

Lecture 10

Linkage Analysis

February 1, 1996

Lecturer: Elizabeth Thompson

Notes: Brendan Mumeey

For details beyond what are provided in the lecture, see Prof. Thompson's recent technical report [3], and related materials accessible from her home page <http://www.stat.washington.edu/thompson/>. Scanned versions of her slides used for this lecture can be found on the course web (<http://www.cs.washington.edu/education/courses/590bi/>).

Although the title of this lecture is “linkage analysis” it is going to be mostly about computing probabilities. Linkage analysis can be performed for a variety of species. We will concentrate on the case where the individuals in the pedigree are humans. Human pedigrees present some unique problems in that the data is often incomplete. In particular some individuals may be unobserved, or the trait genotypes and possibly the phase of homologous pairs may be unknown.

We begin with a quick review of linkage analysis. The underlying problem is to compare alternate hypotheses about the location of genes in the genome and their penetrance (defined in Lecture 7). For example, if two genes are close together (linked) then they will tend to be co-inherited. A likelihood ratio test is used to test this hypothesis given the pedigree data:

$$\frac{L(\text{linkage})}{L(\text{unlinked})} = \frac{P(\text{data}|\text{linked at } r)}{P(\text{data}|r = 0.5)} \quad (10.1)$$

The parameter r is the *recombination probability*: the probability of an odd number of recombinations between the two gene loci. When $r = 0.5$ the genes are unlinked. Note that we are really just interested in the ratio and not the actual likelihoods (which will be quite small, and get smaller with more data).

10.1. Conditional Independence in Genetics

A property of pedigrees which is valuable in simplifying calculations is that the probability of an individual having a particular genotype depends only on the genotypes of the individual's parents, spouse, and offspring, the *neighborhood* of the individual. For any question $?$ about the genotype of an individual,

$$P(?|\text{all pedigree information}) = P(?|\text{neighborhood}).$$

This *Markov* property also holds for linkages between sites on individual chromosomes (a fact that is probably of growing importance with modern multipoint techniques). Let there be a sequence of k sites on a

chromosome and let s_i be a zero-one variable indicating whether site i is inherited from the maternal or paternal copy. In the absence of crossovers, all s_i would be 0, or all 1. With the possibility of crossovers, they will be a mixture of 0's and 1's, and

$$P(s_j | \text{all } s_i) = P(s_j | s_{j\pm 1}).$$

This is actually an approximation. There is a phenomenon called *interference* that causes a crossover event at one site to effect (generally reduce) the probabilities of crossovers at nearby sites. However, the effect is small, and is generally ignored in statistical modeling.

10.2. Calculating the *a posteriori* probability of the data

The now-“classic” analysis based on the independence observation was first presented by Elston and Stewart in 1971 [1]. To compare likelihoods in (10.1) we need to evaluate the probability of the observed data in the pedigree given an underlying hypothesis about the linkage. Let \tilde{Y} represent the observed data in the pedigree and let Y_i represent the observed data for individual i . Likewise define \tilde{G} to be all genotype information in the pedigree and G_i to be the genotype information for individual i . An individual is said to be a *founder* in a pedigree if his/her parents do not belong to the pedigree. If the parents are specified, the individual is a *nonfounder*. Let $m(i)$ and $f(i)$ be the mother and father of nonfounder i . Let $P_\theta(E)$ be the probability of event E conditioned on the hypothesis θ . Then

$$\begin{aligned} P_\theta(\tilde{Y}) &= \sum_{\tilde{G}} P_\theta(\tilde{G}) P_\theta(\tilde{Y} | \tilde{G}) \\ &= \sum_{\tilde{G}} \prod_{i \in \text{observed}} P_\theta(Y_i | G_i) \prod_{i \in \text{founders}} P_\theta(G_i) \prod_{i \in \text{nonfounders}} P_\theta(G_i | \tilde{G}_{m(i)}, \tilde{G}_{f(i)}) \end{aligned} \quad (10.2)$$

The sum above runs through all possible configurations of genotypes in the pedigree. In a large pedigree a brute-force evaluation of the sum would be infeasible. The key is to collect all factors involved in a peripheral *family* (mother, father and offspring). The following example illustrates this process: Consider the pedigree in Figure 10.1. Assume that the genotype “Bb” is observed on individuals 2 and 5. Then the

