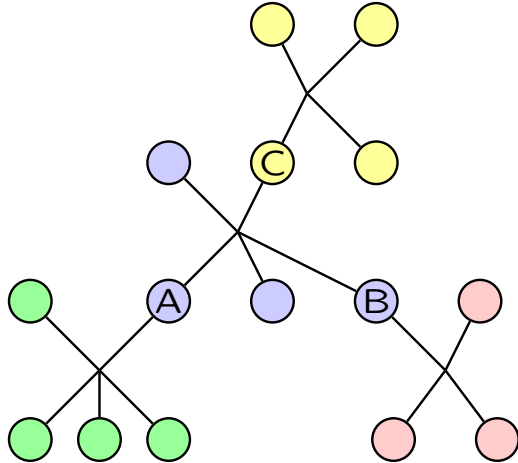


4.5.1 Conditional independence over pedigree members:

- We get our DNA from our parents and copy to our offspring.
All meioses (transmissions of DNA) are independent.
Given the genotypes of parents, sibs are independent, and independent of grandparents, uncles, cousins,

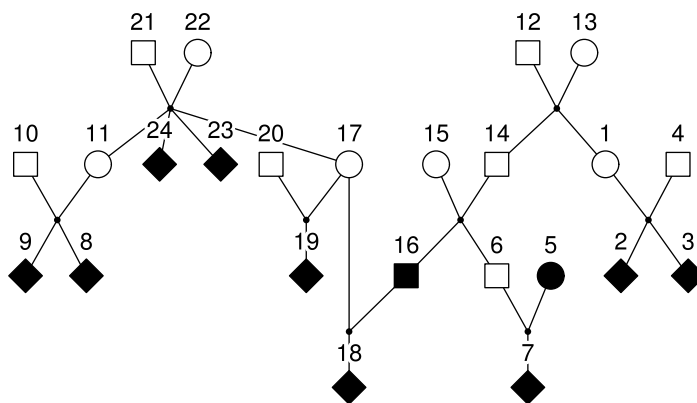


- Sum over the green family for each genotype of A
- Sum over the pink family for each genotype of B
- Using these values, sum over the blue family for each genotype of C.
- Using these values for each C, sum over the yellow family.

- First used by [Haldane and Smith \(1947\)](#) (without computers)
Formalized (and programmed) by [Elston and Stewart \(1971\)](#).

4.5.2 Elston-Stewart algorithm:

- An example pedigree without loops. Shaded individuals are those for whom phenotypic data are assumed to be available



- Unlike a chromosome, pedigrees have a natural direction; probabilities of a set of offspring given the two parental genotypes is easier than of the parents given the offspring.

4.5.3 Pedigree peeling:

- Similar ideas to HMM underlie pedigree peeling. We don't have Markov chain, but, for pedigrees with no loops, we do have that conditionally of genotype (pair of haplotypes) of each individual, the data "above" are independent of data "below".

- For any individual i Define:

$$\text{Forwards/Down: } R_i^*(g) = \Pr(\text{data above, } G_i = g)$$

$$\text{Backwards/Up: } R_i^\dagger(g) = \Pr(\text{data below} \mid G_i = g)$$

- Equations very similar to HMM, at each stage putting in genotype probabilities (if a founder), transmission probabilities from parents to offspring, and probabilities of data marker and/or trait) given the genotype.

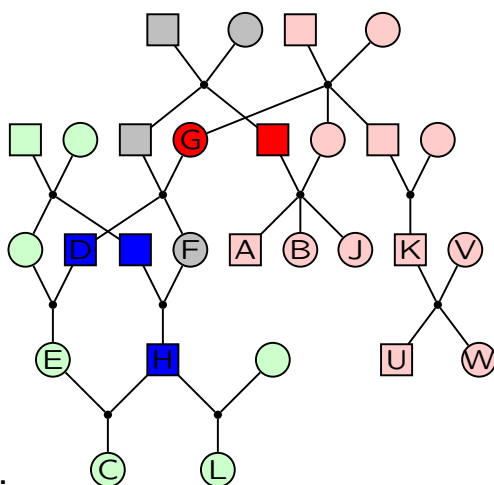
- Hence, we can sum over all the unobserved genotypes to compute the probability of trait and/or marker data; that is, the likelihood $\Pr(\mathbf{Y})$.

- Implemented in LIPED, LINKAGE, FASTLINK, VITESSE

4.5.4 Joint pedigree peeling:

- The main problem with pedigree peeling is that the genotypes we sum over, and the transmissions we consider, are of pairs of haplotypes, jointly over loci.

- So computation is now linear in pedigree size but the number of genotypes each individual can have is exponential in number of loci.



- Methods can be extended to pedigrees with loops, but this is even more computationally intensive.

