Chapter 5. Linkage Analysis

Linkage is an important tool for the mapping of genetic loci and a method for mapping disease loci. With the availability of numerous DNA markers throughout the human genome, linkage analysis has succeeded in mapping the mutations responsible for hundreds of Mendelian diseases. The current challenge is to use linkage to map the susceptibility loci for common diseases with complex genetic and environmental determinants.

Linkage analysis has two different objectives: (1) is to map the genetic distance between markers, at this case, we know the genotypes at all the markers; (2) is to map the linkage between the genetic marker or markers and the disease locus, at this case, we only have the genotypes at a marker locus or markers but not the disease locus. However, we will have some diseased individuals and some normal individuals.

1. Indices of marker information

To be useful as a marker for linkage analysis, a locus should be highly polymorphic so that alleles inherited from different sources are likely distinguishable from each other. An ideal index for marker informativeness should therefore measure not only the number of possible alleles occurring at the locus, but also the frequencies of these alleles. Two such indices are commonly used. The first index is simple the probability that a randomly selected individual from the population under random mating is heterozygous at the locus. This index, called heterozygosity and denoted by H, is defined as follow

$$H = 1 - \sum_{i=1}^{m} p_i^2$$

where p_i is the allele frequency of the ith allele at the locus, and m is the total number of the alleles at this locus. Another popular index, called polymorphism information content and denoted as PIC, is defined as

$$PIC = 1 - \sum_{i=1}^{m} p_i^2 - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} 2p_i^2 p_j^2.$$

2. Map genetic distance between markers

2.1 Linkage analysis Using Fully Informative Gametes

One objective of linkage analysis is to make inference about relative positions of two or more markers (not necessary the disease locus) based on family data. We consider the case of two markers. The linkage analysis here is to infer the recombination fraction (rate) or test the hypothesis of linkage between the two markers. Let θ denote the recombination rate between the two markers. The problem is the estimate θ or test the null hypothesis

$$H_0 : \theta = \frac{1}{2} \text{ vs } H_1 : \theta < \frac{1}{2}.$$

First, we consider some concepts:

- A haplotype is called a recombinant if it is the result of the recombination of the parental haplotypes, other wise it is called non-recombinant.
- Non-informative haplotype: You totally don't know that this haplotype is a recombinant or non-recombinant.
- Fully informative haplotype: you can unambiguous determine that the haplotype is a recombinant or non-recombinant
- Partial informative or informative: If we know that the haplotype may be a recombinant but not certain.

Example 1. Consider the following pedigree

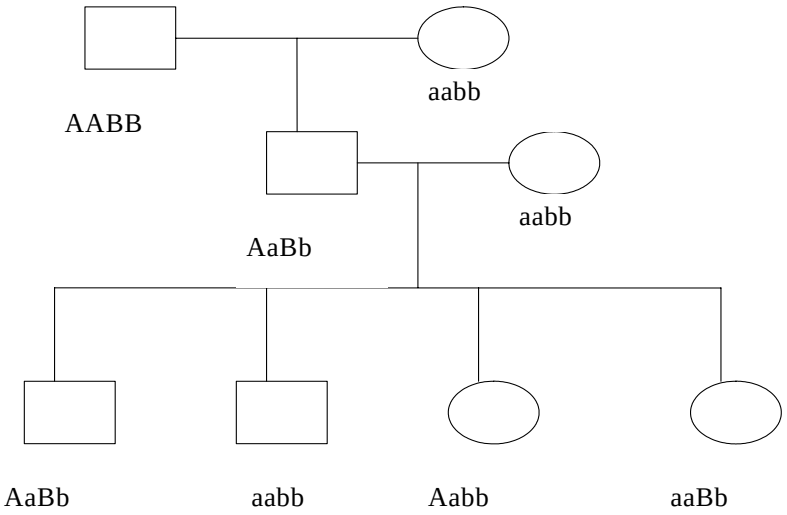

Figure 1. One three-generation pedigree.

The grandfather has two haplotypes AB/AB, grandmother has two haplotype ab/ab. The Father has two haplotypes AB/ab which are non-informative. Because we do not know the haplotype AB (ab) is a recombinant of not. The first son has one haplotype ab from mother and one haplotype AB from father. Haplotype ab is non-informative and the haplotype AB from father is fully informative, since we are certain that AB is a non-recombinant.

