12-5-2006

USE OF HIDDEN MARKOV MODELS FOR QTL MAPPING

Karl W. Broman
The Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, kbroman@biostat.wisc.edu

Use of hidden Markov models for QTL mapping

Karl W Broman
Department of Biostatistics, Johns Hopkins University

December 5, 2006

An important aspect of the QTL mapping problem is the treatment of missing genotype data. If complete genotype data were available, QTL mapping would reduce to the problem of model selection in linear regression. However, in the consideration of loci in the intervals between the available genetic markers, genotype data is inherently missing. Even at the typed genetic markers, genotype data is seldom complete, as a result of failures in the genotyping assays or for the sake of economy (for example, in the case of selective genotyping, where only individuals with extreme phenotypes are genotyped).

In standard interval mapping, one deals with the missing QTL genotype data by performing maximum likelihood under a mixture model, using a version of the EM algorithm. Central to this approach is the calculation of the distribution of QTL genotypes conditional on the observed multipoint marker data. In the pseudomarker algorithm, which uses a form of multiple imputation, one must be able to simulate from the joint distribution of the genotypes at the pseudomarkers, conditional on the observed marker data.

We discuss the use of algorithms developed for hidden Markov models (HMMs) to perform the tasks mentioned above and thus deal with the missing genotype data problem. Simpler approaches can and have been used. For example, in a backcross in the absence of genotyping errors, the QTL genotype probabilities, conditional on the marker data, are a simple function of the genotypes at the nearest flanking markers. The more refined algorithms described here have several advantages. First, we may allow for the presence of genotyping errors. Second, we may more easily deal with partially informative genotypes. (For example, in an intercross, at some markers the heterozygote may not be distinguishable from one of the homozygotes.) Third, the bookkeeping tasks in implementing these algorithms can be more simple. Fourth, the algorithms can be more easily extended to more complex experimental crosses (such as the four-way cross).

In the next section, we define hidden Markov models in the context of the analysis of experimental crosses. In the following sections, we describe the basic algorithms for calculating QTL genotype probabilities, simulating from the joint distribution of QTL genotypes, estimating genetic

Address for correspondence: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205. E-mail: kbroman@jhsph.edu

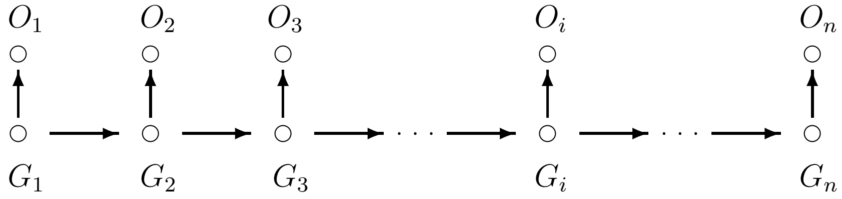

Figure 1: Illustration of a hidden Markov model. G's indicate underlying genotypes; O's indicate observed marker phenotypes.

maps, and identifying genotyping errors. We conclude the chapter with a discussion of a practical issue in the implementation of these algorithms in computer programs.

1 Specification of the model

A Markov chain is a collection of random variables, $\{G_1, G_2, \ldots, G_n\}$, satisfying the Markov property $\Pr(G_{i+1}|G_i, \ldots, G_1) = \Pr(G_{i+1}|G_i)$ for all i. In a Markov chain, for any i, the "past", $\{G_1, \ldots, G_{i-1}\}$, and the "future", $\{G_{i+1}, \ldots, G_n\}$, are conditionally independent, given the "present", G_i. We focus on Markov chains for which the random variables $\{G_i\}$ take values in a common, finite set, $\mathcal{G}$.

A hidden Markov model (HMM) consists of an unobservable underlying Markov chain, $\{G_i\}$, and a set of observable random variables, $\{O_i\}$, where each O_i depends only on G_i. In other words, for each i, O_i, given G_i, is conditionally independent of everything else, $\{O_1$, …, O_{i-1}, O_{i+1}, …, O_n, G_1, …, G_{i-1}, G_{i+1}, …, $G_n\}$. It may be useful to keep in mind the illustration in Figure 1.

