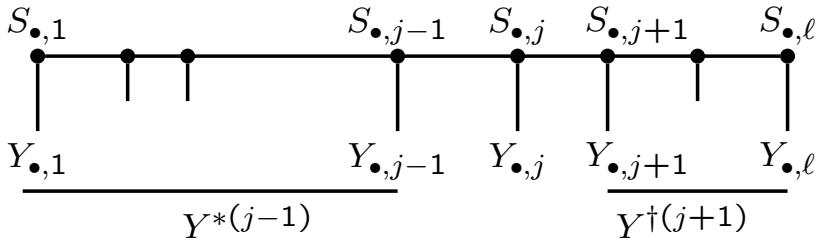
4.3.1 Baum algorithm for total probability:



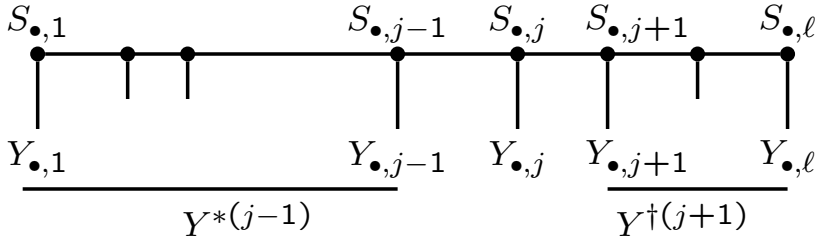
• For data observations $\mathbf{Y} = (Y_{\bullet,j}, j = 1, \ldots, \ell)$, we want to compute $\Pr(\mathbf{Y})$. Due to the first-order Markov dependence of the $S_{\bullet,j}$, we have

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

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

• 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 $R^*_1(s) = \Pr(S_{\bullet,1} = s)$. 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 (2.4.8).
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.

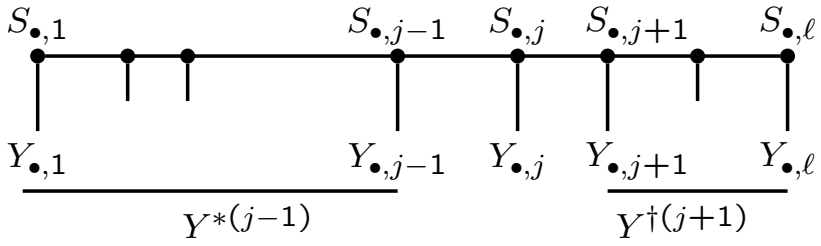
• For Genehunter, distances are relative to first marker at $d = 0$.

• 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.)

4.4.1 The backwards Baum algorithm and $\Pr(S_{\bullet,j} \mid \mathbf{Y})$:



- Now also define

$$R_j^\dagger(s) \;=\; \Pr(Y_k, k = j, \ldots, \ell \mid S_{\bullet,j} = s) = \Pr(Y^{\dagger(j)} \mid S_{\bullet,j} = s).$$

- $$\begin{aligned} R_j^\dagger(s) \;&=\; \Pr(Y_k, k = j, \ldots, \ell \mid S_{\bullet,j} = s) \\ &= \sum_{s^*} \Pr(Y_k, k = j, \ldots, \ell, S_{\bullet,j+1} = s^* \mid S_{\bullet,j} = s) \\ &= \Pr(Y_j \mid S_{\bullet,j} = s) \sum_{s^*} R_{j+1}^\dagger(s^*) \Pr(S_{\bullet,j+1} = s^* | S_{\bullet,j} = s) \end{aligned}$$

- Then $\Pr(S_{\bullet,j} = s \mid \mathbf{Y}) \;=\; \dfrac{\Pr(\mathbf{Y}, S_{\bullet,j} = s)}{\Pr(\mathbf{Y})} \;=\; \dfrac{R_j^*(s) R_j^\dagger(s)}{\Pr(\mathbf{Y})}$

4.4.2 Expected recombination counts: implementing EM:

- Recall from 4.2.3 we want

$$\tilde{X}_{j-1} \;=\; \mathbf{E}(X_{j-1} \mid \mathbf{Y}) \;=\; \sum_i \mathbf{E}(|S_{i,j} - S_{i,j-1}| \mid \mathbf{Y})$$

- Note that

$$\begin{aligned} \Pr(S_{\bullet,j-1}, S_{\bullet,j} \mid \mathbf{Y}) \;&=\; \Pr(S_{\bullet,j-1}, S_{\bullet,j}, \mathbf{Y}) / \Pr(\mathbf{Y}) \text{ and} \\ \Pr(S_{\bullet,j-1}, S_{\bullet,j}, \mathbf{Y}) \;&=\; \Pr(Y^{*(j-2)}, S_{\bullet,j-1}) \Pr(Y_{\bullet,j-1} \mid S_{\bullet,j-1}) \\ &\quad \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) P(Y^{\dagger(j)} \mid S_{\bullet,j}) \end{aligned}$$

- The first term is just the $R_{j-1}^*(S_{\bullet,j-1})$ we had in the forwards Baum algorithm, the second is just a single-locus probability of data given inheritance, the third is just the recombination/non-recombination transitions in I_{j-1} interval, and the final is $R_j^\dagger(S_{\bullet,j})$ from the backwards version of the Baum algorithm.

- On small pedigrees, the EM map estimation can be implemented.

