QTL Mapping II:
Hidden Markov model technology
and
The pseudomarker algorithm

Karl W Broman
Department of Biostatistics
Johns Hopkins University

kbroman@jhsph.edu

www.biostat.jhsph.edu/~kbroman

HMM technology: Outline

• The problems
• A simple solution
• Why a complex solution?
• The hidden Markov model
• Backcross, intercross
• QTL genotype probabilities
• Simulation of QTL genotypes
The problems

- Calculate genotype probability at an arbitrary location, conditional on multipoint marker data.
- Simulate from the joint genotype distribution on a grid, given multipoint marker data.

<table>
<thead>
<tr>
<th>Even grid</th>
<th>Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A A A - H H B</td>
</tr>
<tr>
<td></td>
<td>B B H H H - H</td>
</tr>
<tr>
<td></td>
<td>H H H A H H H</td>
</tr>
</tbody>
</table>

A simple solution

- Under the no interference (NI) model, the genotypes follow a Markov chain.
- Thus, the genotype probability depends only on the nearest flanking typed markers

<table>
<thead>
<tr>
<th>Even grid</th>
<th>Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A A A - H H B</td>
</tr>
<tr>
<td></td>
<td>B B H H H - H</td>
</tr>
<tr>
<td></td>
<td>H H H A H H H</td>
</tr>
</tbody>
</table>
Why a complex solution?

- Allow for the presence of genotyping errors
- Simply deal with partially informative genotypes (e.g., C = H or B)
- Simplify bookkeeping tasks in the implementation
- Easily extend algorithms to more complex experimental crosses (such as the four-way cross)

The hidden Markov model

- The \( \{G_i\} \) (hidden states) form a Markov chain, with values in some finite set, \( S \).
  \[
  \Pr(G_{i+1} \mid G_i, \ldots, G_1) = \Pr(G_{i+1} \mid G_i)
  \]
- The observable random variables, \( \{O_i\} \), take values in another finite set, \( O \).
  \( O_i \) depends only on \( G_i \)
- \( G_i = \) “true” genotype at marker \( i \)
- \( O_i = \) “observed genotype” (marker phenotype) at \( i \)
Model parameters

- **Initiation probabilities:** \( \pi(g) = \Pr(G_1 = g) \)
  for \( g \in \mathcal{G} \)

- **Transition probabilities:** \( t_i(g, g') = \Pr(G_{i+1} = g' \mid G_i = g) \)
  for \( i = 1, \ldots, n - 1 \) and \( g, g' \in \mathcal{G} \)

- **Emission probabilities:** \( e_i(g, o) = \Pr(O_i = o \mid G_i = g) \)
  for \( i = 1, \ldots, n, g \in \mathcal{G}, \text{and } o \in \mathcal{O} \)

(We assume \( e_i(g, o) \equiv e(g, o) \) for all \( i \).)

Joint probability

\[
\Pr(G = g, O = o) = \Pr(G_1 = g_1, \ldots, G_n = g_n, O_1 = o_1, \ldots, O_n = o_n) \\
= \Pr(G_1 = g_1) \Pr(G_2 = g_2 \mid G_1 = g_1) \cdots \\
\cdots \Pr(G_n = g_n \mid G_{n-1} = g_{n-1}) \cdot \Pr(O_1 = o_1 \mid G_1 = g_1) \cdots \\
\cdots \Pr(O_n = o_n \mid G_n = g_n) \\
= \pi(g_1) \prod_{i=1}^{n-1} t_i(g_i, g_{i+1}) \prod_{i=1}^n e(g_i, o_i)
\]
The backcross

\[ S = \{AA, AB\} \quad \emptyset = \{A, H, -\} \quad (\_\_ = \text{missing}) \]

Initiation probabilities:
\[ \pi(AA) = \pi(AB) = \frac{1}{2} \]

Transition probabilities:
\[ r_i = \text{recombination fraction for interval } i. \]
\[ t_i(AA, AB) = t_i(AB, AA) = r_i \]
\[ t_i(AA, AA) = t_i(AB, AB) = 1 - r_i \]

