0.001$. This low value makes very probable a single origin of the disease allele in the

Locus	Probability ×1000 All non-*ibd*				Probability ×1000 4 or more *ibd*			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
trait	937	0	0	0	0	978	997	969
M1	879	988	907	746	0	0	0	0
M2	906	999	879	306	0	0	0	0
M3	975	925	906	10	0	0	0	0
M4	874	863	808	17	0	0	0	425
M5	924	843	755	263	0	0	1	39
M6	931	726	742	532	0	0	0	2
M7	901	971	682	689	0	0	0	0
M8	919	751	414	589	0	0	0	0
M9	864	685	28	458	0	2	508	18
M10	676	539	0	387	4	7	982	30
M11	872	434	13	532	0	0	0	0
M12	879	406	180	598	0	0	0	0
M13	870	643	370	672	0	0	0	0
M14	872	867	589	773	0	0	0	0
M15	988	894	978	988	0	0	0	0
M16	947	916	963	980	0	0	0	0
M17	993	894	981	990	0	0	0	0

TABLE 10.2. *Conditional probabilities of gene identity by descent given the marker data simulated on the modified Icelandic pedigree. Shown are probabilities* ×1000. *For details of the cases (1)-(4), see text*

pedigree. There are in all 39 founders in the pedigree, and hence 78 founder genes at the disease locus, but only the four in the original couple are ancestral to all the six carrier parents of affected individuals. The 203 (potential) patterns were scored marginally at each marker. Several cases were considered:

(1) All markers and the trait locus independently segregating (unlinked). A null trait locus provides the single-locus prior probability of *ibd* given only the pedigree structure

(2) Correct map for marker data. The correct recessive trait model and affected trait status of individuals is assumed, but the trait locus is modeled as unlinked to the markers.

(3) Correct map, with trait locus in correct location between M10 and M11.

(4) Correct marker map, with trait locus in incorrect position between M3 and M4.

For the trait locus, one member of the original couple was specified to be a heterozygous carrier, and the other a non-carrier. No assumptions were made about the trait genotypes of any other founders or ancestors. The affected individuals whose haplotypes were scored were assumed homozygous for the disease allele.

The results are summarized in Table 10.2. Each column refers to the specified one of the four cases (1)-(4) given above. The table consists of probabilities multiplied

by 1000 for ease of presentation. The first set of four columns gives the conditional probabilities of no *ibd* among the six haplotypes. The second block of four columns gives the total conditional probability of *ibd* patterns in which at least four of the six haplotypes are *ibd*. The MCMC incorporates jointly the information from all linked loci, although the conditional probabilities are here summarized marginally for each locus. Since the MCMC runs jointly over loci, scoring of joint realized patterns is also possible.

The sampler used here is the M-sampler (section 8.4), so one MCMC step consists of resampling the meiosis indicators $S_{i,j}$ jointly for all 18 loci for a randomly chosen meiosis i. Each run consists of 10^7 meiosis MCMC steps, and takes about 12 hours on a workstation running a shared LSBatch system. States of *ibd* are only output if the sum over loci of the estimated conditional probability is greater than 0.001. Given the marker data, more states are thus feasible than are given in the output summary: states which were realized in the MCMC with low frequencies do not appear.

Although the marker data alone do not suggest high levels of gene *ibd* among affecteds, the conditional probability of some *ibd* among the six haplotypes in the region of the true trait location (M9,M10,M11) is high. Even independently, column (1), there is evidence of some gene *ibd* in this region, particularly at marker M10. The values in this column may be compared with the single-locus prior. Based only on the pedigree structure, the probability of no gene *ibd* is 0.937. The trait locus itself contains a lot of the information on segregation (Table 10.2). Even in the absence of marker data, the trait information reduces the probability of no gene *ibd* among the six haplotypes from 0.937 to close to 0, and increases the probability of four or more haplotypes *ibd* from close to 0 to 0.978.

When the trait locus is hypothesized in its true position, very high levels of gene *ibd* are estimated at the adjacent marker M10, while the high levels at the trait locus itself are reinforced. Disconcertingly, when the trait locus is hypothesized in an incorrect position, *ibd* at the trait locus is only slightly decreased, while estimated *ibd* probabilities at loci in the region of this incorrect position (M3, M4, M5) are much increased. The strength of the information provided by the segregation pattern of this rare recessive trait makes inference of gene location difficult. Since marker data are very sparse on the pedigree, it is possible for the marker descent patterns to adapt to alternative hypothesized gene locations.

