1 Extensions of the basic model for parametric linkage

Our model regarding the relationship between genotypes and phenotypes thus far has been overly simplistic. We have assumed that each genotype translates into one definite phenotype, and while this is a useful model for studying some genetic disorders, it is not realistic for many other disorders. A good example of where this simple model breaks down is with breast cancer. At least one gene has been found that seems to be associated with developing breast cancer. The relationship here is complicated since not everyone who has this gene develops breast cancer by a certain age, hence we speak of *susceptibility* genes rather than disease genes. Taking this perspective greatly extends the applicability of finding genes associated with disorders. Another possibility is that a "disorder" is actually the norm, hence someone will have the disorder unless the individual has a specific genotype.

In addition, we have assumed all genetic disorders fit into the Mendelian framework of a single disease gene (that is either dominant or recessive), but there is increasing evidence that other modes of transmission provide better models for inheritance. A trait whose mode of genetic transmission does not fit into the Mendelian framework is referred to as a *complex trait*. Additionally, it is possible that genetic transmission of a trait is well modeled by the Mendelian framework at more than one locus, and different affected individuals have genetic anomalies at different loci.

By moving outside of the Mendelian framework, a vast array of models for disease etiology is possible. We caution that while this makes the field exciting, one must be very careful. All we observe is phenotypes (but since we use codominant markers, and we assume no errors in genotyping, we can unequivocally determine marker genotypes). If we relax the manner in which genotypes of parents are transformed into genotypes of offspring and the manner in which genotypes are transformed into phenotypes we run the risk of using nonidentifiable models (since we are proposing 2 models for unobservable phenomena). This risk is very serious because computing the likelihood for pedigree data is not transparent, hence it is difficult to judge if a given model is identifiable. The practice of graphing the likelihood ratio as a function of the recombination fraction is due to the recognition that the likelihood for pedigree data is potentially multimodal and very complicated. This implies that a nonlinear maximization routine may converge to a local mode when in fact the likelihood is unbounded or no unique MLE exists (because the model is nonidentifiable).

1.1 Penetrance

Penetrance is defined as the probability one has a certain phenotype given the genotype. Thus far, we have assumed penetrance is either zero or one. In general, we can allow the penetrance to be a parameter and maximize the likelihood with respect to the recombination fraction and the penetrance simultaneously. One problem with this approach is *ascertainment bias*, that is, the families we have in a typical genetic study are selected because some member of the family has the disease and the genotype. We miss individuals who have the genotype but not the phenotype, hence other methods for estimating penetrance have been devised. Clearly, ignoring the ascertainment bias will result in an overestimate of the penetrance.

If we are not interested in the recombination fraction, and we suppose the disease is a rare, dominant disorder, then we can estimate the penetrance with $2\hat{p}$, where $\hat{p}$ is the proportion of affecteds in the family. For a rare, recessive trait we would use $4\hat{p}$. As above, this estimate will be biased due to the ascertainment bias. *Weinberg's proband method* (1927) attempts to correct for ascertainment bias by leaving the proband out of the calculation, that is, if k_i represents the number of affected individuals in the i^{th} kinship (of size n_i), then

$$\hat{p} = \frac{\sum_i (k_i - 1)}{\sum_i (n_i - 1)}.$$

Some disorders only develop with age, hence we speak of *age dependent penetrance*. Instead of just estimating a single parameter representing the penetrance, we parameterize the probability of having a certain phenotype by age x given a certain genotype. For example, if X represents the age at which an individual with a certain genotype develops some disease, we might use the logistic curve to model the penetrance for a given genotype

$$P(X \leq t|\mu,\sigma) = \frac{1}{1+\exp\{-(x-\mu)/\sigma\}},$$

and estimate the parameters using maximum likelihood. In an analogous fashion, we can allow the penetrance to depend on any characteristic we desire.

1.2 Phenocopies

A *phenocopy* is an individual who displays a disease phenotype but does not have the disease genotype. It is often reasonable to suppose that a certain trait can have environmental sources and genetic sources, and some individuals have the disorder due to environmental sources alone. The *phenocopy rate* is defined as the proportion of phenocopies among affected individuals. Clearly a high phenocopy rate will make it difficult to establish linkage. The best means for getting around this issue is careful definition of the phenotype. Many disorders have been defined with reference to clinical criteria (since medical management is the usual objective), but clinical criteria often lump cases into categories which may not be useful from the perspective of the genetic etiology of the disorder. For example, obesity is a harmful condition, and from the perspective of clinical management, it may not make much of a difference as to the source of the obesity. Yet there seem to be genetic cases of obesity, and if the distinction between genetic and non-genetic cases is not made it will be very difficult to find any obesity susceptibility genes. The solution is to try to use a phenotype definition that distinguishes between the genetic and non-genetic cases, such as some aspect of the metabolism, rather than body mass.