The hidden states, G_i, take values in a common, finite set, $\mathcal{G}$; the observed states, O_i, take values in another finite set, $\mathcal{O}$. The joint distribution of the O_i and G_i in the HMM is parameterized by three sets of probabilities, which we will call the initiation, transition and emission probabilities. The initiation probabilities define the distribution of the initial hidden state: $\pi(g) = \Pr(G_1 = g)$ for $g \in \mathcal{G}$. The transition probabilities complete the specification for the joint distribution of the underlying, hidden Markov chain: $t_i(g, g') = \Pr(G_{i+1} = g'|G_i = g)$ for $i = 1, \ldots, n-1$ and $g, g' \in \mathcal{G}$. The emission probabilities concern the conditional distribution of the observed states given the hidden states: $e_i(g, o) = \Pr(O_i = o|G_i = g)$ for $i = 1, \ldots, n$, $g \in \mathcal{G}$, and $o \in \mathcal{O}$. We will assume here that the emission probabilities are homogeneous, with $e_i(g, o) \equiv e(g, o)$ for all i, g, o.

We now begin to consider the application of HMMs to experimental crosses. Below, we will describe the backcross and intercross specifically, but first we define the relevant HMM in some generality.

One may focus on the genotypes for a single individual at a set of loci on a single chromosome. (We will focus on an autosome.) We let G_i, $i = 1, \ldots, n$ denote the true underlying genotypes for the individual at a set of n ordered loci, and let the O_i denote the observed marker "phenotype" at locus i.

These loci may be genetic markers, or they may be "pseudomarkers," under consideration as putative QTL. The genotypes are often assumed to be phase-known genotypes, though for the intercross they need not be, as we will see below. Under the assumption of no crossover interference in meiosis, for many types of crosses, the G_i form a Markov chain. The set $\mathcal{G}$ corresponds to the possible values of these phase-known genotypes. The initiation probabilities correspond to a segregation model at a single locus; the transition probabilities are a function of the recombination fractions, r_i, between adjacent markers.

The set $\mathcal{O}$ corresponds to the set of possible observed marker phenotypes, which will include the possibility of missing values and partially informative phenotypes (such as in the case of a dominant or recessive marker). The emission probabilities involve a model for errors in genotyping, which we will assume to be common across markers, though in reality, some markers are considerably more error-prone than others. It is important to point out, further, that one conditions on the observed pattern of missing data. This will become more clear below.

1.1 The backcross

Consider a backcross individual derived from two inbred strains, A and B, where the F$_1$ parent was crossed back to the A strain. We let $\mathcal{G} = \{AA, AB\}$, the possible genotypes at a locus. The set of possible marker phenotypes is $\mathcal{O} = \{A, H, -\}$, with the last symbol corresponding to a missing value. Note our attempt to use different symbols for the underlying genotypes and the observed marker phenotypes.

The initiation probabilities, assuming Mendel's rules, are simply $\pi(AA) = \pi(AB) = 1/2$. The transition probabilities are $t_i(AA, AB) = t_i(AB, AA) = r_i$, where r_i denotes the recombination fraction between loci i and $i + 1$. Of course, $t_i(AA, AA) = t_i(AB, AB) = 1 - r_i$.

In forming the emission probabilities, we assume a constant error rate in genotyping, ϵ, so that $e(AA, A) = e(AB, H) = 1 - \epsilon$, and $e(AA, H) = e(AB, A) = \epsilon$. We condition on the observed pattern of missing data, and so $e(AA, -) = e(AB, -) = 1$. One may consider $- = \{A \text{ or } H\}$, so that $e(AA, -) = e(AA, A) + e(AA, H) = 1$.

One may consider, in forming the emission probabilities, more refined models for genotyping errors. For example, one may consider a locus-specific error rate, and one may allow the chance of a heterozygote being erroneously observed as a homozygote to be somewhat different than the converse. However, we have seen little benefit in such refinements.

1.2 The intercross

Consider a single individual in the F$_2$ generation from an intercross between two inbred strains, A and B. One may consider the hidden states, G_i, to be either phase-known genotypes (with four possible states, $\{AA, AB, BA, BB\}$) or phase-unknown genotypes (with three possible states, $\{AA, AB, BB\}$). It is an interesting and useful fact that in either case the G_i form a Markov chain (under the assumption of no crossover interference).

We will focus on the phase-unknown case, with $\mathcal{G} = \{AA, AB, BB\}$. The initiation probabilities are again those implied by Mendel's rules: $\pi(AA) = \pi(BB) = 1/4$, $\pi(AB) = 1/2$.

The transition probabilities are displayed in the Table 1, where r_i denotes the recombination fraction between markers i and $i+1$. Note that we assume that there are no sex differences in the recombination fractions.