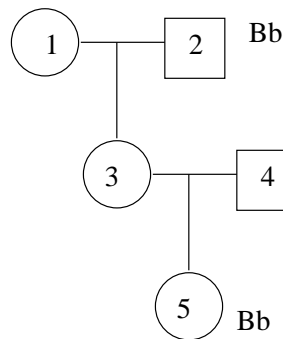


Figure 10.1: Simple pedigree example

above sum becomes

$$\begin{aligned}
 P_{\theta}(\tilde{Y}) &= \sum_{G_1 \dots G_5} P_{\theta}(Y_5 = \text{Bb} | G_5) P_{\theta}(Y_2 = \text{Bb} | G_2) P_{\theta}(G_1) P_{\theta}(G_2) P_{\theta}(G_4) P_{\theta}(G_3 | G_1, G_2) P_{\theta}(G_5 | G_3, G_4) \\
 &= \sum_{G_1 \dots G_3} P_{\theta}(Y_2 = \text{Bb} | G_2) P_{\theta}(G_1) P_{\theta}(G_2) P_{\theta}(G_3 | G_1, G_2) R_{\theta}(G_3)
 \end{aligned}$$

where R is called an r -function and is found by collecting all factors involving the genotypes of the family $\{3, 4, 5\}$ and summing over the ranges for G_4 and G_5 .

This example illustrates the simple case where there are no *loops* in the pedigree. Loops arise in pedigrees when there is inbreeding in the population. For example, if first cousins marry, a loop is created as shown in Figure 10.2. Brothers marrying sisters also causes a loop.

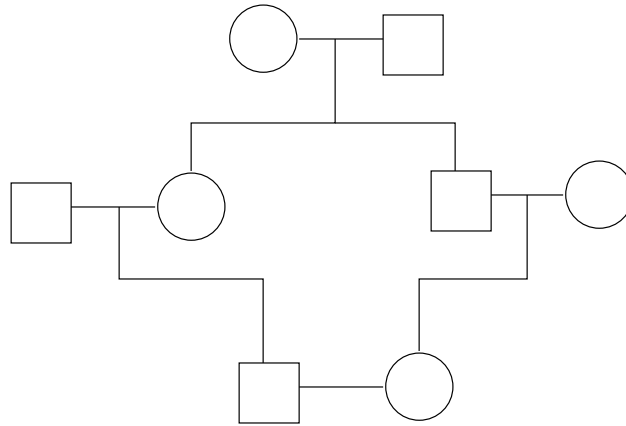


Figure 10.2: Inbreeding creates pedigree loops

The bottom line on the method introduced above is that loop-free pedigrees are computationally easy, at least when there are relatively few loci.

10.3. Complex Pedigrees

A pedigree containing loops is called a *complex pedigree*. The difficulty with complex pedigrees is that it is no longer possible to simplify the sum in (10.2) by finding peripheral families which connect to the rest of the pedigree through exactly one individual (G_3 in the example). If a family is part of a loop then it is connected through at least two individuals. In this case the r -function depends on the genotypes of these two individuals. The order in which sub-components of the pedigree are evaluated is called the *peeling sequence*. In a large pedigree with many loops finding a good peeling sequence is critical to doing the computation. Unfortunately, this problem is very similar to the *graph partitioning problem* which is known to be NP-hard [2]. Although finding an optimal peeling sequence is NP-hard on general graphs, it is not clear that the problem is NP-hard in practice. (For example, age structure of the population and other effects constrain

the graphs in various ways.) Nevertheless, all current algorithms content themselves with heuristics and/or approximations to optimal peeling sequences on “real” data.

Figure 10.3 illustrates this technique applied to a pedigree of an isolated population from the island of Tristan de Cunhan in the South Atlantic. Dashed lines indicate the cuts made in the graph to form the peeling sequence.

The bottom line on the “peeling sequence” approach to complex pedigrees seems to be that it can handle a single locus if the number of cuts is smaller than about 15, but it is hopelessly inadequate (by itself) for handling more complex and/or multi-locus pedigrees.

10.4. Monte-Carlo Methods

For larger pedigrees, it is too costly to evaluate the sum in (10.2) exactly. The speaker has developed some Monte-Carlo approaches to estimating this probability. A naive approach is to take N random samples of \tilde{G} , evaluate the resulting terms in (10.2) and average. This estimate of $P_\theta(\tilde{Y})$ can then be used to calculate the likelihood of the linkage hypothesis θ :

$$\hat{L}(\theta) = \frac{1}{N} \sum_{i=1}^N P_\theta(\tilde{Y} | \tilde{G}^{(i)}) P_\theta(\tilde{G}^{(i)})$$

Unfortunately the sample space is so large that this estimator is no good for computationally feasible values of N .

A more sophisticated approach uses the idea of *Gibbs sampling*. In point form the approach is:

- a. Start at a configuration of the genotypes which is consistent with the data (actually another problem).
- b. Select an individual j at random. Re-sample G_j given \tilde{Y} and $\tilde{G}_{\text{all else}}$. By the Markov property, the probability distribution to sample is

$$P_\theta(G_j | Y_j, \tilde{G}_{\text{neighborhood}(j)}).$$

- c. Repeat b. *ad nauseum*.
- d. “Eventually” \tilde{G} is a realization from $P_\theta(\tilde{G}, \tilde{Y})$.

