4.2 Multilocus linkage analysis
4.2.1 Meiosis indicators at multiple loci

- For multiple loci, j, j=1,…,*L*
  S_ij = 0 if gene at meiosis i locus j is parent's maternal
  = 1 if gene at meiosis i locus j is parent's paternal.

- We define Sv(j) = { S_ij; i=1,… , m }, for j=1,… , *L*
  Sm(i) = { S_ij; j=1,… , *L*}, for i=1,… , m
  where m is the number of meioses in the pedigree, and *L* the number of loci along the chromosome.

- Dependence of the {S_ij}:
  Sm(i) are independent over i , i=1,...,m.
  S_ij are independent for loci j on different chromosome pairs
  Sv(j) are dependent among loci j on the same chromosome pair

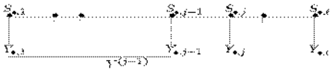
4.2.2 Conditional independence (no interference)

- Assume that *L* loci are ordered 1,…,*L* along the chromosome
- Let the intervals between successive loci be I(1),…,I(*L*– 1).
- Let T(i, j)=1 if a gamete resulting from meiosis i is recombinant on interval I(j), and T(i, j) =0 otherwise (j=1,…,*L*-1).
- Then, in a given meiosis i
  T(i, j) = 1 if S_ij ≠ S_i,j+1,
  and T(i, j) = 0 if S_ij = S_i,j+1, for j=1,…,*L*-1.
- A model for Sm(i) is equivalent to a model for (T(i,1}, … , T(i,*L*-1)).
- The simplest models for meiosis assume *no interference*.
- In this case the T(i,j) are independent over i and j.
- Then the S_ij are first-order Markov over loci j, with meioses i always being independent.
- One way to express this is that
  P(S_ij | S_i1,..., S_i,j-1) = P(S_ij | S_i,j-1) so that
  P(Sm(i)) = P(S_i,1) Π_{j=2}^*L* P(S_ij | S_i,j-1) or,
  combining the meioses
  P({S_ij}) = P(Sv(1)) Π_{j=2}^*L* P(Sv(j) | Sv(j-1)) (see also 4.2.4).

- Another way of expressing this Markov dependence is through the probability of any given indicator S_ij conditional on all the others.
- S_ij given S_{-(i,j)} = { S_kl; (kl) ≠ (i,j)}, depends only on the indicators for the same meiosis and the two neighboring loci.
- For s=0,1, P(S_ij =s | S_{-(i,j)}) = P(S_ij=s | S_i,j+1, S_i,j-1) which is proportional to
  ρ(j-1)^|s-S_i,j-1| × (1 - ρ(j-1))^{1-|s-S_i,j-1| } ×
  × ρ(j)^|s-S_i,j+1| × (1 - ρ(j))^{1-|s-S_i,j+1|} where
  ρ(j) = P(T(i,j) = 1) = P(S_ij ≠ S_i,j+1) is the recombination frequency in I(j).
- Note that the equation just counts the recombination/non-recombination events in intervals I(j-1) and I(j), implied by the three indicators (S_i,j-1, S_ij=s, S_i,j+1).

- Recall in Chapter 2 we discussed for a single locus the equations
  P(Y) = Σ_{S_ij} P(Y | {S_ij}) P({S_ij}) = Σ_{S_ij} P(Y | J({S_ij}) P({S_ij})
  = Σ_J P(Y | J) P(J), where J was the ibd pattern determined by {S_ij}.
- There are fewer ibd patterns than values of {S_ij}. However, although the component S_ij are Markov over loci j, gene ibd patterns are not.
- Different values of Sv(j) may give rise to the same ibd pattern at locus j.
  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_ij}.

4.2.3 The hidden Markov structure



- The conditional independence structure of data, in the absence of genetic interference.
- The figure shows the Markov dependence of the Sv(j).
- Also the data Yv(j) at locus j depends only on the inheritance Sv(j) at that locus, (and on allele frequencies etc. for locus j).
- Given Sv(j), {Yv(k), k=(j+1), … *L*}, Yv(j), and Sv(j-1) are mutually independent.
- OR, given Sv(j), {Yv(k),k=1, … ,j-1} = Y^(j-1), Yv(j), and Sv(j+1) are mutually independent.