Table 1: The transition probabilities, $t_i(g, g') = \Pr(G_{i+1} = g'|G_i = g)$, for a phase-unknown intercross.

	g'		
g	AA	AB	BB
AA	$(1-r_i)^2$	$2r_i(1-r_i)$	r_i^2
AB	$r_i(1-r_i)$	$(1-r_i)^2 + r_i^2$	$r_i(1-r_i)$
BB	r_i^2	$2r_i(1-r_i)$	$(1-r_i)^2$

As possible observed marker phenotypes, we let $\mathcal{O} = \{A, H, B, C, D, -\}$, with A, B, and H corresponding to the two homozygotes and the heterozygote, respectively, $-$ corresponding to a completely missing value, and with C and D allowing the treatment of dominant marker loci: we define C and D as in the popular computer software, MapMaker (LANDER *et al.* 1989), with $C = \{\text{not } A\} = \{B \text{ or } H\}$ and $D = \{\text{not } B\} = \{A \text{ or } H\}$.

The emission probabilities, for a simple genotyping error model, are shown in Table 2, where we let ϵ denote the genotyping error rate. Note that we again condition on the pattern of missing genotype data, and so, for example, $\Pr(O_i = C|G_i) = \Pr(O_i = B|G_i) + \Pr(O_i = H|G_i)$.

Table 2: The emission probabilities, $e(g, o) = \Pr(O_i = o|G_i = g)$, for a phase-unknown intercross.

	o					
g	A	H	B	C	D	$-$
AA	$1-\epsilon$	$\epsilon/2$	$\epsilon/2$	$\epsilon/2$	$1-\epsilon/2$	1
AB	$\epsilon/2$	$1-\epsilon$	$\epsilon/2$	$1-\epsilon/2$	$1-\epsilon/2$	1
BB	$\epsilon/2$	$\epsilon/2$	$1-\epsilon$	$1-\epsilon/2$	$\epsilon/2$	1

2 QTL genotype probabilities

Having set up the hidden Markov model for experimental crosses, we now begin our discussion of the basic algorithms used in order to deal with missing genotype data in QTL mapping. We begin with the calculation of the conditional QTL genotype probabilities given multipoint marker data, which are critical for standard interval mapping with a single QTL model. Using the notation developed in the previous section, we seek $\Pr(G_i = g|\boldsymbol{O})$, where $\boldsymbol{O} = (O_1, \ldots, O_n)$.

The brute-force approach for calculating this probability is to simply sum over all possible genotypes at the other loci.

$$\begin{aligned}
\Pr(G_i = g_i|\boldsymbol{O}) &= \sum_{g_1} \dots \sum_{g_{i-1}} \sum_{g_{i+1}} \dots \sum_{g_n} \Pr(G_1 = g_1, \dots, G_n = g_n|\boldsymbol{O}) \\
&\propto \sum_{g_1} \dots \sum_{g_{i-1}} \sum_{g_{i+1}} \dots \sum_{g_n} \pi(g_1) \prod_{j=1}^{n-1} t_j(g_j, g_{j+1}) \prod_{j=1}^{n} e(g_j, O_j)
\end{aligned}$$

For the phase-known intercross, with three possible genotypes, this is a sum with 3^{n-1} terms; clearly this is unwieldy and unnecessary. That there are simple algorithms for this calculation, which make critical use of the conditional independence structure of the HMM, is the primary motivation for the use of HMMs in experimental crosses.

The approach we follow makes use of the following two sets of probabilities.

$$\begin{aligned}
\alpha_i(g) &= \Pr(O_1, \dots, O_i, G_i = g) \\
\beta_i(g) &= \Pr(O_{i+1}, \dots, O_n|G_i = g)
\end{aligned}$$

Note that, once the α's and β's have been calculated, the probability that is the focus of this section follows directly:

$$\begin{aligned}
\Pr(G_i = g|\boldsymbol{O}) &= \Pr(G_i = g, \boldsymbol{O}) / \Pr(\boldsymbol{O}) \\
&= \alpha_i(g)\beta_i(g) / \textstyle\sum_{g'} \alpha_i(g')\beta_i(g').
\end{aligned}$$