An estimator of $\hat{L}(\theta)$ based on the resulting \tilde{G} sample is much better than the naive approach of using random samples of \tilde{G} . Here better means that the distribution of estimates is unbiased and has low variance. See [3] for more information.

References

- [1] R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human Heredity*, 21(6):523–542, 1971.

THE GENES OF THE TRISTAN DA CUNHANS 103

Figure 28. A marriage node graph of the major section of the ancestral genealogy of the Tristan da Cunha population sampled in 1961. L^* and L^{**} are specific sampled individuals (see text). Reprinted from Thompson 1978.

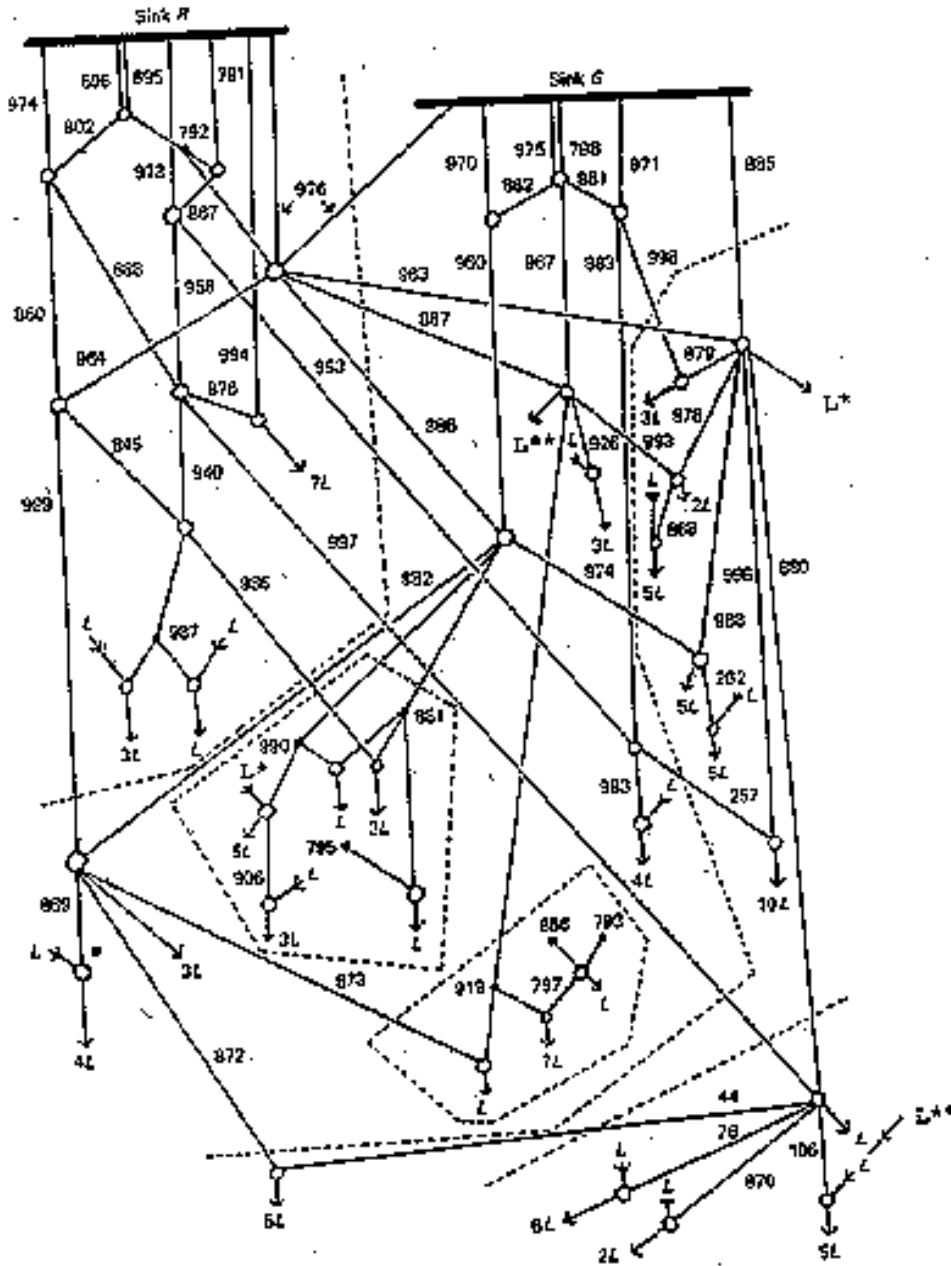


Figure 10.3: Complex pedigree example

- [2] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.
- [3] E. A. Thompson. Monte Carlo in genetic analysis. Technical Report 294, Department of Statistics, University of Washington, Sept. 1995. (PostScript version obtainable via `ftp://evolution.genetics.washington.edu/pub/thompson/Papers/genepi_mcmc_rept.95.`).