10.3 Likelihoods and log-likelihoods

We then attempted a Monte Carlo estimate of the full location lod score, assuming each of the six parents of affected individuals to be heterozygous for a very rare recessive trait allele. However, the Monte Carlo likelihood estimation methods of Chapter 9 failed to converge. A plot of the expected base-e complete-data log-likelihoods (equation (9.14)) from this same Monte-Carlo run reveals why (Figure 10.2). The MCMC was performed at hypothesized trait locus positions γ_0 in the center of each marker interval, at positions linked but outside the span of the markers, and also with the trait locus unlinked. The complete-data log-likelihood

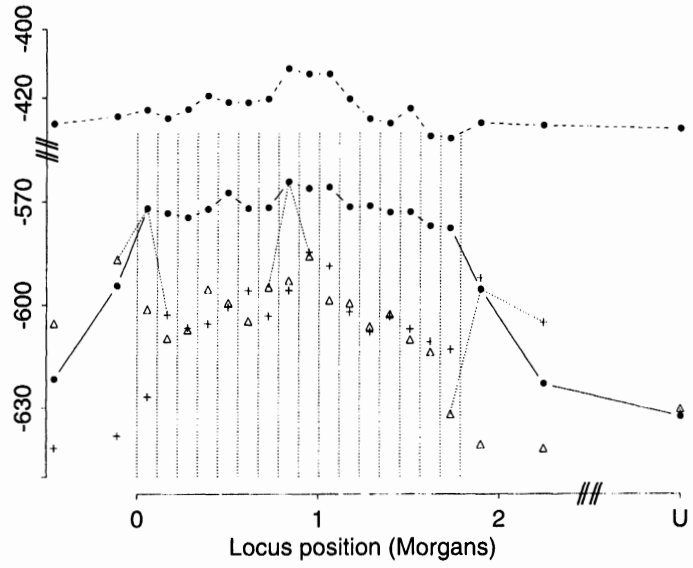

FIGURE 10.2. *Expected complete-data log-likelihood components for the simulated data on the modified Icelandic pedigree. Shown are* $\mathrm{E}_{\gamma_0}(\log_e \Pr(\mathbf{Y} \mid \mathbf{S}) \mid \mathbf{Y})$ *(upper curve), and* $\mathrm{E}_{\gamma_0}(\log_e P_\gamma(\mathbf{S}) \mid \mathbf{Y})$ *for* $\gamma = \gamma_0$ *(•, lower curve), and for* γ *to the left (△) and right (+) of* γ_0*. The location U denotes unlinked. For additional details see text*

is partitioned into segregation (equation (9.15)) and penetrance (equation (9.16)) parts. The figure shows first the penetrance contribution $\mathrm{E}_{\gamma_0}(\log \Pr(\mathbf{Y} \mid \mathbf{S}) \mid \mathbf{Y})$ to the expected complete-data log-likelihood for each simulation position (upper curve). This conditional probability does not directly depend on the hypothesized trait location, γ, although the expected log-probability does so through the realized $\mathbf{S}$. The segregation contribution $\mathrm{E}_{\gamma_0}(\log P_\gamma(\mathbf{S}) \mid \mathbf{Y})$ depends both on the simulation location γ_0, and on the evaluation location γ. The figure shows the values for each simulation position γ_0 (lower curve), with evaluations at γ_0 and at positions one step to the left (△) and to the right (+). Shown also are three example connections of realizations at a given γ_0, shown as •, with the same realizations evaluated to the left (△) and right (+). These log-likelihood differences are of order 25, indicating that $\mathbf{S}$-values realized at a given γ_0 are of order e^{25} less probable under neighboring values: it is unsurprising the Monte Carlo estimation of the likelihood is infeasible.

The expected complete-data log-likelihood is not only useful in diagnosing failure; it also provides some evidence regarding alternative models. For the four cases (1)–(4) considered in section 10.2, the complete-data base-e log-likelihoods averaged over each run are -1704, -1061, -982 and -998 respectively. Clearly, the assumption that the marker loci are unlinked (-1704) is unwarranted. The other three runs assume the correct marker map, with the trait locus unlinked, correctly positioned, and incorrectly positioned, respectively. The largest expected complete-date log-

likelihood is obtained when the model is correct, while the value under the model that the trait locus is unlinked is almost 80 units smaller. Summing the two curves, for the penetrances, $\log \Pr(\mathbf{Y}|\mathbf{S})$, and segegrations, $\log \Pr(\mathbf{S})$ in Figure 10.2, we see that the maximum expected complete-data log-likelihood is obtained for trait locations between marker M8 and marker M11. Within this range there is little discrimination, but outside these three marker intervals both segregation and penetrance contributions decrease markedly.