We consider a simple case, that is, all the informative haplotypes are fully informative as the third generation of the pedigree given in Figure 1. Let n denote the total number of the informative haplotypes; X denote the number of recombinants. Then $X \sim B(n, \theta)$, i.e.

$$P(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Then the likelihood function up to a constant is given by

$$L(\theta) = \theta^X (1-\theta)^{n-X}.$$

The MLE of θ is given by

$$\hat{\theta} = \begin{cases} \frac{X}{n} & \frac{X}{n} \leq \frac{1}{2} \\ 0.5 \, \text{otherwise.} \end{cases}$$

.

The likelihood ratio test statistics

$$G^2(X) = 2(\log(L(\hat{\theta}) - \log(L(0.5)).$$

Here, the likelihood ratio test is not a χ^2 distribution but a 50:50 mixture of a point probability at 0 and a chi-square distribution with degrees of freedom one.

[*Note on mixture distribution: Let* X, Y *and* Z *denote three random variables with density functions (or probability functions)* $f_1(x), f_2(x)$, *and* $f_3(x)$. *We say that* X *is* $p_1 : p_2 (p_{1+} p_2 = 1)$ *mixture of* Y *and* Z *if*

$$f_1(x) = p_1 f_2(x) + p_2 f_3(x).$$

So,

$$F_X(x) = P(X \leq x) = p_1 P(Y \leq x) + p_2 P(Y \leq x) = p_1 F_Y(x) + p_2 F_Z(x).$$

]

To Calculate the p-value of the test, let Y denote a one point distribution at 0 i.e $P(Y = 0) = 1$. Since G^2 is a 50:50 mixture of Y and χ^2_1, then

$$P(G^2 \leq t) = 0.5 P(Y \leq t) + 0.5 P(\chi^2_1 \leq t)$$

or

$$\begin{aligned} P(G^2 > t) &= 1 - (0.5 P(Y \leq t) + 0.5 P(\chi^2_1 \leq t)) \\ &= 0.5(1 - P(\chi^2_1 \leq t)) \\ &= 0.5 P(\chi^2_1 > t). \end{aligned}$$

Let $g^2 = G^2(x)$, then the p-value of the likelihood ratio test will be

$$\text{p-value} = P(G^2 > g^2) = 0.5 P(\chi^2_1 > g^2).$$

2.2. Linkage Analysis of General Pedigree

For an general pedigree, it is far from realistic to assume that you can know unambiguous that each informative haplotype is either a recombinant or a non-recombinant. For example,

the pedigree we considered in Figure 1 with missing grandparents. At this case, we do not know the father's two haplotypes are AB/ab or Ab/aB. So, we are not certain that the first child's haplotype AB is a recombinant or not. However, we can calculate the probability (or likelihood) given some parameters.

The two marker genotypes are usually denoted by AaBb, aabb. The the genotype with known haplotype (phase) information denoted by AB/ab, Ab/aB, ab/ab are called ordered genotypes. Sometimes, we also need the information of parental origin of the haplotype, such as AB/ab means that haplotype AB comes from father and ab comes from mother.

For simplicity of the notation, we consider first two children. Denote the genotypes of the father, mother, and the two children by $y = (\ y_f, y_m, y_1, y_2)$ respectively, the likelihood or probability of the pedigree's genotype is given by $P(y)$.

Denote the ordered genotype of the father, mother and two children by $g = (\ g_f, g_m, g_1, g_2)$. g may have many possible, for example, $g_f = AB/ab$ or Ab/aB (we consider parental origin here). If we assume H-W equilibrium and linkage equilibrium (the alleles in the two markers are statistically independent), then $P(AB/ab) = P(AB)P(ab) = p(1-p)q(1-q)$ (here we assume the allele frequency of A is p and allele frequency of B is q). Let G denote the all possible compatible combinations of the ordered genotype for the parents. Then,

$$
\begin{aligned}
& P(y) \\
= & \sum_{(g_f, g_m) \in G} P(g_f, g_m, y_1, y_2) \\
= & \sum_{(g_f, g_m) \in G} P(g_f) P(g_m) \prod_{i=1}^{2} P(y_i | g_f, g_m) \\
= & P(AB/ab) P(ab/ab) \prod_{i=1}^{2} P(y_i | g_f = AB/ab, g_m = ab/ab) \\
& + P(Ab/aB) P(ab/ab) \prod_{i=1}^{2} P(y_i | Ab/aB, ab/ab) \\
= & pq(1-p)^3 1-q)^2 (\prod_{i=1}^{2} p(y_i | AB/ab, ab/ab) + \prod_{i=1}^{2} p(y_i | Ab/aB, ab/ab)
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
& P(y_1 | AB/ab, ab, ab) \\
= & P(AaBb | AB/ab, ab/ab) \\
= & (1-\theta)/2
\end{aligned}
$$