Emission probabilities:
\[ \epsilon = \text{genotyping error rate} \]
\[ e(AA, A) = e(AB, H) = 1 - \epsilon, \quad e(AA, -) = e(AB, -) = 1 \]
\[ e(AA, H) = e(AB, A) = \epsilon \]

The intercross

We’ll consider phase-unknown genotypes.

\[ S = \{AA, AB, BB\} \]

Initiation probabilities:
\[ \pi(AA) = \pi(BB) = \frac{1}{4}, \quad \pi(AB) = \frac{1}{2} \]

Transition probabilities, \( t_i(g, g') = \Pr(G_{i+1} = g' \mid G_i = g) \):

<table>
<thead>
<tr>
<th>( g )</th>
<th>( AA )</th>
<th>( AB )</th>
<th>( BB )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>((1 - r_i)^2)</td>
<td>(2r_i(1 - r_i))</td>
<td>(r_i^2)</td>
</tr>
<tr>
<td>AB</td>
<td>(r_i(1 - r_i))</td>
<td>((1 - r_i)^2 + r_i^2)</td>
<td>(r_i(1 - r_i))</td>
</tr>
<tr>
<td>BB</td>
<td>(r_i^2)</td>
<td>(2r_i(1 - r_i))</td>
<td>((1 - r_i)^2)</td>
</tr>
</tbody>
</table>
The intercross (cont.)

\[ \mathcal{O} = \{A, H, B, C, D, -\} \]
\[ - = \text{missing} = \{A \text{ or } H \text{ or } B\} \]
\[ C = \text{not} \ A = \{H \text{ or } B\} \]
\[ D = \text{not} \ B = \{A \text{ or } H\} \]

Emission probabilities, \(e(g, o) = \Pr(O_i = o \mid G_i = g)\):

<table>
<thead>
<tr>
<th>(g)</th>
<th>(o)</th>
<th>(A)</th>
<th>(H)</th>
<th>(B)</th>
<th>(C)</th>
<th>(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(AA)</td>
<td>(1 - \epsilon)</td>
<td>(\epsilon/2)</td>
<td>(\epsilon/2)</td>
<td>(\epsilon)</td>
<td>(1 - \epsilon/2)</td>
<td>1</td>
</tr>
<tr>
<td>(AB)</td>
<td>(\epsilon/2)</td>
<td>(1 - \epsilon)</td>
<td>(\epsilon/2)</td>
<td>(1 - \epsilon/2)</td>
<td>(1 - \epsilon/2)</td>
<td>1</td>
</tr>
<tr>
<td>(BB)</td>
<td>(\epsilon/2)</td>
<td>(\epsilon/2)</td>
<td>(1 - \epsilon)</td>
<td>(1 - \epsilon/2)</td>
<td>(\epsilon)</td>
<td>1</td>
</tr>
</tbody>
</table>

QTL genotype probabilities

We seek to calculate \(\Pr(G_i = g \mid O)\). where \(O = (O_1, O_2, \ldots, O_n)\) is the observed multipoint marker data.

Brute force:

\[
\Pr(G_i = g_i | O) = \sum_{g_1} \cdots \sum_{g_{i-1}} \sum_{g_i} \cdots \sum_{g_n} \Pr(G_1 = g_1, \ldots, G_n = g_n | O)
\]

\[
\propto \sum_{g_1} \cdots \sum_{g_{i-1}} \sum_{g_i} \cdots \sum_{g_n} \pi(g_1) \prod_{j=1}^{n-1} t_j(g_j, g_{j+1}) \prod_{j=1}^{n} e(g_j, O_j)
\]

For the phase-unknown intercross, this is a sum with \(3^{n-1}\) terms; clearly this is unwieldy and unnecessary. But, of course, there is a simpler way!
The forward and backward equations

\[ O_1 \quad O_2 \quad O_3 \quad O_i \quad O_n \]
\[ \circ \quad \circ \quad \circ \quad \circ \quad \circ \]
\[ G_1 \quad G_2 \quad G_3 \quad G_i \quad G_n \]

Our approach makes use of the following two sets of probabilities:

\[ \alpha_i(g) = \Pr(O_1, \ldots, O_i, G_i = g) \]
\[ \beta_i(g) = \Pr(O_{i+1}, \ldots, O_n | G_i = g) \]

Note that once the \( \alpha \)'s and \( \beta \)'s have been calculated, the probability that is our focus follows directly:

\[
\Pr(G_i = g | O) = \frac{\Pr(G_i = g, O)}{\Pr(O)} = \frac{\alpha_i(g)\beta_i(g)}{\sum_{g'} \alpha_i(g')\beta_i(g')}
\]

The forward equations

The \( \alpha \)'s are calculated inductively.