10.4 Gene *ibd* in a smaller example

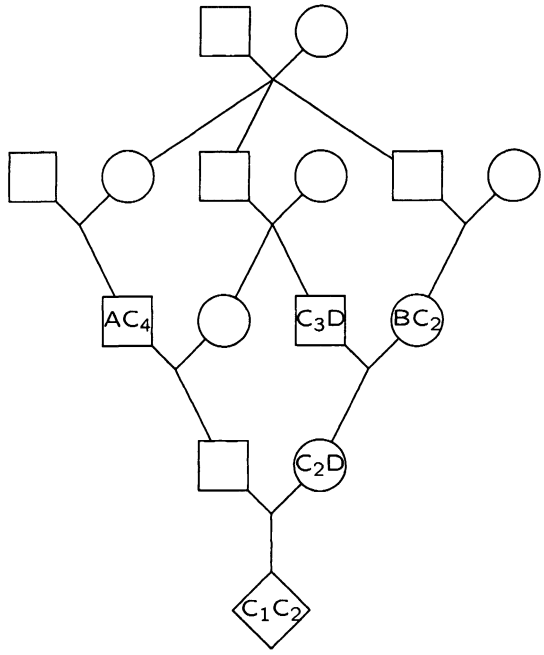

FIGURE 10.3. *Hypothetical phenotypic data assumed at each marker locus on the pedigree of Figure 1.1. The four potentially distinct C alleles are labeled C_1 to C_4*

To examine the performance of the MCMC method in more detail, we consider a smaller example, returning again to the pedigree of Figure 1.1. We suppose marker data as in Figure 3.5 at each of five marker loci, with recombination frequency 20% between adjacent markers (genetic distance 25.54 cM under a no-interference meiosis model). The trait data, for a rare recessive trait, is only that the final

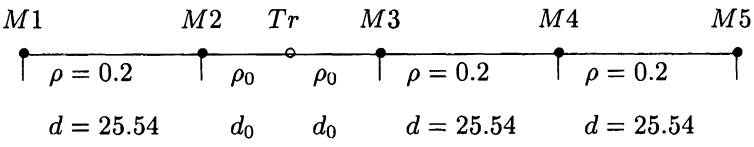

FIGURE 10 4 *Marker (M1 to M5) and trait (Tr) locations for the example of Figure 10.3. The trait locus is at the midpoint of the (M2, M3) interval, so* $d_0 = 12.77cM$ *and* $\rho_0 = 0.1187$

gene *ibd* pattern	pedigree prior	marker loci M3	marker loci M5	trait locus $q = 0.001$	trait locus $q = 0.05$
all 4 genes *ibd*	29	182	127	275	189
3 of 4 genes *ibd*	156	381	355	400	317
2 pairs of *ibd* genes	84	129	119	85	88
2 of 4 genes *ibd*	484	250	303	238	327
all 4 non-*ibd*	247	58	96	2	79
complete-data log-likelihood: segregations				-44.7	-45.1
complete-data log-likelihood: penetrances				-40.2	-37.1

TABLE 10 3. *Conditional probabilities* (×1000) *of gene ibd among the four C alleles on the pedigree of Figure 10.3, with five equally spaced marker loci, M1 to M5, and for a recessive trait unlinked to the markers*

gene *ibd* pattern	trait with $q = 0.001$ trait	M3	M5	trait with $q = 0.05$ trait	M3	M5
all 4 genes *ibd*	390	344	155	361	326	152
3 of 4 genes *ibd*	530	394	372	504	446	370
2 pairs of *ibd* genes	27	78	122	40	84	121
2 of 4 genes *ibd*	53	176	279	86	130	283
all 4 non-*ibd*	0	8	72	9	14	74
complete-data log-likelihood						
segregations		-38.5			-38.6	
penetrances		-39.1			-35.6	

TABLE 10 4 *Conditional probabilities* (×1000) *of gene ibd among the four C alleles on the pedigree of Figure 10.3, with five equally spaced marker loci, M1 to M5. The trait is now in the map, midway between M2 and M3*

individual of the pedigree is affected. Initially a trait allele frequency of $q = 0.001$ was assumed, although a value $q = 0.05$ was also considered.

At each of the five marker loci, frequencies 0.2, 0.2, 0.4 and 0.2 were assumed for alleles A, B, C, and D, respectively. Of particular interest in this test example is the potential for gene *ibd* among 4 potentially distinct C alleles, labeled C_1 to

C_4 in Figure 10.3. At each marker locus, given these marker phenotypes, all 15 possible patterns of gene *ibd* among these four C-alleles are possible. The C allele was given a relatively high frequency in order to give the possibility of four distinct origins non-negligible probability, while in contrast the trait was assumed rare to give high conditional probability that the affected individual is autozygous (has two *ibd* genes) at the trait locus.