Similarly

$$\begin{aligned} & P(y_2|AB/ab, ab, ab) \\ =& P(aabb|AB/ab, ab/ab) \\ =& (1-\theta)/2 \end{aligned}$$

$$\begin{aligned} & P(y_1|Ab/aB, ab, ab) \\ =& P(AaBb|AB/ab, ab/ab) \\ =& \theta/2 \end{aligned}$$

$$\begin{aligned} & P(y_2|AB/ab, ab, ab) \\ =& P(aabb|Ab/aB, ab/ab) \\ =& \theta/2 \end{aligned}$$

So, the likelihood contribution of this pedigree is

$$p(y) = \frac{1}{4}pq(1-p)^3 1-q)^2[\theta^2 + (1-\theta)^2]$$

and log-likelihood contribution of this pedigree, up to a constant, is

$$\begin{aligned} \log(\theta, p, q) =& \log P(y) \\ =& \log[pq(1-p)^3(1-q)^3] + \log[\theta^2 + (1-\theta)^2] \end{aligned}$$

If there are many pedigrees, noting that the pedigrees are independent, the total log-likelihood is the sum of all the pedigree's log-likelihood contributions.

For a general pedigree with f founders and $n-f$ non-founders, let the first f member of the pedigree denote founders, and $y = (y_1, \ldots, y_n)$ and $g = (g_1, g_2, \ldots, g_n)$ denote the genotype and ordered genotype of all members of the pedigree, respectively. Then, the likelihood contribution of this pedigree is given by

$$\begin{aligned} P(y) =& \sum_g P(g_1, \ldots, g_f, g_{f+1}, \ldots, g_n) \\ =& \sum_g \prod_{i=1}^{f} p(g_i) \prod_{j=f+1}^{n} p(g_j|g_{j,f}, g_{j,m}) \end{aligned}$$

where $g_{j,f}$ and $g_{j,m}$ are the ordered genotypes of ith individual's father and mother, respectively. As an example, consider the pedigree in Figure 2.

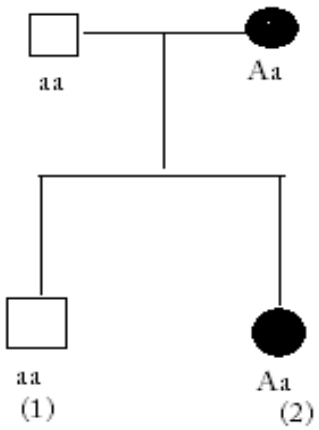

Figure 2.

The mother and daughter are affected. Suppose that this is a rare Mendelian dominant disease decided by a biallelic locus with alleles D and d. So, we can assume the genotypes of father, mother, son and daughter at disease locus are dd, Dd, dd and Dd, respectively. Now, denote the recombination rate between the marker and the disease locus by θ. To test if the marker has linkage with the disease locus or not is equivalent to test null hypothesis $H_0 : \theta = 0.5 \; vs \; H_1 : \theta < 0.5$.

To construct the likelihood test, we need to calculate the likelihood function: (Let p and q denote the allele frequencies of A and D)

$$P(Y) = \sum_g P(g_f)P(g_m)P(g_1|g_f, y_m)P(g_2|g_f, g_m)$$

where $g = (g_f, g_m, g_1, g_2) = (ad/ad, AD/ad, ad/ad, AD/ad)$ or $(ad/ad, Ad/aD, ad/ad, AD/ad)$. Fro each (g_f, g_m), we have

$$P(g_f)P(g_m) = 4pg(1-p)^3(1-q)^3$$

and

$$\begin{aligned} P(ad/ad|ad/ad, AD/ad) &= (1-\theta)/2 \\ P(AD/ad|ad/ad, AD/ad) &= (1-\theta)/2 \\ P(ad/ad|ad/ad, Ad/aD) &= \theta/2 \\ P(AD/ad|ad/ad, Ad/aD) &= \theta/2. \end{aligned}$$

So, likelihood function of this pedigree is

$$P(Y) = pg(1-p)^3(1-q)^3[(1-\theta)^2 + \theta^2].$$