The α's and β's are calculated inductively, using what are called the forward and backward equations, respectively. We begin with the forward equations. First, note that

$$\alpha_1(g) = \Pr(O_1, G_1 = g) = \pi(g)\ e(g, O_1).$$

Now, assume that we have calculated $\alpha_i(g)$ for each $g \in \mathcal{G}$. Then we have

$$\begin{aligned}
\alpha_{i+1}(g) &= \Pr(O_1, \dots, O_i, O_{i+1}, G_{i+1} = g) \\
&= \textstyle\sum_{g'} \Pr(O_1, \dots, O_i, O_{i+1}, G_i = g', G_{i+1} = g) \\
&= \textstyle\sum_{g'} \Pr(O_1, \dots, O_i, G_i = g') \Pr(G_{i+1} = g|G_i = g') \Pr(O_{i+1}|G_{i+1} = g) \\
&= e(g, O_{i+1}) \textstyle\sum_{g'} \alpha_i(g')\ t_i(g', g).
\end{aligned}$$

In the third line above, we made use of the conditional independence structure of the HMM, noting that

$$\Pr(G_{i+1} = g|G_i = g', O_1, \dots, O_i) = \Pr(G_{i+1} = g|G_i = g')$$

and

$$\Pr(O_{i+1}|G_{i+1} = g, G_i = g', O_1, \dots, O_i) = \Pr(O_{i+1}|G_{i+1} = g).$$

Calculation of the β's proceeds similarly, though starting at the other end of the chain. We define $\beta_n(g) = 1$ for all $g \in \mathcal{G}$. Assuming that we have calculated $\beta_i(g)$ for all g, we have

$$\begin{aligned}
\beta_{i-1}(g) &= \Pr(O_i, \dots, O_n|G_{i-1} = g) \\
&= \textstyle\sum_{g'} \Pr(O_i, \dots, O_n, G_i = g'|G_{i-1} = g) \\
&= \textstyle\sum_{g'} \Pr(O_{i+1}, \dots O_n|G_i = g') \Pr(G_i = g'|G_{i-1} = g) \Pr(O_i|G_i = g') \\
&= \textstyle\sum_{g'} \beta_i(g')\ t_{i-1}(g, g')\ e(g', O_i).
\end{aligned}$$

Again, in the third line above, we made use of the conditional independence structure of the HMM.

In summary, in order to calculate the QTL genotype probabilities, conditional on multipoint marker data, $\Pr(G_i = g|\boldsymbol{O})$, we make use of the forward and backward equations to first calculate, for each i and g, $\alpha_i(g) = \Pr(O_1, \dots, O_i, G_i = g)$ and $\beta_i(g) = \Pr(O_{i+1}, \dots, O_n|G_i = g)$. These algorithms are extremely efficient and can accommodate partially missing genotypes (such as are observed at dominant markers in an intercross) and a model for errors in genotyping.

3 Simulation of QTL genotypes

Central to the multiple imputation approach to QTL mapping is the simulation of QTL genotypes via their joint distribution conditional on the observed multipoint marker data. In this section, we describe how this is done. One considers a single chromosome and a single individual at a time. As will be seen, the simulation algorithm makes use of the β's defined in the previous section. Thus, one must first perform the backward equations described above.

The algorithm is quite simple. One first draws $g_1^\star$ from the distribution

$$\Pr(G_1 = g|\boldsymbol{O}) = \frac{\alpha_1(g)\beta_1(g)}{\sum_{g'} \alpha_1(g')\beta_1(g')}.$$

Genotypes for further loci are drawn iteratively: having drawn $g_1^\star, \dots, g_i^\star$, one draws $g_{i+1}^\star$ from the distribution

$$\begin{aligned} \Pr(G_{i+1} = g|\boldsymbol{O}, G_i = g_i^\star) &= \frac{\Pr(G_{i+1} = g, G_i = g_i^\star|\boldsymbol{O})}{\Pr(G_i = g_i^\star|\boldsymbol{O})} \\ &= \frac{\alpha_i(g_i^\star)\ t_i(g_i^\star, g)\ e(g, O_{i+1})\ \beta_{i+1}(g)}{\alpha_i(g_i^\star)\beta_i(g_i^\star)} \\ &= \frac{t_i(g_i^\star, g)\ e(g, O_{i+1})\ \beta_{i+1}(g)}{\beta_i(g_i^\star)}. \end{aligned}$$

We are again making critical use of the conditional independence structure of the HMM.