Tables 10.3 and 10.4 summarize the conditional probabilities, when markers are run unlinked to the trait locus, and when the locus is in the mid-point of the second of the four marker intervals (Figure 10.4). Trait allele frequencies of $q = 0.001$ and $q = 0.05$ were each used. Each run consists of 10^7 M-sampler steps, each step selecting a random meiosis for update. We see a similar pattern to the example of section 10.2. Table 10.3 shows that each of the marker and trait data separately increases the conditional probability of gene *ibd* among the like alleles and decreases the probability of non-identity, relative to the prior based only on the pedigree structure. When the trait locus is within the marker map, the trait and marker data together reinforce the inference of gene *ibd* (Table 10.4). However, the effect of hypothesizing a trait location within the map on the inference of *ibd* at the marker loci is not nearly as strong as for the example of section 10.2. The effects are stronger for a rarer trait, both when unlinked (Table 10.3) and when linked (Table 10.4). However, the 50-fold change in allele frequency from $q = 0.001$ to $q = 0.05$ has a relatively minor effect. Of course, when the trait is unlinked, changing trait allele frequency does not impact marker *ibd*. When the trait is linked, the impact of trait data on marker *ibd* is larger for the adjacent marker M3 than for the terminal marker M5. The 50-fold difference in trait allele frequency (Table 10.4) has a moderate impact at the adjacent marker M3, but almost no impact at the terminal marker M5. The total complete-data log-likelihoods are larger when the trait is in the map, indicating evidence for linkage. The penetrance terms differ by about 3 between $q = 0.001$ and $q = 0.05$, the latter value giving higher probabilities.

10.5 MCMC lod score estimation

For the example of section 10.4, exact lod scores can be computed using GENEHUNTER 2 (Kruglyak et al., 1996; Kruglyak and Lander, 1998). These are shown in Figure 10.5. The two solid lines show the base-10 lod scores for trait locus position when the previous the marker data are assumed on five members of the pedigree. The higher curve corresponds to a trait allele frequency $q = 0.001$, and the lower to $q = 0.05$. The two broken curves show the base-10 lod scores when the marker data consist only of the final individual being homozygous for marker allele C with allele frequency 0.4, at each of the five marker loci. Again the upper curve is for $q = 0.001$ and the lower for $q = 0.05$. Note that the differences between the lod score curves for $q = 0.001$ and $q = 0.05$ are not large, although there is more evidence for linkage when a rarer trait frequency is assumed. This is a 50-fold change in allele frequency, and thus a 2500-fold change in the frequency of the recessive phenotype. Since there are only five founders in the pedigree, even

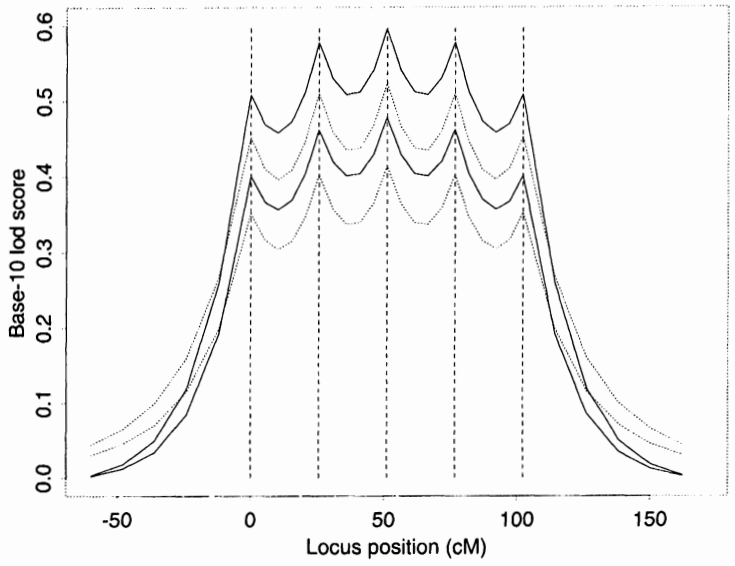

FIGURE 10.5. *Exact base-10 location lod scores computed using GENEHUNTER 2. The solid lines correspond to having marker data on five pedigree members, and the broken lines to having marker data on only the final affected inbred individual. In each pair, the upper curve corresponds to a trait allele frequency* $q = 0.001$, *and the lower to* $q = 0.05$

at this much higher trait allele frequency the probability of two separate origins of the allele in the pedigree is small.

We note that the shape of this location lod score curve is atypical, with maxima at the markers due to the assumption of the same marker data at each locus. This complete concordance of the data, and its consistency with absence of recombination between trait and markers, leads to the symmetry of the curve and to local maxima of the lod score which occur at rather than between the markers. We see that most of the information for linkage is in the data on the final individual; this is the power of homozygosity mapping for a rare recessive trait, as discussed in section 4.6. However, at loose linkage to the marker loci, the marker data on the additional four individuals do impact the lod score curve. Due to the particular marker data assumed, whereby the C_3 allele is known not to be transmitted to the final individual (Figure 10.3), lod scores are sharply decreased outside the map, and in fact are slightly negative at looser linkage.