- Given genotypes on a cutset, the data on individuals each side of the cutset are independent. – e.g. the red/pink and blue/green sets.

- Above/below no longer well-defined. Individuals may be Below some but above others in the same cutset.

4.5.5 Monte Carlo given trait data – SIMLINK:

- In early 1980s, often we had trait data; trait data had simple known models (e.g. dominant, recessive)
- Marker maps were just starting; marker typing was expensive. How to persuade NIH to fund a linkage study? How to show the potential data would have power to detect linkage, if present?
- We can simulate trait and marker data jointly quite easily, assuming some trait model. But it is much better to simulate what marker data would look like conditional on the trait data we have.
- [Ploughman and Boehnke \(1989\)](#) solved this problem: see next slide. Their SIMLINK program produced Elods given the observed data, and assuming a marker of some informativeness (e.g. 4 equifrequent alleles), at some distance from the trait locus (e.g. 10 cM).
- If this Elod is “big enough” then this indicates that the trait data are sufficient for marker typing to be worth doing. For some time, NIH required SIMLINK evidence in grant proposals.

4.5.6 SIMLINK – cf markerdrop for steps (1) to (4):

- (1) Peel up at the trait locus, saving the $R_i^\dagger(g)$ for each trait-locus genotype g of each individual i : $R_i^\dagger(g) = \Pr(\text{data below} \mid G_i = g)$.
- (2) Assign founder genotypes at the top, in accordance with trait model and upward peeling.
- (3) Simulate genotypes back down at the trait locus, using the saved partial sums — see 4.4.4 for HMM analogy.
- (4) Simulate marker genotypes at loci linked to the trait, at some assumed marker allele frequencies and recombination fraction.
- (5) Do the lod score for each simulated data set.
- (6) Average over simulated data sets to compute an empirical Elod.

Note the analogy to the HMM realization of S (see 4.4.4). Here we first compute up (backwards in time), and then simulate down (forwards). (Note 2: Could peel any direction, but data are at bottom.)

Note 3: MORGAN/markerdrop is similar, but uses $S_{\cdot,j}$ not genotypes.

4.6.1 What about big pedigrees with many markers??:

- The Lander-Green (or more generally HMM) approach is restricted to relatively small pedigrees ($m \leq 27$).
- The Elston-Stewart approach (or more generally pedigree-peeling) is restricted to just a few loci ($\ell \leq 4$).
- What about m and ℓ both large?

(1) Break up the pedigrees (to reduce m), and pretend they are unrelated. So compute lod scores independently for each piece, and sum them. **This loses information.**

(2) Break up the chromosome (to very small ℓ) – e.g. interval mapping, in which we use just two markers at a time, and hypothesize trait locus positions between them. **This loses information** and requires very tedious piecing together many lod-score segments (which don't match at the marker locations).

4.6.2 Monte Carlo Estimation of lod scores:

- This formulation due to [Lange and Sobel \(1991\)](#). It is similar to version implemented in MORGAN/Im_linkage.
- Let Z be trait data, and \mathbf{Y} the marker data. Let γ be hypothesized position of trait locus. Then

$$\frac{L(\gamma)}{L(\infty)} = \frac{P_\gamma(Z, \mathbf{Y})}{\Pr(Z) \cdot \Pr(\mathbf{Y})} = \frac{P_\gamma(Z | \mathbf{Y})}{\Pr(Z)}$$

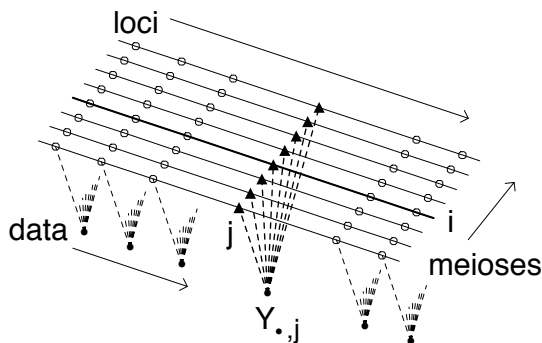
- Now let \mathbf{S} be the inheritance patterns $S_{i,j}$ at the marker loci.

$$P_\gamma(Z | \mathbf{Y}) = \sum_{\mathbf{S}} P_\gamma(Z | \mathbf{S}) \Pr(\mathbf{S} | \mathbf{Y})$$

- **Suppose** we could sample \mathbf{S} from $\Pr(\mathbf{S} | \mathbf{Y})$: giving realizations $\mathbf{S}^{(k)}$, $k = 1, \dots, N$. Then we could estimate $P_\gamma(Z | \mathbf{Y})$ by averaging the N values of $P_\gamma(Z | \mathbf{S}^{(k)})$. (These computed by pedigree peeling.)
- We can do this with one set of $\mathbf{S}^{(k)}$ for all the hypothesized γ and hence get a Monte Carlo estimate of the map-specific lod score.

