Chapter 4: Multiple Marker Loci

4.1 Markov dependence of inheritance vectors 4-1
4.2 Estimating genetic maps 4-6
4.3 HMM computations across chromosomes 4-10
4.4 The probabilities of latent \textit{ibd} given $Y$ 4-14
4.5 Computations for big pedigrees 4-20
4.6 Big pedigrees with many markers 4-26

4.1.1 MULTILOCUS INHERITANCE SPECIFICATION:

\[
\begin{pmatrix}
\rho_1 & \rho_2 \\
S_{i,1} & S_{i,2} & S_{i,3} & \cdots & S_{i,j-1} & S_{i,j} & S_{i,j+1} & \cdots & S_{i,\ell} \\
T_{i,1} & T_{i,2} & \cdots & T_{i,j-1} & T_{i,j}
\end{pmatrix}
\]

- Assume that $\ell$ loci are ordered $1, \ldots, \ell$ along the chromosome. Let the intervals between successive loci be $I_1, \ldots, I_{\ell-1}$.
- $S_{i,j} = 0$ or $1$ specifies inheritance at locus $j$ in meosis $i$.
- $\rho_j$ is probability of recombination between locus $j$ and locus $j + 1$.
- $S_{i,*} = \{S_{i,j}, i = 1, \ldots, m\}$ is the inheritance vector at locus $j$.
- $S_{i,*} = \{S_{i,j}, j = 1, \ldots, \ell\}$ is vector specifying meiosis or gamete $i$.
- Let $T_{i,j} = 1$ if a gamete $i$ is recombinant on interval $I_j$, and $T_{i,j} = 0$ otherwise ($j = 1, \ldots, \ell - 1$). Then, in meiosis $i$,

\[
\begin{align*}
T_{i,j} & = 1 \text{ if } S_{i,j} \neq S_{i,j+1}, \text{ and} \\
T_{i,j} & = 0 \text{ if } S_{i,j} = S_{i,j+1}, \quad j = 1, \ldots, \ell - 1. \\
\Pr(T_{i,j} = 1) & = \Pr(S_{i,j} \neq S_{i,j+1}) = \rho_j.
\end{align*}
\]
4.1.2 MULTILOCUS INHERITANCE; NO INTERFERENCE:

- A model for \( S_{i,*} = \{ S_{i,j}, j = 1, ..., \ell \} \) is equivalent to a model for \( (T_{i,1}, ..., T_{i,\ell-1}) \); for example, some genetic interference model.
- The simplest models for meiosis assume no interference: that is, that the \( T_{i,j} \) are independent, for all \( i \) and \( j \).
- Then the \( S_{i,j} \) are first-order Markov over loci \( j \), with meioses \( i \) being independent.
- One way to express this is that
  \[
  \Pr(S_{i,j} | S_{i,1}, ..., S_{i,j-1}) = \Pr(S_{i,j} | S_{i,j-1})
  \]
  so that
  \[
  \Pr(S_{i,*}) = \Pr(S_{i,1}) \prod_{j=2}^{\ell} \Pr(S_{i,j} | S_{i,j-1})
  \]
- Combining the meioses
  \[
  \Pr(S) = \Pr(S_{*,1}) \prod_{j=2}^{\ell} \Pr(S_{*,j} | S_{*,j-1})
  \]
where \( S = \{ S_{i,j}; i = 1, ..., m, j = 1, ..., \ell \} \).

4.1.3 CONDITIONAL INDEPENDENCE OF \( S \):

- The Markov dependence may also be expressed as:
  Given \( S_{i,j}, S_{i,j-1} \) is independent of \( S_{i,j+1} \).
- Another useful way is to consider the probability of any given indicator \( S_{i,j} \) conditional on all the others, \( S_{-(i,j)} = \{ S_{k,l}; (k, l) \neq (i, j) \} \).
- Then \( S_{i,j} \) depends only on the indicators for the same meiosis and the two neighboring loci. For \( s = 0, 1 \),
  \[
  \Pr(S_{i,j} = s | S_{-(i,j)}) = \Pr(S_{i,j} = s | S_{i,j+1}, S_{i,j-1}) \propto \rho_j^{|s-S_{i,j-1}|} (1-\rho_{j-1})^{1-|s-S_{i,j-1}|} \rho_j^{|s-S_{i,j+1}|} (1-\rho_j)^{1-|s-S_{i,j+1}|}
  \]
where \( \rho_j = \Pr(S_{i,j} \neq S_{i,j+1}) \) is the recombination frequency in \( I_j \).
- Note that the equation just indicates the recombination/non-recombination events in intervals \( I_{j-1} \) and \( I_j \), implied by the three indicators \( (S_{i,j-1}, S_{i,j} = s, S_{i,j+1}) \).
4.1.4 THE LOCUS \( j \) DATA PROBABILITIES:

Recall in slides 2.5.1 to 2.5.5, we computed the single-locus computation of observed data on a set of individuals, in terms either of \( ibd \) states \( J \), or using the inheritance \( S \).