Attempting estimation of this lod score curve, using the M-sampler as before, gave improvement over the example of section 10.3, although not fully satisfactory results. The expected complete-data base-e log-likelihoods for the case $q = 0.001$ are shown in Figure 10.6, again separated into the penetrance and segregation contributions. As before, the average log-probability of meioses sampled under hypothesized trait location γ_j is much larger under that location than under

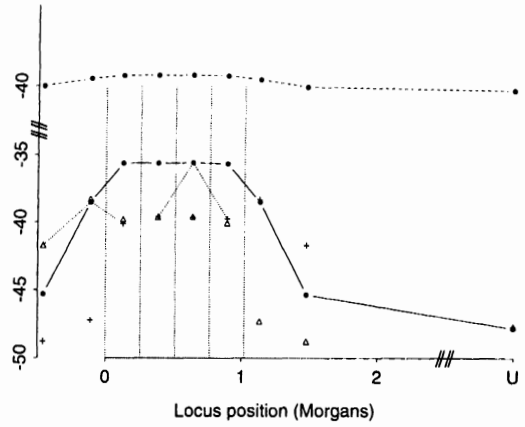

FIGURE 10.6. *Expected complete-data log-likelihoods with the hypothetical data of Figure 10.3 assumed at each of five equally spaced linked marker loci. The notation is as in Figure 10.2*

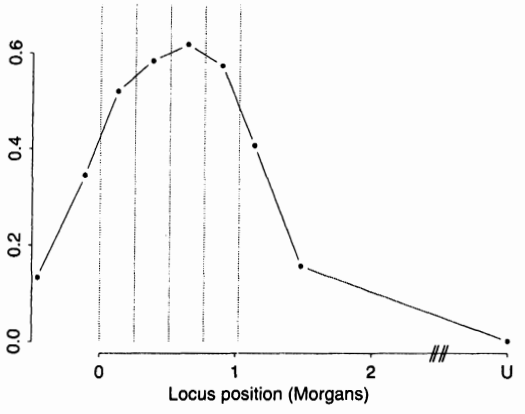

FIGURE 10.7. *Estimated Monte Carlo location base-10 lod score curve for the hypothetical data of Figure 10.3*

locations in the neighboring marker intervals to the left and right. However, now the difference is less than 5 rather than 25. Although e^5 is two orders of magnitude, estimation of the location score curve is now feasible in the sense that the methods converge to provide an estimate.

The method of equation (9.4) of section 9.2 is used, estimating only likelihood ratios at the two points adjacent to the simulation value in each estimating equation. Thus ideally, solution of equation (9.4) should provide an eigenvalue of 2. However, although gene *ibd* probabilities appear to be reliably estimated,

diagnostics suggested the sampler was not mixing well, and different runs gave substantially different lod score estimates. By comparison with exact values (Figure 10.5), the lod score was overestimated. One resulting lod score estimate is shown in Figure 10.7.

In the hope of improving performance, the assumed trait allele frequency was increased to $q = 0.05$. The true lod score is not much affected (Figure 10.5). Unfortunately, neither is the Monte Carlo estimate; curves very similar to that of Figure 10.7 were again obtained. However, the MCMC performance was more robust at the higher trait allele frequency, with much better agreement between runs. For a run giving final estimates indistinguishable from those of Figure 10.7, the relevant eigenvalue of equation (9.4) was 1.94, apparently close to the "perfect" value 2. This indicates indicates good agreement of the ratios provided by simulations at adjacent points.

Despite this apparent success, the absolute values of the log-likelihood differences are still overestimated. As seen in the next section this is primarily due to an insufficient number of simulation points for the MCMC. Additionally, the method of combining the likelihood ratio estimates into an overall lod score appears often to give a positive bias. The Monte Carlo estimator based on equation (9.1), of the ratio of the likelihood at an adjacent point to that at the simulation point, is unbiased. However, the statistical properties of the estimation method based on equation (9.4) are unclear. All that is guaranteed is that the resulting estimator of the lod score is consistent, as the number of realizations at each simulation point becomes infinite. Finally, the value 1.94, although "close" to 2, was less close than with better MCMC samplers sampling at more trait locations. Then, a value in the range 1.98 to 2.02 is typically obtained.