First, note that

\[ \alpha_1(g) = \Pr(O_1, G_1 = g) = \pi(g) \, e(g, O_1) \]

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

\[ \alpha_{i+1}(g) = \Pr(O_1, \ldots, O_i, O_{i+1}, G_{i+1} = g) \]
\[ = \sum_{g'} \Pr(O_1, \ldots, O_i, O_{i+1}, G_i = g', G_{i+1} = g) \]
\[ = \sum_{g'} \Pr(O_1, \ldots, O_i, G_i = g') \Pr(G_{i+1} = g | G_i = g') \Pr(O_{i+1} | G_{i+1} = g) \]
\[ = e(g, O_{i+1}) \sum_{g'} \alpha_i(g') \, t_i(g', g) \]
The backward equations

The $\beta$’s are calculated similarly, but moving backward.

First, we define $\beta_n(g) \equiv 1$ for all $g \in \mathcal{G}$.

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

$$
\beta_{i-1}(g) = \Pr(O_i, \ldots, O_n | G_{i-1} = g)
$$

$$
= \sum_{g'} \Pr(O_i, \ldots, O_n, G_i = g' | G_{i-1} = g)
$$

$$
= \sum_{g'} \Pr(O_{i+1}, \ldots, O_n | G_i = g') \Pr(G_i = g' | G_{i-1} = g) \Pr(O_i | G_i = g')
$$

$$
= \sum_{g'} \beta_i(g') t_{i-1}(g, g') e(g', O_i)
$$

QTL genotype probabilities

1. Calculate the $\alpha$’s and $\beta$’s, simultaneously, via the forward and backward equations.

2. Calculate, for each $i$ and $g$,

$$
\Pr(G_i = g | O) = \frac{\Pr(G_i = g, O)}{\Pr(O)}
$$

$$
= \frac{\alpha_i(g) \beta_i(g)}{\sum_{g'} \alpha_i(g') \beta_i(g')}
$$
Simulation of QTL genotypes

We seek to simulate from the joint distribution, \( \Pr(G_1, \ldots, G_n \mid O) \)

[Why? We’ll explain shortly.]

First draw \( g_1^* \) from the distribution

\[
\Pr(G_1 = g \mid O) = \frac{\alpha_1(g)\beta_1(g)}{\sum_{g'}\alpha_1(g')\beta_1(g')}
\]

Genotypes for further loci are drawn iteratively:

having drawn \( g_1^*, \ldots, g_i^* \), draw \( g_{i+1}^* \) from

\[
\Pr(G_{i+1} = g \mid O, G_i = g_i^*) = \frac{\Pr(G_{i+1} = g, G_i = g_i^* \mid O)}{\Pr(G_i = g_i^* \mid O)}
= \frac{\alpha_i(g_i^*)t_i(g_i^*, g)\epsilon(g, O_{i+1})\beta_{i+1}(g)}{\alpha_i(g_i^*)\beta_i(g_i^*)}
= t_i(g_i^*, g)\epsilon(g, O_{i+1})\beta_{i+1}(g)/\beta_i(g_i^*)
\]

Note that we need to first calculate the \( \beta \)'s (via the backward equations).

A practical issue

In the case of many genetic markers (or pseudomarkers), the direct calculation of \( \alpha \) and \( \beta \), as described above, will result in underflow.

\( \alpha_n(g) = \Pr(O_1, O_2, \ldots, O_n, G_n = g) \) can be extremely small!

One method to deal with this is to work with \( \alpha' = \log \alpha \) and \( \beta' = \log \beta \).