Note that the α's are not needed, except for $\alpha_1(g) = \pi(g)\ e(g, O_1)$. Thus the forward equations need not be performed. For each individual, one first uses the backward equations to calculate the β's and then simulates the chain from left to right, using the equations above. It should be no surprise that one may instead use the forward equations to calculate the α's, and then simulate the chain from right to left, using formulas analogous to those above.

4 Joint QTL genotype probabilities

In multiple interval mapping (MIM) with multiple linked QTL, it is important to calculate joint QTL genotype probabilities, conditional on the observed multipoint marker data.

We begin by describing the calculation of $\Pr(G_i = g, G_j = g'|\boldsymbol{O})$ for all i, j with $i < j$. As will be seen, one must first calculate the α's and β's defined above. One may start by calculating the case $j = i + 1$ for each $i = 1, \dots, n-1$, as follows.

$$\begin{aligned}
\Pr(G_i = g, G_{i+1} = g'|\boldsymbol{O}) &\propto \Pr(G_i = g, G_{i+1} = g', \boldsymbol{O}) \\
&= \Pr(O_1, \dots, O_i, G_i = g) \Pr(G_{i+1} = g'|G_i = g) \\
&\qquad \times \Pr(O_{i+1}|G_{i+1} = g') \Pr(O_{i+2}, \dots, O_n|G_{i+1} = g') \\
&= \alpha_i(g)\; t_i(g, g')\; e(g', O_{i+1})\; \beta_{i+1}(g')
\end{aligned}$$

One uses the final line above and rescales the results so that they sum to 1.

The rest of the pairwise probabilities follow with the standard technique, using induction.

$$\begin{aligned}
\Pr(G_i = g, G_j = g''|\boldsymbol{O}) &= \sum_{g''} \Pr(G_i = g, G_{j-1} = g'', G_j = g'|\boldsymbol{O}) \\
&= \sum_{g''} \Pr(G_i = g, G_{j-1} = g''|\boldsymbol{O}) \Pr(G_j = g'|G_{j-1} = g'', \boldsymbol{O})
\end{aligned}$$

Finally, one may wish to calculate the joint probabilities for multiple linked loci, conditional on the observed multipoint marker data. Again, the conditional independence structure of the HMM makes this a simple task: the joint distribution may be calculated based on pairwise probabilities whose calculation was described above. Consider $i_1 < i_2 < \dots < i_k$, with each $i_j \in \{1, \dots, n\}$; we have

$$\Pr(G_{i_1} = g_1, \dots, G_{i_k} = g_k|\boldsymbol{O}) = \\ \Pr(G_{i_1} = g_1, G_{i_2} = g_2|\boldsymbol{O}) \prod_{j=2}^{k-1} \Pr(G_{i_{j+1}} = g_{j+1}|G_{i_j} = g_j, \boldsymbol{O}).$$

The equations in this section do get a little bit complicated, but they are all formed of quite simple pieces. The central calculation is the use of the forward and backward equations to obtain the α's and β's.

5 The Viterbi algorithm

In some cases, it is useful to impute the underlying genotype data, calculating $\hat{\boldsymbol{G}} = \arg\max_{\boldsymbol{G}} \Pr(\boldsymbol{G}|\boldsymbol{O})$. The Viterbi algorithm solves this problem via dynamic programming.

First, define

$$\gamma_k(g) = \max_{g_1, \cdots, g_{k-1}} \Pr(G_1 = g_1, \dots G_{k-1} = g_{k-1}, G_k = g_k, O_1, \dots, O_k).$$

These are calculated inductively, by an approach similar to that used in the forward equations (Sec. 2). Let $\gamma_1(g) = \Pr(G_1 = g, O_1) = \pi(g)e(g, O_1)$. Given $\gamma_k(g)$ for all k and g, we have

$$\gamma_{k+1}(g) = e(g, O_{k+1}) \max_{g'} t_k(g', g)\gamma_k(g').$$

At the same time, we keep track of the values at which the maxima occurred: define $\delta_{k+1}(g) = \arg\max_{g'} t_k(g', g)\gamma_k(g')$. If the maximum is not unique, we can keep track of each of them or pick a random one. (We do the latter in R/qtl.)

To obtain the most probable sequence of underlying genotypes, we then take $\hat{G}_n = \arg\max_g \gamma_n(g)$ and, working backwards, $\hat{G}_{k-1} = \delta_k(\hat{G}_k)$.

The inferred genotypes obtained by the Viterbi algorithm should be used with great caution. If one treated the inferred genotypes as if they were the true values, an important source of uncertainty would be ignored.