2 Heterogeneity in the recombination fraction

It is well documented that there is heterogeneity in the recombination fraction. For example, there are more crossovers in women than in men, hence we should have a recombination fraction for men and women. In *Drosophilia* the males show no recombination at all. Moreover, there is heterogeneity in the recombination fraction between families. Since normal physiology is the outcome of many inter-related biological pathways, there are many opportunities for something to go wrong. For this reason, distinct genetic defects can have the same outcome in terms of physiology, that is, distinct genotypes can lead to the same disease phenotype. Although there are a number of examples, a good example was provided by a case in which 2 parents with a recessive form of albinism had 4 offspring without albinism. One explanation for this is the parents had the albinism gene at different loci (mutation is unlikely with 4 offspring, but germ-line mosaicism could provide another explanation). Often this sort of heterogeneity is referred to as *locus heterogeneity* in contrast to *allelic heterogeneity*. Allelic heterogeneity is the situation in which more than one allele at the same locus can lead to the disorder. Cystic fibrosis is an example of a disorder characterized by allelic heterogeneity.

2.1 Testing for linkage when there is heterogeneity

If heterogeneity is ignored, it will be difficult to establish linkage, or it may appear that several genes are involved in the disorder (leading to the incorrect conclusion that the disorder is complex). For this

reason, a number of statistical tests have been devised to test for heterogeneity and linkage in the presence of heterogeneity. With large families, a sensible first step is to conduct a linkage analysis for each family separately and all families together. Additionally, one should examine all the phenotypic data to see if heterogeneity is present with regard to the phenotype.

2.1.1 Morton's test

Suppose the families can be grouped into c classes based on the phenotype data. Alternatively, there may be c families. Morton proposed to test the null hypothesis

$$H_0 : \theta_1 = \cdots = \theta_c < \frac{1}{2}$$

against the alternative

$$H_1 : \theta_1 \neq \cdots \neq \theta_c.$$

If $\hat{\theta}_i$ is the MLE for the i^{th} class, $\hat{\theta}$ the MLE ignoring the classes, Z_i is the log 10 likelihood ratio for the i^{th} class, and Z is log 10 likelihood ratio ignoring the classes, then Morton proposed the following maximum likelihood ratio test

$$X^2 = 2\log(10)\Big[\sum_i Z_i(\hat{\theta}_i) - Z(\hat{\theta})\Big].$$

Under H_0, X^2 is asymptotically χ^2 with $c-1$ degrees of freedom (since we can obtain H_1 from H_0 by imposing $c-1$ linear constraints). It is possible to use this test and reject homogeneity, yet fail to establish linkage (especially since conventional significance levels are often used when conducting this test), hence one suggestion is to use a significance level for this test of 0.0001.

2.1.2 The test for admixture

Other methods don't require one to break up the families into classes, or suppose the families are all in their own class. The basic idea is to suppose there is some distribution from which each family has a draw that determines the recombination fraction for that family, that is, we use a *hierarchical model*. The simplest example is to suppose there are 2 family types: those with linkage of the disease trait to a marker and those without linkage. In this case, the likelihood for the i^{th} family can be expressed

$$L_i(x|\alpha,\theta) = \alpha L_i(x|\theta) + (1-\alpha)L_i(x|\theta = 0.5).$$

We then test the hypothesis $H_0 : \alpha = 1$ against the alternative $H_1 : \alpha < 1, \theta < 1/2$. If we define the test statistic as

$$X^2 = 2[\log L(\hat{\alpha},\hat{\theta}) - \log L(1,\hat{\theta})],$$

then X^2 has a distribution under the null hypothesis that is a χ^2 variable with one degree of freedom with probability 0.5, and is zero with probability with 0.5. To get a p-value, look up the value for X^2 in a χ^2 table (with 1 degree of freedom), then divide that probability by 2. There are other tests for admixture based on more sophisticated hierarchical models. One can incorporate a variety of factors into such models, such as age of onset.

2.1.3 The hlod

Rather than testing for heterogeneity, we may want to test for linkage given there is heterogeneity. In the framework of the test for admixture, we are now interested in the test statistic

$$X^2 = 2[\log L(\hat{\alpha}, \hat{\theta}) - \log L(\hat{\alpha}, 0.5)].$$

The distribution of this test statistic is well approximated as the maximum of 2 independent χ^2 variables with 1 degree of freedom. If we take the log to the base 10 of the maximized likelihood ratio, then we get the *hlod*.