\[
\Pr(Y) = \sum_s \Pr(Y \mid S) \Pr(S) = \sum_s \Pr(Y \mid J(S)) \Pr(S) = \sum_J \Pr(Y \mid J) \Pr(J).
\]

- In examples we used the \( ibd \) states, because there are fewer \( ibd \) patterns \( J \) than values of \( S \). For example, just \((k_0, k_1, k_2)\) for two non-inbred individuals, regardless of what pedigree gave rise to them.
- However, although the component \( S_{i,j} \) are Markov over loci \( j \), gene \( ibd \) patterns are not. Different values of \( S_{*,j} \) may give rise to the same \( ibd \) pattern. Grouping the states of a Markov chain does not, in general, produce a Markov chain. So to use the Markov dependence, we have to use \( S \).
- Now let \( Y_{*,j} \) denote all the data corresponding to locus \( j \).

4.1.5 THE HMM ACROSS LOCI FOR PEDIGREE DATA:

\[
\begin{align*}
\Pr(S) &= \Pr(S_{*,1}) \prod_{j=2}^{l} \Pr(S_{*,j} \mid S_{*,j-1}) \\
\Pr(Y \mid S) &= \prod_{j=1}^{\ell} \Pr(Y_{*,j} \mid S_{*,j}).
\end{align*}
\]

- As before \( S_{*,j} \) determines the \( ibd \) at locus \( j \), and hence \( \Pr(Y_{*,j} \mid S_{*,j}) \).
- Note that, given \( S_{*,j} \), \( Y^{*(j-1)}, Y_{*,j} \), and \( Y^{*(j+1)} \) are mutually independent.

Also, given \( S_{*,j} \), \( Y^{*(j-1)}, Y_{*,j} \), and \( S_{*,j+1} \) are independent.
Also, given \( S_{*,j} \), \( Y^{*(j+1)}, Y_{*,j} \), and \( S_{*,j-1} \) are independent.
### 4.2.1 Counting recombinants if $S$ is observed:

- If $S$ is observed, we can count recombinants. Let $X_{m,j} = \sum_{i \text{ male}} |S_{i,j+1} - S_{i,j}|$ be the number of recombinations in interval $I_j$ in male meioses, and $M_m$ is the total number of male meioses scored in the pedigree. Similarly for female meioses.

- $Y$ is irrelevant to $\rho$-estimation, and the log-likelihood is

$$
\log P_{\ell}(S) = \log(P_{\ell}(S_{1,1})) + \sum_{j=1}^{\ell-1} \log(P_{\ell}(S_{j+1,j} \mid S_{j,j}))
$$

- Recombination parameters $\rho_{m,j}$ and $\rho_{f,j}$ enter only in

$$
\log(P_{\ell}(S_{j+1,j} \mid S_{j,j})) = X_{m,j} \log(\rho_{m,j}) + (M_m - X_{m,j}) \log(1 - \rho_{m,j})
+ X_{f,j} \log(\rho_{f,j}) + (M_f - X_{f,j}) \log(1 - \rho_{f,j})
$$

- $\widehat{\rho}_{m,j} = X_{m,j}/M_m$, and $\widehat{\rho}_{f,j} = X_{f,j}/M_f$.

### 4.2.2 $S$ unobserved: An EM algorithm for genetic maps:

- $\rho_{m,j}$ and $\rho_{f,j}$ occur only in the term $\log(P_{\ell}(S_{j+1,j} \mid S_{j,j}))$ of the complete-data log-likelihood

$$
\log(P_{\ell}(S_{1,1})) + \sum_{j=1}^{\ell-1} \log(P_{\ell}(S_{j+1,j} \mid S_{j,j})) + \sum_{j=1}^{\ell} \log(P_{\ell}(Y_{j,j} \mid S_{j,j}))
$$

- E-step: The expected complete-data log-likelihood requires only computation of $E(\log(P_{\ell}(S_{j+1,j} \mid S_{j,j})) \mid Y)$ or

$$
\bar{X}_{m,j} = E(X_{m,j} \mid Y) = \sum_{i \text{ male}} E(|S_{i,j+1} - S_{i,j}| \mid Y)
$$

and similarly $\bar{X}_{f,j}$.

- M-step: The new estimate of $\rho_{m,j}$ is $\bar{X}_{m,j}/M_m$, and similarly for all intervals $j = 1, 2, 3, \ldots, \ell - 1$ and for both the male and female meioses.

- The EM algorithm is thus readily implemented to provide estimates of recombination frequencies for all intervals and for both sexes, provided E-step can be done. (See 4.4.2 for how we do this.)
4.2.3 Given S: Ordering loci and testing for interference:

- Suppose we have three loci \( j = 1, 2, 3 \) at which \( S_{.,j} \) is observed. Assume recombination rates are the same for male and female meioses.