But in the forward equations, we need

\[
\alpha'_{i+1}(g) = \log \epsilon(g, O_{i+1}) + \log\{\sum_{g'}\alpha_i(g')t_i(g', g)\}
\]

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 with the following trick:

\[
\log(f_1 + f_2) = \log(e^{g_1} + e^{g_2})
= \log\{e^{g_1}(1 + e^{g_2-g_1})\}
= g_1 + \log(1 + e^{g_2-g_1})
\]

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 \( \log(f_1 + f_2) \approx g_2 \).
The pseudomarker algorithm: Outline


- Data structure and notation
- Basic idea
- Advantages and cautions
- An example

Data structure and notation

\[ y = \text{phenotypes} \]
\[ m = \text{observed marker genotypes} \]
\[ q = \text{unobserved QTL genotypes} \]
\[ \mu = \text{model parameters} \]
\[ \gamma = \text{QTL locations} \]
\[ H = \text{QTL model} \]
The factorization

\[
\Pr(y, m, q, \mu, \gamma) = \{\Pr(y \mid q, \mu) \Pr(\mu)\} \{\Pr(q \mid m, \gamma) \Pr(m) \Pr(\gamma)\}
\]

- \(\Pr(y \mid q, \mu) \Pr(\mu)\) = genetic model part
- \(\Pr(q \mid m, \gamma) \Pr(m) \Pr(\gamma)\) = linkage part

The unobserved QTL genotypes play a central role.

If the QTL genotypes were known, the problem reduces to

model selection in regression

The basic idea

- **Simulate** multiple realizations of the joint genotypes on a uniform grid, conditional on the observed multipoint marker data.

- **Fit a QTL model** with each realization, one at a time.

- **Combine the realizations** to get an estimate of the posterior probability of the QTL model.
Advantages

- Simple computation (just regression)
- Handle missing genotype data
- Covariates
- Any phenotype distribution
- Multi-dimensional genome scans
- Linked QTL; interacting QTL
- Modular algorithm
- No MCMC worries

Cautions

- Monte Carlo error (number of imputations)
- Numerical integration error (density of pseudomarker grid)
- Model selection (as usual)
- Relatively slow, for one- or two-dimensional genome scans
An example

Sugiyama et al. (2001) Genomics 71:70–77

Salt-induced hypertension in the mouse.

Backcross with 250 individuals.

174 markers (for most, only genotyped the extremes).

Phenotype distribution

Blood pressure
All chromosomes

Chromosome 1
Chromosome 4

Map position (cM)
## Drop-one-term table

<table>
<thead>
<tr>
<th>Term</th>
<th>df</th>
<th>LOD</th>
<th>% variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1@37</td>
<td>1</td>
<td>1.9</td>
<td>2.3</td>
</tr>
<tr>
<td>c1@80</td>
<td>1</td>
<td>3.1</td>
<td>3.8</td>
</tr>
<tr>
<td>c4@30</td>
<td>1</td>
<td>9.5</td>
<td>12.3</td>
</tr>
<tr>
<td>c6@60</td>
<td>2</td>
<td>5.7</td>
<td>7.1</td>
</tr>
<tr>
<td>c7@54</td>
<td>2</td>
<td>2.0</td>
<td>2.4</td>
</tr>
<tr>
<td>c15@18</td>
<td>3</td>
<td>7.6</td>
<td>9.6</td>
</tr>
<tr>
<td>c6@60 : c15@18</td>
<td>1</td>
<td>3.8</td>
<td>4.6</td>
</tr>
<tr>
<td>c7@54 : c15@18</td>
<td>1</td>
<td>1.7</td>
<td>2.1</td>
</tr>
</tbody>
</table>
  The first paper on hidden Markov models.

  A quite readable review of HMMs.

  Review of HMMs.

  The first application of HMMs in biology.

  First use of HMMs for genetic mapping.

  Paper describing how to deal with genotyping errors in experimental crosses.

• 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
  An alternative approach for dealing with missing and partially missing genotype data.

  The paper on the imputation method (the “pseudomarker algorithm”).

  The salt-induced hypertension example.