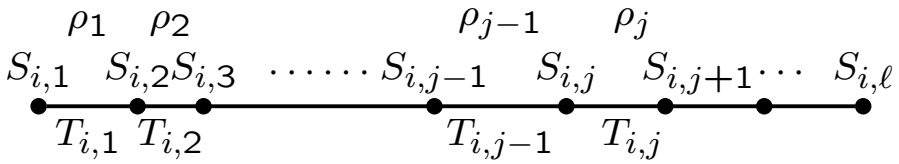
Chapter 4: Multiple Marker Loci

4.1.1 MULTILOCUS INHERITANCE SPECIFICATION:



• Assume that ℓ 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.
ρ_j is probability of recombination between locus j and locus $j+1$.

• $S_{\bullet,j} = \{S_{i,j}, i = 1, ..., m\}$ is the inheritance vector at locus j.
$S_{i,\bullet} = \{S_{i,j}, j = 1, ..., \ell\}$ 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,

$$\begin{aligned} T_{i,j} &= 1 \text{ if } S_{i,j} \neq S_{i,j+1}, \text{ and} \\ T_{i,j} &= 0 \text{ if } S_{i,j} = S_{i,j+1}, \quad j = 1, \ldots, \ell - 1. \\ \Pr(T_{i,j} = 1) &= \Pr(S_{i,j} \neq S_{i,j+1}) = \rho_j. \end{aligned}$$

4.1.2 MULTILOCUS INHERITANCE; NO INTERFENCE:

• A model for $S_{i,\bullet} = \{S_{i,j},\ j = 1, ..., \ell\}$ is equivalent to a model for $(T_{i,1}, \ldots, T_{i,\ell-1})$; 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

$$\Pr(S_{i,j} \mid S_{i,1}, ..., S_{i,j-1}) \ = \ \Pr(S_{i,j} \mid S_{i,j-1})$$

so that
$$\Pr(S_{i,\bullet}) \ = \ \Pr(S_{i,1}) \prod_{j=2}^{\ell} \Pr(S_{i,j} \mid S_{i,j-1})$$

• Combining the meioses

$$\Pr(\mathbf{S}) \ = \ \Pr(S_{\bullet,1}) \prod_{j=2}^{\ell} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1})$$

where $\mathbf{S} \ = \ \{S_{i,j};\ i = 1, ..., m,\ j = 1, ..., \ell\}$.

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, $\mathbf{S}_{-(i,j)} = \{S_{k,l};\ (k,l) \neq (i,j)\}$.

• Then $S_{i,j}$ depends only on the indicators for the same meiosis and the two neighboring loci. For $s = 0, 1$,

$$\Pr(S_{i,j} = s \mid \mathbf{S}_{-(i,j)}) \ = \ \Pr(S_{i,j} = s \mid S_{i,j+1}, S_{i,j-1})$$
$$\propto \ \rho_{j-1}^{|s-S_{i,j-1}|}(1-\rho_{j-1})^{1-|s-S_{i,j-1}|}\ \rho_j^{|s-S_{i,j+1}|}(1-\rho_j)^{1-|s-S_{i,j+1}|}$$

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 PROBABILITIRS:

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 $\mathbf{J}$, or using the inheritance $\mathbf{S}$.

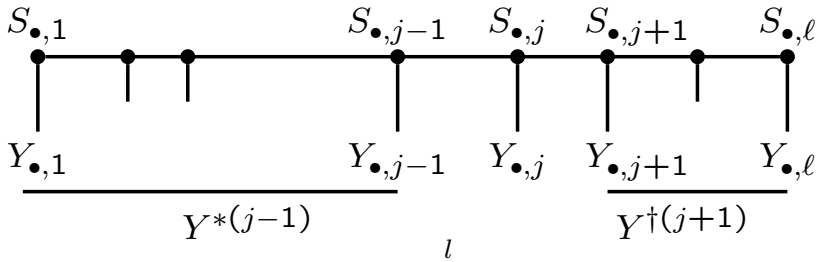
$$\Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) \Pr(\mathbf{S})$$
$$= \sum_{\mathbf{J}} \Pr(\mathbf{Y} \mid \mathbf{J}) \Pr(\mathbf{J}).$$

• In examples we used the *ibd* states, because there are fewer *ibd* patterns $\mathbf{J}$ than values of $\mathbf{S}$. 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 $S_{\bullet,j}$ 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 $\mathbf{S}$.

• Now let $Y_{\bullet,j}$ denote all the data corresponding to locus j.

4.1.5 THE HMM ACROSS LOCI FOR PEDIGREE DATA:



$$\Pr(\mathbf{S}) = \Pr(S_{\bullet,1}) \prod_{j=2}^{l} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1})$$

• As before $S_{\bullet,j}$ determines the *ibd* at locus j, and hence $\Pr(Y_{\bullet,j}|S_{\bullet,j})$.

$$\text{Then} \quad \Pr(\mathbf{Y} \mid \mathbf{S}) = \prod_{j=1}^{\ell} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}).$$

• Note that, given $S_{\bullet,j}$,
$Y^{*(j-1)}$, $Y_{\bullet,j}$, and $Y^{\dagger(j+1)}$ are mutually independent.