Moreover, if inter-marker positions are included and genotyping error is allowed, the results of the Viterbi algorithm can vary according to the density of inter-marker positions that are used. The Viterbi algorithm identifies the most likely sequence of genotypes, but this sequence may have quite low probability and may exhibit features which are themselves unlikely.

For example, consider three markers at a 10 cM spacing and a single backcross individual with observed marker genotypes *AA–AB–AB* at the three markers. If the Viterbi algorithm is applied with a genotyping error rate of 1%, and using just the three marker positions, the most likely sequence of underlying genotypes matches those observed. If, however, one considers positions at 1 cM steps across the region, the most likely sequence of underlying genotypes is such that the individual is heterozygous across the entire region. While it is probable that the individual is recombinant across the first interval and that the observed genotype at the first marker is not in error, if many inter-marker positions are considered, this event is split across multiple sequences of genotypes (each corresponding to a different position for the recombination event), and so the sequence in which the initial genotype is in error and there is no recombination event ends up being most likely.

This issue leads us to recommend the use of simulation to impute genotypes (as described in Sec. 3), rather than using the Viterbi algorithm to calculate the most probable sequence of underlying genotypes.

6 Estimation of inter-marker distances

The calculations described above depend crucially on the order of the genetic markers and the recombination fractions between adjacent markers (i.e., the inter-marker distances). In this section, we describe the derivation of joint maximum likelihood estimates (MLEs) of the recombination fractions between genetic markers, assuming that the order of the genetic markers is known. We omit from consideration the more difficult problem of determining marker order.

Taking the order of the genetic markers as fixed and known, the probability of the observed marker data for an individual, $\Pr(\boldsymbol{O})$, still depends on the recombination fractions between adjacent markers. For the sake of simplicity, this dependence has been neglected in our notation heretofore. Moreover, we have been considering a single individual at a time. In our discussion of the estimation of inter-marker distances, however, it will be important to make this dependence clear. Let $\boldsymbol{r} = (r_1, \ldots, r_{n-1}$ denote the set of recombination fractions, and let $\boldsymbol{O}_k$ denote the observed marker data for individual k, for $k = 1, \ldots, N$.

We seek the MLE of $\boldsymbol{r}$, defined to be the value of $\boldsymbol{r}$ for which the likelihood is maximized, $\hat{\boldsymbol{r}} = \arg\max L(r)$, where $L(r) = \prod_{k=1}^{N} \Pr(\boldsymbol{O}_k|\boldsymbol{r})$. These estimates are obtained using a version of the EM algorithm (DEMPSTER *et al.* 1977).

We begin with initial estimates of the recombination fractions, $\hat{\boldsymbol{r}}^{(0)}$. The EM algorithm is an iterative algorithm: the estimated recombination fractions are successively improved, increasing the likelihood at each stage, until convergence. In each iteration, the updated estimates of the recombination fractions are the expected proportions of recombination events, across the N individuals, in each marker interval, given the current estimates of the recombination fractions.

At each iteration, we first perform the forward and backward equations for each individual, using the current estimates of the recombination fractions, $\hat{\boldsymbol{r}}^{(s)}$. We then calculate, for each interval i, $\gamma_{ki}(g, g'|\hat{\boldsymbol{r}}^{(s)}) = \Pr(G_{k,i} = g, G_{k,i+1} = g'|\boldsymbol{O}, \hat{\boldsymbol{r}}^{(s)})$. This is the probability that individual k has genotypes g and g' at markers i and $i + 1$, given its multipoint marker data, and given the current estimates of the recombination fractions. The calculation of the γ's, based on the α's and β's for the corresponding individual, appears in Sec. 4.

The updated estimate of the recombination fraction for interval i is then $\hat{r}_i^{(s+1)} = \sum_k \sum_{g,g'} \gamma_{ki}(g, g'|\hat{\boldsymbol{r}}^{(s)}) \;\; p(g, g')/N$, where $p(g, g')$ is the proportion of recombination events across the interval (i.e., 0, 1/2, or 1) if the individual has genotypes g and g' at the markers defining the interval. Note that, in estimating the inter-marker distances for an intercross, we use the phase-known (4-state) version of the HMM, so that the function $p(g, g')$ is well defined.

7 Detection of genotyping errors

Successful QTL mapping requires high quality phenotype and genotype data. In this section, we describe an approach for identifying errors in the genotype data. For each marker and each individual, we calculate a LOD score, with large LOD scores indicating likely errors.