4.2.4 Baum algorithm for total probability

- For data observations Y=(Yv(j), j=1,…,*L*), we want to compute P(Y).
- Due to the first-order Markov dependence of the Sv(j), we have
  P(Y) = Σ_{S_ij} P({S_ij}, bY) = Σ_{S_ij} P(Y | {S_ij}) P({S_ij})
  = Σ_{S_ij} (P(Sv(1)) Π_{j=2}^*L* P(Sv(j) | Sv(j-1)))
  (Π_{j=1}^*L* P(Yv(j) | Sv(j))).
- We can go forwards. Let Y^(j) = (Yv(1), …, Yv(j)), the data along the chromosome up to and including locus j. Note Y = Y^(*L*).
- Now define the joint probability
  R*_j(s) = P(Yv(k), k=1,... j-1, Sv(j)=s) = P(Y^(j-1), Sv(j)=s)
  with R*_1(s) = P(Sv(1) =s).
- Then for j=1,2,…,*L*-1
  R*_{j+1}(s) = Σ_s* (P(Sv(j+1) =s | Sv(j)=s*)
  P(Yv(j) | Sv(j) =s*) R*_j(s*)),
- With P(Y) = Σ_s* P(Yv(*L*) | Sv(*L*) =s*) R*_*L* (s*).

4.2.5 Lander-Green algorithm

- We can compute P(Yv(j) | Sv(j)) for simple traits– recall the example at end of Chapter 2.Then the computation method of 4.2.4 can be applied.
- However this exact computation is limited to small pedigrees. If there are m meioses on the pedigree, then Sv(j) can take 2^m values. Computations involve, for each locus, transitions from the 2^m values of Sv(j) to the 2^m values of Sv(j+1).
- Computation is of order *L* 2^m 2^m= *L* 4^m. For Genehunter, for a pedigree with n individuals, f of whom are founders, m = 2 n -3 f, and m ≤ 16.
- Additionally, for each locus and for each value of Sv(j), we must compute P(Yv(j) | Sv(j)). Although this is easy for given Sv(j), this limits size of pedigree.
- Actually better algorithms using independence of meioses give us a *factored HMM* which means we can get an algorithm of order *L* m 2^m but it is still exponential in pedigree size.
- The map-specific lod score is log_10 (L(d)/L(∞)), where d is the hypothesized chromosomal location of the trait locus measured in genetic distance, and d=∞ corresponds to ρ=1/2, or absence of linkage. (For Genehunter, distances are relative to first marker at d=0.)
- The *location score* is defined as 2 log_e (L(d)/L(∞)). Under appropriate conditions, this statistic has approximately a chi-squared distribution in the absence of linkage.
- We consider lod scores for the location d, rather than location scores.
- Genehunter, Allegro, and Merlin are packages using this general approach.

4.2.6 EM algorithm
for estimating genetic maps

- Consider the complete-data log-likelihood
  log P({S_ij}, Y) = log (P(Sv(1)) +Σ_{j=2}^*L* log (P(Sv(j) | Sv(j-1)))
  + Σ_{j=1}^*L* log (P(Yv(j) | Sv(j)))
- Now recombination parameters enter through
  log (P(Sv(j) | Sv(j-1))) =
  R_{m,j-1} log(ρ_{m,j-1}) + (M_m – R_{m,j-1}) log (1- ρ_{m,j-1})
  + R_{f,j-1} log(ρ_{f,j-1}) +(M_f – R_{f,j-1}) log (1- ρ_{f,j-1})
- where R_{m,j-1} = Σ_{i male} | S_{i,j} - S_{i,j-1} | is the number of recombinations in interval I(j-1) in male meioses, ρ_{m,j-1} the recombination rate, and M_m is the total number of male meioses scored in the pedigree.
- and similarly R_{f,j-1}, ρ_{f,j-1}) and M_f for female meioses.
- The expected complete-data log-likelihood requires only computation of
  R*_{m,j-1} = E (R_{m,j-1} | Y) = Σ_{i male} E(| S_ij - S_i,j-1 | | Y)
  and similarly R*_{f,j-1}.

- Since the complete-data log-likelihood is a simple binomial log-likelihood, the M-step sets the new estimate of ρ_{m,j-1} to R*_{m,j-1}/M_m, and similarly for all intervals j=2,3,…,*L* and for both the male and female meioses.
- Note that P(Sv(j-1), Sv(j) | Y) = P(Sv(j-1), Sv(j) , Y) / P(Y)
  and P(Sv(j-1), Sv(j) , Y) = P(Y^(j-2), Sv(j-1)) P(Yv(j-1) | Sv(j-1))
  P(Sv(j) | Sv(j-1)) P(Yv(j) | Sv(j)) P(Yv(j+1), …., Yv(*L*) | Sv(j))
- The first term is just the R*_{j-1}(Sv(j-1)) we had in the Baum algorithm, the second and fourth are just single-locus probabilities of data given inheritance, the third is just the recombination/non-recombination transitions, and the final one can be computed by backwards (conditional) version of the Baum algorithm.
- Note there are many different forms of the Baum algorithm 4.2.4, all closely related but providing probabilities of slightly different events.
- The EM algorithm is thus readily implemented to provide maximum likelihood (MLE) estimates of recombination frequencies for all intervals and for both sexes.