Also, given $S_{\bullet,j}$, $Y^{*(j-1)}$, $Y_{\bullet,j}$, and $S_{\bullet,j+1}$ are independent.
Also, given $S_{\bullet,j}$, $Y^{\dagger(j+1)}$, $Y_{\bullet,j}$, and $S_{\bullet,j-1}$ are independent.

4.2.1 Counting recombinants if S is observed:

• If $\mathbf{S}$ 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.

• $\mathbf{Y}$ is irrelevant to ρ-estimation, and the log-likelihood is

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

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

$$\begin{aligned} \log(\Pr(S_{\bullet,j+1}|S_{\bullet,j})) &= \\ X_{m,j} \log(\rho_{m,j}) &+ (M_m - X_{m,j}) \log(1-\rho_{m,j}) \\ + X_{f,j} \log(\rho_{f,j}) &+ (M_f - X_{f,j}) \log(1-\rho_{f,j}) \end{aligned}$$

• $\widehat{\rho_{m,j}} = X_{m,j}/M_m$, and $\widehat{\rho_{f,j}} = X_{f,j}/M_f$,

4.2.2 $\mathbf{S}$ unobserved: An EM algorithm for genetic maps:

• $\rho_{m,j}$ and $\rho_{f,j}$ occur only in the term $\log(\Pr(S_{\bullet,j+1} \mid S_{\bullet,j})$ of the complete-data log-likelihood $\log \Pr(\mathbf{S}, \mathbf{Y}) =$

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

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

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

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

• M-step: The new estimate of $\rho_{m,j}$ is $\tilde{X}_{m,j}/M_m$, 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 $\mathbf{S}$: Ordering loci and testing for interference:

• Suppose we have three loci $j = 1, 2, 3$ at which $S_{\bullet,j}$ 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 $S_{i,\bullet} = (0, 1, 0)$ or $(1, 0, 1)$ or $T_i = (1, 1)$.

• More generally, for ℓ loci known to be linked, we can seek the ordering of columns j of $\mathbf{S}$ 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.)

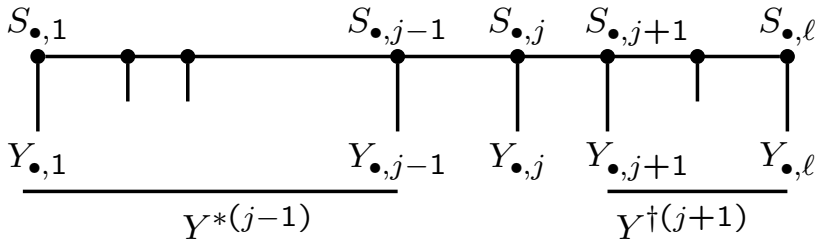
$T_{i,k}$	$T_{i,j}$		
	0	1	
0			$(1-\rho_k)$
1		$\rho_j\rho_k$?	ρ_k
	$(1-\rho_j)$	ρ_j	1

• So to test for interference between I_j and I_k, we could just use a 2×2 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.

blank slide:

4.3.1 Baum algorithm for total probability:



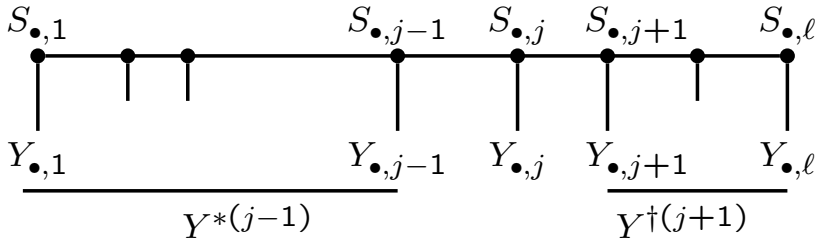
• For data observations $\mathbf{Y} = (Y_{\bullet,j}, j = 1, \ldots, \ell)$, we want to compute $\Pr(\mathbf{Y})$. Due to the first-order Markov dependence of the $S_{\bullet,j}$, we have

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{S}, \mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S})$$

$$= \sum_{\mathbf{S}} \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{\ell} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \prod_{j=1}^{\ell} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right).$$

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

4.3.2 The forwards Baum algorithm:



• Now define the joint probability

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

with $R^*_1(s) = \Pr(S_{\bullet,1} = s) = (1/2)^m$. Then

$$R^*_{j+1}(s) = \sum_{s^*} \left[\Pr(S_{\bullet,j+1} = s \mid S_{\bullet,j} = s^*) \Pr(Y_{\bullet,j} \mid S_{\bullet,j} = s^*)\ R^*_j(s^*) \right]$$

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

$$\Pr(\mathbf{Y}) = \sum_{s^*} \Pr(Y_{\bullet,\ell} \mid S_{\bullet,\ell} = s^*)\ R^*_\ell(s^*).$$

• That is, we can compute the likelihood $\Pr(\mathbf{Y})$.

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 $S_{\bullet,j}$ can take 2^m values. Computations involve, for each locus, transitions from the 2^m values of $S_{\bullet,j}$ to the 2^m values of $S_{\bullet,j+1}$.

• 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 $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ 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 $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ is easy for given $S_{\bullet,j}$, this must be done for each locus and for each value of $S_{\bullet,j}$.

• 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):

$$L(d) \;=\; \Pr(\mathbf{Y} \mid \text{trait locus is at } d)$$

$d = \infty$ 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.)