10.6 Better MCMC lod scores

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. Conversely, the L-sampler (section 8.3) works well on extended pedigrees, but mixes poorly with multiple linked loci. Combining the two samplers, say in the ratio of 10 M-steps to 1 L-step, can achieve more robust and reliable estimates with higher Monte Carlo precision (Heath and Thompson, 1997). The estimation of conditional *ibd* probabilities of section 10.4 was repeated using the LM-sampler, with an L-sampler proportion of 20%. This means that every step updates either a randomly chosen meiosis (M-step), or a randomly chosen locus (L-step), and each step is independently chosen to be an L-step with probability 0.2. For an equal number of total steps (in this example, 10^7), the MCMC runs took three times as much CPU, but the results of Tables 10.3 and 10.4 were unchanged both for $q = 0.001$ and $q = 0.05$. However, it is likely that the LM-sampler would achieve the same results with a smaller number of total steps.

Using an LM-sampler, and assuming now the higher trait allele frequency of $q = 0.05$, more accurate Monte Carlo lod score estimates are obtained for the example of section 10.5. Shown in Figure 10.8 as a solid line is the exact base-10

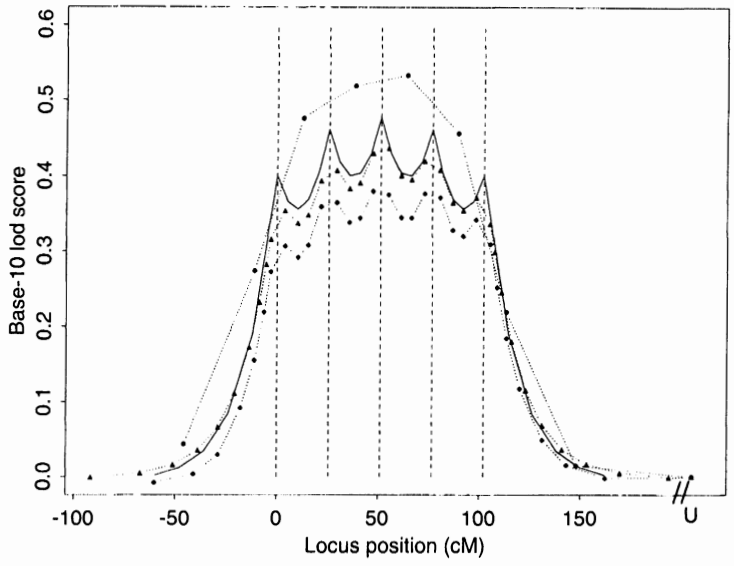

FIGURE 10.8. *Base-10 location score curves for the example of section 10.5 re-estimated, shown also with the exact value*

L-sampler probability	0.0	0.2	0.2	0.2	0.2
MCMC sample points					
unlinked	1	1	1	1	1
each end	2	2	3	7	10
each interval	1	1	3	4	4
Total	9	9	19	31	37
eigenvalue of (9.4)	1.942	1.979	1.993	1.999	2.002
MCMC realizations/point	10^6	10^6	10^6	10^6	10^6
CPU time (secs)	5,237	29,466	60,507	96,153	116,443
Shown in Figure	10.7	10.8	—	10.8 (*)	10.8

TABLE 10.5. *Summary of LM-sampler runs on the example of section 10.5. The penultimate run, designated (*), is the run also used for the results of Figures 10.9 and 10.10. The first column shows the M-sampler run discussed in section 10.5. The runs were done on a DEC alpha workstation 400-233, with 192 MB memory*

lod score computed using GENEHUNTER 2 (Kruglyak et al., 1996; Kruglyak and Lander, 1998). We note again that this lod score is atypical with local maxima at every marker, due to the assumption of the same marker data at each locus and its consistency with absence of recombination between trait and markers (Figure 10.5). This makes this lod score curve a challenge for Monte Carlo estimation, even though this pedigree is small. Also shown are three Monte Carlo estimates of the lod score,

with the MCMC done using the LM-sampler. As in previous sections, likelihood ratios were estimated only relative to adjacent trait locations, and the lod score estimation method of equation (9.4) was used to combine these into a single set of lod scores. Only lod scores at the simulation points are estimated. There is no attempt to interpolate between these points, which are connected by broken lines in Figure 10.8 for clarity only.

For clarity and easier comparison, we show here only three curves, each done with an L-sampler proportion of 20%. The MCMC is performed with the trait locus in each of the positions indicated, starting with the trait locus unlinked. When the hypothesized trait locus location is changed, the first step is to update the trait locus meiosis indicators. The initial set-up is done using the L-sampler set-up for unlinked loci (Heath, 1997). On a large pedigree, with extensive marker data, some burn-in for the linked marker loci should therefore be included, but this was ignored in this example. The marker loci were not tightly linked (see Figure 10.4).