3 Quantitative traits

Often phenotypic data is in the form of continuous measurements, hence it makes sense to think of *quantitative traits*. Since we conceptualize genes as discrete entities, it is not clear how we could have quantitative traits arising from a single gene. Typically, geneticists think of quantitative traits as arising from a single major gene (the *quantitative trait locus* or QTL) with minor contributions from many less important genes. The goal is then to find this single major gene. Alternatively, we can think of the quantitative trait as due to Mendelian inheritance and our measurement includes some measurement error or the effect of incomplete penetrance. In practice we could never distinguish between a trait being a quantitative trait or a Mendelian trait subject to incomplete penetrance unless we were able to localize these other, less important genes (and this is usually considered impossible because each has such a small effect).

The basic idea behind most contemporary methods for gene localization that assume the existence of a quantitative trait is that family members who have a similar set of genes influencing the quantitative trait will have similar phenotypes. Hence we should examine how the correlation between values of the quantitative trait depends on allele sharing in a family. If individuals with similar values for the quantitative trait share a set of alleles on some contiguous section of the genome, then the QTL(s) must lie in this (these) shared region(s). Typically multivariate normal models are used to model the distribution of values for the quantitative trait (each QTL acts as a random effect). In these models, the covariance between the measurements for 2 individuals is a function of how close the individuals are in the family, and the variance of the QTLs. One then tests for the effect of a QTL by testing the hypothesis that the variance associated with each QTL is zero.

4 Multipoint parametric linkage analysis

We saw that the objective of 2 point linkage analysis was to localize a disease gene. By computing the genetic distance between many markers and genes we could construct a genetic map. The objective of multipoint linkage analysis is the same as 2 point linkage analysis, but since we use more than one marker, we can compute recombination rates more precisely (and order markers more precisely). One account (Lathrop (1985)) found the efficiency went up by a factor of 5 by conducting 3 point linkage analysis as opposed to 2 point linkage analysis.

4.1 Quantifying linkage

When there is more than one marker, there is more than one way in common use to quantify linkage of a disease locus to the set of markers. We briefly outline these here. All of these measures are log base 10

likelihood ratios under 2 different models (in analogy with the lod score).

1. **Global support**

   Here we compare the likelihoods when a certain locus is in the map to when it is not on the map. We say there is linkage with the map when the global support exceeds 3.

2. **Interval support**

   Here we compare the likelihood when a locus is in a certain interval to the likelihood when it is in any other interval.

3. **Support for a given order**

   Here we compare the likelihood given a certain order to the most likely order.

4. **Generalized lod score**

   Here we compare the likelihood given a certain order to the likelihood we get assuming no linkage. For example, if there are 3 loci, we compute

   $$\frac{L(\theta_{12}, \theta_{23})}{L(0.5, 0.5)}.$$

   This measure doesn't make sense when you are mapping a loci onto a known map of markers (because then the denominator doesn't make sense).

5. **Map specific multipoint lod score**

   Here we express the ratio of the likelihood as a function of locus location to the likelihood when the locus is off the map. A related quantity is the *location score*. This measure uses the usual 2 times the natural logarithm of this likelihood ratio.

4.2 Interference

For markers that are close, it is unlikely that they will be separated by 2 crossovers, hence the recombination fraction is just the probability of a crossover. This implies genetic distance is directly related to physical distance. As the distance grows, the genetic distance departs from the physical distance since genetic distance is between 0 and 0.5. To understand the nature of the problem, realize that if one typed 3 markers on a single chromosome and assumed that genetic distances add in the same manner that physical distance on a line is additive, then one could conclude that the recombination fraction between 2 of the markers exceeds 0.5.

Consider 3 loci, A, B, and C, and denote the recombination fractions as θ_{AB}, θ_{AC} and θ_{BC}. Suppose the order of these loci along the chromosome is given by A, B, C. Then

$$\theta_{AC} = \theta_{AB} + \theta_{BC} - 2\gamma\theta_{AB}\theta_{BC},$$

where

$$\gamma = P(BC|AB)/\theta_{BC},$$

if $P(BC|AB)$ means the probability of recombination between B and C given there is recombination between A and B. Here γ is referred to as the *coefficient of coincidence*. A related quantity, known as the *interference* is defined as $I = 1 - \gamma$. We expect γ to be less than one due to interference. One can estimate γ by

using the estimated recombination fractions, but usually estimates are very imprecise (requiring hundreds or thousands of meioses). It is typically assumed that γ does not depend on genomic location, but only the physical distances between the loci involved. If γ is known, then we can relate physical distance to recombination probabilities.