- We can choose the order that minimizes “double recombinants”: i.e. meioses \( i \) in which \( S_{i,.} = (0, 1, 0) \) or \( (1, 0, 1) \) or \( T_i = (1, 1) \).

- More generally, for \( \ell \) loci known to be linked, we can seek the ordering of columns \( j \) of \( S \) that minimizes recombination events.

- For any two locus intervals, \( I_j \) and \( I_k \), say, in the absence of interference \( T_{i,j} \) and \( T_{i,k} \) are independent if \( j \neq k \). (And the meioses \( i \) are independent.)

- So to test for interference between \( I_j \) and \( I_k \), we could just use a \( 2 \times 2 \) table for the counts of \( (T_{j}, T_{k}) \) over meioses.

- More generally (beyond the scope of this class!) we could fit a map function to the patterns of recombination we see.

<table>
<thead>
<tr>
<th>( T_{i,k} )</th>
<th>( T_{i,j} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>( \rho_j \rho_k )</td>
</tr>
<tr>
<td>(1 - ( \rho_j ))</td>
<td>( \rho_j )</td>
</tr>
</tbody>
</table>
4.3.1 Baum algorithm for total probability:

- For data observations $Y = (Y_{*,j}, j = 1, \ldots, \ell)$, we want to compute $Pr(Y)$. Due to the first-order Markov dependence of the $S_{*,j}$, we have

$$Pr(Y) = \sum_s Pr(S, Y) = \sum_s Pr(Y | S) Pr(S)$$

$$= \sum_s \left( Pr(S_{*,1}) \prod_{j=2}^\ell Pr(S_{*,j} | S_{*,j-1}) \prod_{j=1}^\ell Pr(Y_{*,j} | S_{*,j}) \right).$$

- Let $Y^{*(j)} = (Y_{*,1}, \ldots, Y_{*,j})$, the data along the chromosome up to and including locus $j$. Note $Y = Y^{*(\ell)}$.

4.3.2 The forwards Baum algorithm:

- Now define the joint probability

$$R_j^*(s) = Pr(Y_{*,k}, k = 1, \ldots, j - 1, S_{*,j} = s) = Pr(Y^{*(j-1)}, S_{*,j} = s)$$

with $R_1^*(s) = Pr(S_{*,1} = s) = (1/2)^m$. Then

$$R_{j+1}^*(s) = \sum_{s^*} [Pr(S_{*,j+1} = s | S_{*,j} = s^*) Pr(Y_{*,j} | S_{*,j} = s^*) R_j^*(s^*)]$$

for $j = 1, 2, \ldots, \ell - 1$, with

$$Pr(Y) = \sum_{s^*} Pr(Y_{*,\ell} | S_{*,\ell} = s^*) R_{\ell}^*(s^*).$$

- That is, we can compute the likelihood $Pr(Y)$. 

4.3.3 The Lander-Green algorithm: Lander and Green (1987):

- The Genehunter algorithm is the forwards algorithm of 4.3.2.
- If there are \( m \) meioses on the pedigree, then \( S_{*,j} \) can take \( 2^m \) values. Computations involve, for each locus, transitions from the \( 2^m \) values of \( S_{*,j} \) to the \( 2^m \) values of \( S_{*,j+1} \).
- Overall computation is order \( \ell 2^{2m} \).

For Genehunter, for a pedigree with \( n \) individuals, \( f \) of whom are founders, \( m = 2(n - f) - f = 2n - 3f \), and \( m \leq 16 \).
- We can compute \( \Pr(Y_{*,j} \mid S_{*,j}) \) for genetic marker data (2.5.3-5).
  Also for data at a trait locus, where we observe only phenotypes not genotypes, although this is (a bit) harder.
- Even if computation of \( \Pr(Y_{*,j} \mid S_{*,j}) \) is easy for given \( S_{*,j} \), this must be done for each locus and for each value of \( S_{*,j} \).
- The exact Lander-Green computation is limited to small pedigrees. Although better algorithms using independence of meioses give us a factored HMM which means we can get an algorithm of order \( m \ell 2^m \) but is is still exponential in pedigree size. (MERLIN: \( m \leq 27 \).)

4.3.4 The linkage map-specific lod score:

- We hypothesize the trait locus at some position \( d \) on the chromosome, measured in genetic distance (cM):
  \[
  L(d) = \Pr(Y \mid \text{trait locus is at } d)
  \]
  \( d = \infty \) corresponds to \( \rho = \frac{1}{2} \), or absence of linkage.
- For Genehunter, distances are relative to first marker at \( d = 0 \).
- The map-specific lod score is \( \log_{10}(L(d)/L(\infty)) \), measured in genetic distance.
- The location score is defined as \( 2 \log_e(L(d)/L(\infty)) \). Under appropriate conditions, this statistic has approximately a chi-squared distribution in the absence of linkage.
- Software for map-specific lod scores is implemented in Genehunter, Allegro, and MERLIN (recommended for small pedigrees). (Monte Carlo and/or MCMC versions are implemented in SIMWALK-2 and in MORGAN.)