4.6.3 MCMC sampling of S given marker data Y :

- The “SIMLINK” peeling approach (4.5.6) allows us to sample $S_{\cdot,j}$ given $Y_{\cdot,j}$ at a single locus j , but not all loci at once.
- The HMM approach (4.4.4) allows us to sample S given Y for a small number of meioses, but not all meioses at once.
- The independence of meioses i and Markov dependence of vectors $S_{\cdot,j}$ provide good MCMC block Gibbs samplers: for ℓ loci and m meioses.



L-sampler: resample $S_{\cdot,j}$ given Y and $S_{\cdot,j'}, j \neq j'$ Heath 1997.
Uses pedigree computations, $O(\ell m)$.

MM-sampler: resample $\{S_{i,\cdot}; i \in I^*\}$ $|I^*| = m^*$ given Y and $\{S_{i',\cdot}; i' \notin I^*\}$
Uses HMM computation,
 $O(\ell m^* 2^{m^*})$, $m^* \ll m$.
Tong & Thompson 2008

4.6.4 Added notes re MCMC:

- **L-sampler** is irreducible, but does not mix well.
Depending on data Y , **MM-sampler** may not be irreducible.
- Together (random choice at each step) they do fine, and can be used to provide estimates of the map-specific lod score across the chromosome.
- This is the form of MCMC implemented in MORGAN. SIMWALK-2 uses quite a different framework for the MCMC – but the same lod score estimator.
- For older, more spaced, microsatellite markers, we compute the lod score at points in marker intervals (see Lab 5). With newer denser SNP markers, we compute at a subset of the marker positions.
- The “lm” of `lm_auto` and `lm_linkage` programs in MORGAN indicates these are MCMC programs using the **LM-sampler** - this combination of locus-sampler (L) and meiosis-sampler (M).

4.6.5 Estimating *ibd* probabilities by MCMC:

- Recall the inheritance patterns S determine the *ibd* pattern among any set of “proband gametes” (maternal/paternal gametes of specified individuals).
- Since we have MCMC realizations of S given marker data Y , we can use these to estimate patterns of *ibd* given Y .
- We can estimate at each locus j using $S_{\cdot,j}$, or we can estimate jointly at sets of loci – e.g., over windows of a few loci.
- Data Z at a trait locus t also provide inheritance information (especially if for a rare almost recessive trait, such as in Lab 4).
- It is possible to compute $\Pr(Z | S_{\cdot,t})$ —not in this class!
So, we can include the trait locus and trait data Z in the MCMC, although computations are heavier.
- This is what is implemented in MORGAN/Im_auto, either for markers only, or for markers with a simple trait locus included (e.g. Lab 4).

blank slide:

References

- Bernstein F (1925) Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Zeitschrift für induktiv Abstammungs- und Vererbungslehre* 37:237–270. English translation in: *Selected contributions to the literature of blood groups and immunology (Dunsford Memorial). I. The ABO system*. U.S. Army Medical Research Laboratory; Fort Knox (1966)
- Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. *Advances of Applied Probability* 10:26–61
- Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution* 7:111–122
- Elston RC, Stewart J (1971) A general model for the analysis of pedigree data. *Human Heredity* 21:523–542
- Goddard KA, Yu CE, Oshima J, Miki T, Nakura J, Piusan C, Martin GM, et al. (1996) Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. *American Journal of Human Genetics* 58:1286–1302
- Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* 8:229–309
- Haldane JBS, Smith CAB (1947) A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* 14:10–31

Dr Elizabeth A Thompson

Stat 550: StatGen I: 2013

Bibliography - 2

- Hardy GH (1908) Mendelian proportions in a mixed population. *Science* 28:49–50
- Kosambi DD (1944) The estimation of map distances from recombination values. *Annals of Eugenics* 12:172–175
- Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences USA* 84:2363–2367
- Lange K, Sobel E (1991) A random walk method for computing genetic location scores. *American Journal of Human Genetics* 49:1320–1334
- Mather K (1938) Crossing-over. *Biological Reviews of the Cambridge Philosophical Society* 13:252–292
- Mendel G (1866) Experiments in Plant Hybridisation. In JH Bennett, ed., *English translation and commentary by R. A. Fisher*. Oliver and Boyd, Edinburgh, 1965
- Ploughman LM, Boehnke M (1989) Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* 44:543–551
- Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 78:629–644
- Wahlund S (1928) Zusammensetzung von population und korrelationserscheinung vom standpunkt der vererbungslehre aus betrachtet. *Hereditas* 11:65–106

Dr Elizabeth A Thompson

Stat 550: StatGen I: 2013