The run characteristics and results are summarized in Table 10.5. The first column shows the M-sampler run of section 10.5 for comparison, but this curve is not shown in the figure. As can be seen from a comparison of Figure 10.7 and Figure 10.8, the results are similar when the same simulation points are chosen. This wide point spacing, with only a single point in each marker interval, leads to an overestimate of the lod score. With the LM-sampler (second column of Table 10.5), the upward bias is less, and the eigenvalue of the estimating equation increases from 1.942 to 1.979—closer to the idealized value of 2. Of greater relevance may be that the run takes almost 6 times as much CPU. On the positive side, the LM-sampler gives more consistent results. In fact, both runs were the first run at these computational settings. However, there was greater variability among runs using the M-sampler alone. With 10^6 MCMC steps at each simulation position for the trait locus, results using the LM-sampler were almost identical in repeat runs.

The three curves shown in Figure 10.8 correspond to the second and to the last two columns of Table 10.5. Other runs, including some not listed in this table gave comparable results. Using the sample LM-sampler settings, but increasing the number of points for MCMC and likelihood-ratio estimation (Table 10.5), we obtain much better lod score estimates (Figure 10.8). With more points for estimation and evaluation, the bias in the estimated lod score is reduced or even eliminated. The eigenvalue of the estimating equation becomes increasingly close to the ideal value of 2.000. All curves with several points within the marker intervals managed to mimic the atypical dips of the true curve. More difficulty was encountered in getting the precise level of the curve, relative to the null hypothesis that the trait locus is unlinked. Even with seven linked evaluation and simulation points at each end of the map (Table 10.5), there are still adjacent simulation points at which the likelihood ratio is too large to be well estimated. The final run, with 10 evaluation points at each end of the map did well, even mimicking the true very slightly negative lod scores at each end of the map when the trait locus is close to unlinked. However, even here, there is a slight asymmetry and downward bias as the trait locus crosses the first marker. All the runs show this asymmetry when the trait locus is moved from across the map from left to right, and it is reversed when the direction is reversed. Possibly, more burn-in as the trait locus gets close to the

marker loci would resolve this.

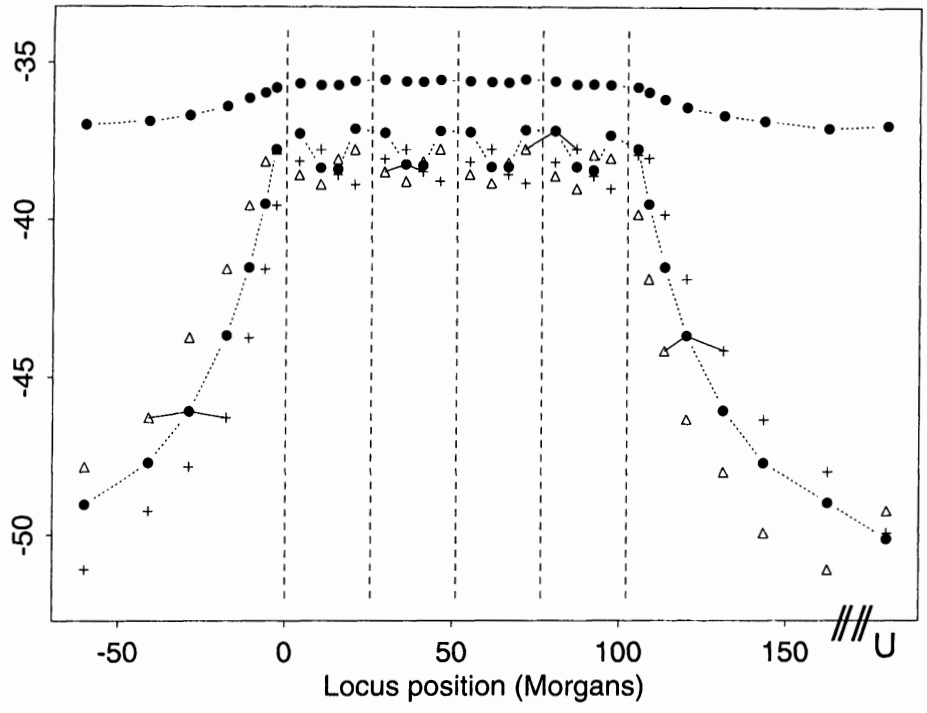

FIGURE 10.9. *Expected complete-data log-likelihoods for the example of section 10.5, shown for the penultimate run of Table 10.5. The notation is as in Figure 10.2. As in that figure, the contribution from penetrance terms is shown separately from that for segregation terms*

For a given L-sampler proportion, the CPU time is almost directly proportional to the number of simulation points, or more generally to the total number of MCMC steps. An L-step appears to take about 20 times as long as an M-step; of course, this ratio is highly data-set and pedigree dependent. For comparison purposes, all runs were done on a 1995 DEC alpha workstation 400-233, upgraded to have 192 MB memory. This machine is about four times slower than newer single-processor DEC alpha workstations with 1GB memory. In addition to computing the likelihood ratios, the program also produced the expected complete-data log-likelihoods (section 9.6) and the conditional probabilities of recombination in all intervals, in both male and female meioses. The added computational cost of producing these useful diagnostics is slight.