A method due to Haldane for understanding the relationship between map distance and recombination fractions supposes the relationship between map distance is given by the identity map for small distances. First, given the above definition of γ, we can write

$$\gamma = \frac{1}{2}\frac{\theta_{AB} + \theta_{BC} - \theta_{AC}}{\theta_{AB}\theta_{BC}}.$$

Now suppose the physical distance between A and B is x, and suppose C is close to B, and of distance Δx. Use $\theta = M(x)$ to denote the mapping of physical distances into genetic distances, then, if $M(0) = 0$, and we take Δx to zero,

$$\gamma = \frac{1}{2M}\Big(1 - \frac{M'(x)}{M'(0)}\Big).$$

If the map is linear with slope one for small x, i.e./ $M'(0) = 1$, then we obtain the differential equation

$$\gamma_0 = \frac{1}{2M}\Big(1 - M'(x)\Big).$$

We use γ_0 here to recognize the role of the assumptions $M(0) = 0$ and $M'(0) = 1$, and we call γ_0 *Haldane's marginal coincidence coefficient*. Hence we have

$$\frac{dM}{dx} = 1 - 2M\gamma_0.$$

Since $M(0) = 0$ we can solve this equation once we specify γ_0. For example, if $\gamma_0 = 1$, $M(x) = (1 - \mathrm{e}^{-2x})/2$, a map function known as *Haldane's map function*. The assumption $\gamma = 1$ is interpreted as $P(BC|AB) = P(BC)$, i.e. independence of crossovers (a case known as no interference). (The Lander-Green algorithm assumes no interference.) If $\gamma_0 = 0$ then there are no crossovers close to a known crossovers, a case known as complete interference (see Exercise 2 for more on this case). If $\gamma_0 = 2\theta$, then $M(x) = \tanh(2x)/2$, which is *Kosambi's map function*. Many other map functions in the literature can be viewed as arising in this way, and since we can interpret γ_0, we can interpret the meaning of a map function. Given a map function, one can compute recombination fractions between 2 loci A and C given the recombination fractions between each locus and some intermediate locus. The resulting formulas are known as *addition rules*, and depend on the form of the map function.

One problem with map functions generated by Haldane's method is they do not necessarily give rise to valid probabilities for observing gametes, that is, they are not necessarily *multi-locus feasible*. Liberman and Karlin (1984) show that a map is multi-locus feasible if

$$(-1)^k M^{(k)}(x) \leq 0,$$

for all $k \geq 1$ and all $x \geq 0$. For example, Kosambi's map is not multi-locus feasible. A way to construct maps that are multi-locus feasible is to fix the maximum number of crossovers, but let the locations be random. If $f(s)$ is the probability generating function for the number of crossovers, i.e.

$$f(s) = \sum_k p_k s^k,$$

where p_k is the probability of k crossovers, then, if μ denotes the mean number of crossovers, we define the map

$$M(x) = \left[1 - f\left(1 - \frac{2x}{\mu}\right)\right] \Big/ 2.$$

As an example, suppose the number of crossovers is distributed according to the Binomial distribution with parameters n and p. Then

$$f(s) = (ps + 1 - p)^n,$$

so

$$f\left(1 - \frac{2x}{\mu}\right) = \left(1 - \frac{2x}{n}\right)^n,$$

hence

$$M(x) = \left[1 - \left(1 - \frac{2x}{n}\right)^n\right] \Big/ 2.$$

If $n = 1$ we find $M(x) = x$, while if we take the limit in n we obtain Haldane's map.

Exercises

1. Consider a family with 2 parents and 1 child. Suppose one parent and the child have a rare, dominant genetic disorder. We genotype the family at one marker locus and find the affected parent has genotype (1,2), while the other parent and the child have genotype (2,2). Compute the likelihood assuming incomplete penetrance as a function of the recombination fraction, genotype frequencies and the penetrance.

2. Morgan's map function can be found by supposing $\gamma_0 = 0$. Use Haldane's method to find the map, $M(x)$, to map physical distance into genetic distance for this marginal coincidence coefficient.

3. Derive the addition formula for the Kosambi map function.

References

Lathrop, G., Lalouel, J., Julier, C., Ott, J. (1985), "Multilocus linkage analysis in humans: Detection of linkage and estimation of recombination", *Am J Hum Genet*, 37:482–498.

Liberman, U., Karlin, S. (1984), "Theoretical models of genetic map functions", *Theor Popul Biol*, 25:331–346.