The presence of partially informative genotypes (e.g., at dominant markers in an intercross) makes this slightly tricky. Let us assume that the observed marker phenotypes, $o \in \mathcal{O}$ are subsets of the possible underlying marker genotypes, $\mathcal{G}$. For example, in the case of an intercross, where $\mathcal{G} = \{AA, AB, BB\}$, the set of possible marker phenotypes is $\mathcal{O} = \{A, H, B, C, D, -\}$, with, for example, $A = \{AA\}$ and $C = \{AB, BB\}$.

Let G_{ki} denote the true underlying genotype for individual k at marker i, and let O_{ki} denote the corresponding marker phenotype. We assume the simple model for genotyping errors that was described in Sec. 1, and we assume the genotyping error rate, ϵ, is known. We seek to calculate

$$\begin{aligned} \text{LOD}_{ki} &= \log_{10}\left\{\frac{\Pr(\boldsymbol{O}|G_{ki} \notin O_{ki}, \epsilon)}{\Pr(\boldsymbol{O}|G_{ki} \in O_{ki}, \epsilon)}\right\} \\ &= \log_{10}\left\{\frac{\Pr(G_{ki} \notin O_{ki}|\boldsymbol{O}, \epsilon)}{\Pr(G_{ki} \in O_{ki}|\boldsymbol{O}, \epsilon)} \times \frac{1-\epsilon}{\epsilon}\right\} \end{aligned}$$

Note that the calculation of the probabilities in the above formula was described in Sec. 2.

These LOD scores depend on the specified genotyping error rate, ϵ, but typical values, in the range 0.001 – 0.02, do give similar results. Genotyping error LOD scores below 3 or 4 are generally benign. Only when the LOD scores exceed 4 should they be given much consideration. It should be noted genotyping errors can only be detected in the case of quite dense markers. At the same time, however, genotyping errors have little effect on the result of QTL mapping if the markers are not dense. Finally, if a particular marker gives many large error LOD scores, it may be that a problem with marker order is the cause (though, of course, the marker may also have a greater than typical frequency of errors.)

8 A practical issue

In the case of many genetic markers (or pseudomarkers), the direct calculation of α and β, as described above, will result in underflow: $\alpha_n(v) = \Pr(O_1, \ldots, O_n, G_n = v)$ can be extremely small. One method to deal with this is to calculate $\alpha' = \log \alpha$ and $\beta' = \log \beta$. In the forward equations, we must obtain $\alpha'_{i+1}(g) = \log e(g, O_{i+1}) + \log\{\sum_{g'} \alpha_i(g') t_i(g', g)\}$. This leads to the problem of calculating $\log(f_1 + f_2)$ on the basis of $g_i = \log f_i$, which may be facilitated by the following trick:

$$\begin{aligned} \log(f_1 + f_2) &= \log(e^{g_1} + e^{g_2}) \\ &= \log\{e^{g_1}(1 + e^{g_2 - g_1})\} \\ &= g_1 + \log(1 + e^{g_2 - g_1}) \end{aligned}$$

A problem occurs when $g_2 \gg g_1$: the above formula will result in an overflow. In such a case one simply notes that $\log(f_1 + f_2) \approx g_2$.

9 Further reading

BAUM *et al.* (1970) were the first to describe estimation for hidden Markov models, and derived the forward and backward equations. For other expositions of the use of HMMs, see RABINER (1989) or LANGE (1999).

CHURCHILL (1989) was the first to use HMMs explicitly in biology. HMMs have been used for a variety of biological applications, including the analysis of patch-clamp recordings for the study of ion channels, multiple sequence alignment, and protein structure prediction.

LANDER and GREEN (1987) described the multipoint estimation of genetic maps; their method was implemented for experimental crosses in the software MapMaker (LANDER *et al.* 1987). JIANG and ZENG (1997) described an alternative approach for dealing with missing and partially missing genotype data. LINCOLN and LANDER (1992) developed the LOD scores, defined above, for identifying genotyping errors in experimental crosses.

10 References

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat **41**: 164–171.

Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bulletin of Mathematical Biology **51**: 79–94.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B **39**: 1–38.

Jiang C, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101**: 47–58.

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA **84**: 2363–2367.

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1**: 174–181.

Lange K (1999) *Numerical analysis for statisticians*. Springer, New York, Sec 23.3.

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics **14**: 604–610.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**: 257–286.