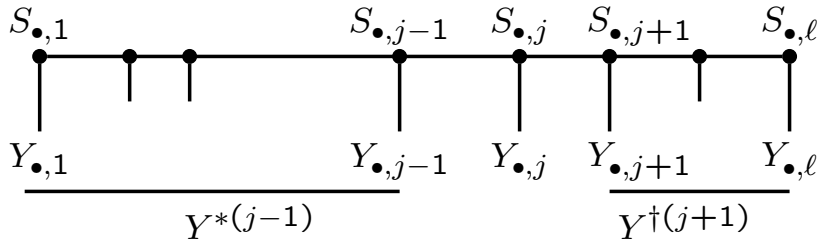
4.4.3 The joint pattern of S over loci:

• 4.4.1 gives us probabilities of $S_{\bullet,j}$ given $\mathbf{Y}$ and hence probabilities of *ibd* at each locus j. Each $S_{\bullet,j}$ can take 2^m values.

• 4.4.2 gives us pairwise probabilities of $(S_{\bullet,j-1}, S_{\bullet,j})$, and hence expected recombination counts, given $\mathbf{Y}$. Each $(S_{\bullet,j-1}, S_{\bullet,j})$ can take $2^m \times 2^m = 4^m$ values.

• But suppose we want $\mathbf{S}$ jointly over all the loci; this is infeasible to compute eactly, even on small pedigrees. $\mathbf{S}$ can take $2^{m\ell}$ values – there are too many possible $\mathbf{S}$.

4.4.4 Monte Carlo realization of S given Y:



• Compute $R^*_j(s) = \Pr(Y^{*(j)}, S_{\bullet,j} = s)$, $j = 1, 2, 3, ...\ell$ as before.

• First, $S_{\bullet,\ell}$ is sampled from $\propto \Pr(\mathbf{Y}, S_{\bullet,\ell}) = \Pr(Y_{\bullet,\ell}|S_{\bullet,l})R^*_\ell(S_{\bullet,\ell})$.
(All sampling probabilities will be normalized over 2^m s-values.)

• Then, given a realization of $(S_{\bullet,j+1} = s^*, S_{\bullet,j+2}, \ldots, S_{\bullet,\ell})$,

$$\Pr(S_{\bullet,j} = s \mid S_{\bullet,j+1} = s^*, S_{\bullet,j+2}, \ldots, S_{\bullet,\ell}, \mathbf{Y}) =$$
$$\Pr(S_{\bullet,j} = s|S_{\bullet,j+1} = s^*, Y^{*(j)}) \propto \Pr(S_{\bullet,j+1} = s^*|S_{\bullet,j} = s) R^*_j(s)\Pr(Y_{\bullet,j}|S_{\bullet,j})$$

• Normalizing these probabilities, we realize each $S_{\bullet,j-1}$, for $j = \ell, \ell-1, \ldots, 4, 3, 2$ in turn, providing an overall realization $\mathbf{S} = (S_{\bullet,1}, \ldots, S_{\bullet,\ell})$ from $\Pr(\mathbf{S} \mid \mathbf{Y})$.

4.4.5 Monte Carlo estimates of *ibd* and genetic maps:

• Instead of computing $\Pr(S_{\bullet,j} \mid \mathbf{Y})$ we can sample it, and hence get estimates of *ibd* patterns at each locus j.

• We can make estimates of *ibd* jointly over loci – for example, the probability that an individual is autoygous over a set of loci, not just the separate probability of each.

• Intead of computing

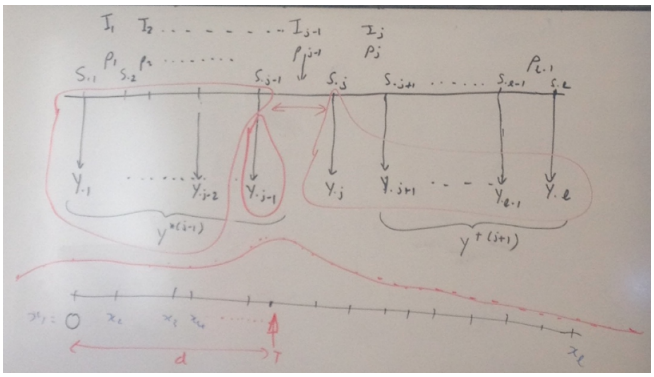
$$\tilde{X}_{j-1} \;=\; \mathbf{E}(X_{j-1} \mid \mathbf{Y}) \;=\; \sum_i \mathbf{E}(|S_{i,j} - S_{i,j-1}| \mid \mathbf{Y})$$

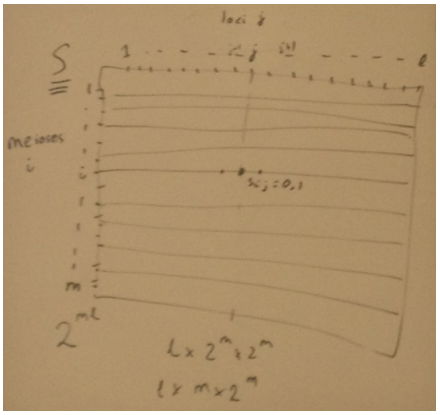
we can count recombination events in N realized $\mathbf{S}$ for all map intervals and each gender.

• Hence we can do Monte Carlo EM, replacing the E-step by these Monte Carlo estimates at each stage.

• Generally, Monte Carlo EM works as well as regular EM, at least for nitial steps. Initially, the Monte Carlo sample size N need not be large, although for the final EM steps it should be increased.

Photos from Qian:





• Left: Above: Showing forwards (R^*) and backwards ($R^\dagger$) components of the HMM lkelihood computation. Below: a map-specific lod score computed at locations along a chromosome.

• Right: The grid of $2^{m\ell}$ components $S_{,j}$ of $\mathbf{S}$, showing the Markov dependence across loci j, and the computational complexity of HMM computations.