The expected complete-data log-likelihoods are shown in Figure 10.9, for the penultimate run shown in Table 10.5. The notation is the same as in Figure 10.2. The contribution from the penetrance terms $\Pr(\mathbf{Y}|\mathbf{S})$ is the upper curve, while

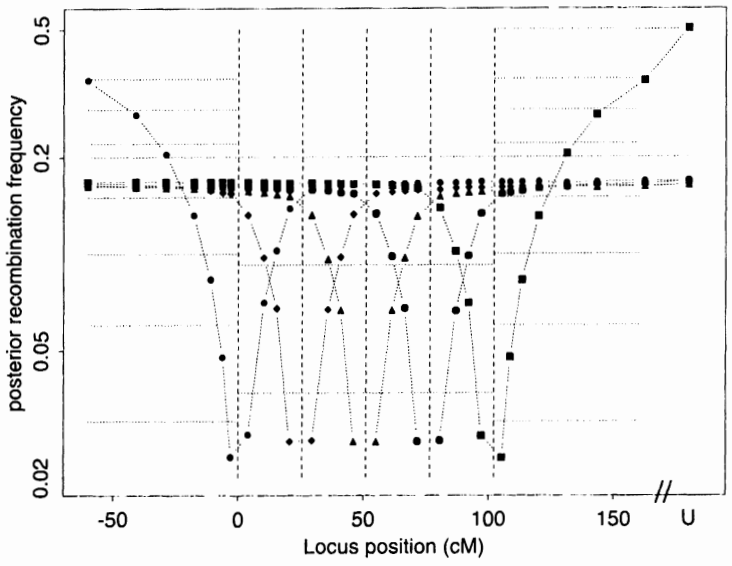

FIGURE 10.10. *Estimated conditional probabilities of recombination in the five map intervals for the example of section 10.5, shown for the penultimate run of Table 10.5. For details, see text*

the lower curve gives the expected value of $P_\gamma(\mathbf{S})$. Each point is plotted at the coordinate corresponding to the trait location γ for which the probability is evaluated. For the penetrance curve, and the main segregation-probability curve (indicated by $\bullet$), the simulation point and evaluation point are the same. A $\triangle$ indicates an evaluation point to the left of the simulation point and a $+$ indicates an evaluation point to the right. As in Figures 10.2 and 10.6, a few corresponding $(\triangle - \bullet - +)$ triplets are connected by lines. By comparison with the Figure 10.6, we see that differences are now small between evaluations at adjacent locations of the log-probabilities of realizations at a given point: the $(\triangle - \bullet - +)$ triplets. As expected, the log-probabilities are highest where the simulation point is also the evaluation point. However, for some evaluation points outside the marker map, we see that the probability is up to seven times (e^2) larger for realizations at an adjacent point than at the point itself—the vertical $(\triangle - \bullet - +)$ differences in the figure. Ideally, for accurate estimation of Monte Carlo lod score curves, both sets of log-probability differences should be small. The results suggested that more simulation points outside the marker map are needed, as also suggested by a comparison of the estimated and exact lod score curves of Figure 10.8. This led to subsequent production of the final run shown in the figure, and as the last column of Table 10.5.

Figure 10.10 shows the conditional probabilities of recombination, given the marker and trait data, for each trait location, in each of the five intervals of the marker and trait locus map. For consistency, these are shown for the same penultimate run of Table 10.5. Each symbol represents a different interval; the

interval containing the trait locus changes as the trait locus moves across the marker map. For greater clarity the frequencies are shown on a log scale. The program estimates frequencies for male and female meioses separately, but these have been combined in the current figure. Even where, as here, the prior recombination frequencies are the same in male and female meioses, the frequencies conditional on data are not. The conditional probabilities depend on the specific marker data and the gender of individuals in whose meioses recombinations are imputed. Also shown in the figure, by broken horizontal lines, are the prior recombination frequencies between markers (20%), at trait locations outside the map, and for two of the four locations for the trait locus within each marker interval. Except for an unlinked trait locus, or very loose linkage, the concordant data at all the markers and at the trait locus depresses the conditional probabilities of recombination below their prior expectation. Even with these fully concordant data, however, the conditional probabilities are not small: each is about 85% to 90% of the prior value.