1 Mapping genomes

There have been tremendous strides in the mapping of genomes, the most notable being the recent completion of the first stage of the human genome project. The completion of these projects allows for isolation and manipulation of individual genes, and by cataloguing all genes in an organism we can start to take new approaches to answering some of the mysteries surrounding genomes, such as the function of non-coding DNA and the co-regulation of expression patterns of genes. Additionally, if we know the location of disease genes, we can advise potential parents with regard to risk for disease or attempt to develop treatments based on knowledge of the gene products. Knowledge of the location of the genes along the genome is knowledge of the *physical map* of the organism. A *genetic map* relates how close 2 loci on a genome are in terms of risk for a certain trait based on knowledge of the allele at the other locus. Either sort of map provides information about the locations of genes.

For small genomes the *shotgun method* has provided a means of determining the sequence of a genome. To understand the motivation for this approach, one must appreciate that a sequence longer than 750 base pairs (bp) can't be sequenced in a single experiment, yet no organisms have such short sequences. To overcome this, we can break up the sequence into smaller, overlapping segments, sequence these, then put it all back together-this is the shotgun method. Unfortunately, for complex organisms, putting all the segments back together correctly is very difficult since there is a lot of repetitive DNA and the sequences are very long (the human genome is about 3 billion bp, or 3,000 Mb). One approach is to break up the genome into *known* short segments that can be sequenced, sequence these (using the shotgun method), then put all these *known* short segments back together (this is the basis of the *clone contig* approach). Another approach uses the basic idea of the shotgun approach, but uses the known locations a certain number of genes to help us put the sequence back together (this process is known as the *directed shotgun*). For this to be a viable method, one needs to know the location of some genes to serve as landmarks. We call these landmarks *markers*, a term we also use to refer to any known feature of a genome. There are 2 broad categories for methods aimed at finding the location of genes: *genetic mapping* and *physical mapping*.

1.1 Types of markers

At first, genes were the only known markers. Genes are suboptimal markers for mapping genomes because they are often too spread out over the genome (for complex organisms, like mammals and flowering plants). In addition, it is often hard to distinguish between the different alleles based on the phenotype. For this reason, other markers have been identified.

1.1.1 Restriction fragment length polymorphisms (RFLPs)

When DNA is treated with a *restriction endonuclease* the DNA segment is cut in specific places. If there is variation between individuals in the locations of the sites where the cutting takes place (i.e. polymorphisms in certain *restriction sites*), then DNA from different individuals will be broken up into different numbers of fragments. There are approximately 10^5 RFLPs in the human genome, and each has 2 alleles (either a cut takes place there or it doesn't). Since these markers only have 2 alleles they are often not useful for genetic mapping since everyone may have the same allele at some locus in a given study.

1.1.2 Simple sequence length polymorphisms (SSLPs)

Often there are repeated sequences of a certain unit (e.g. CACACA), and these can be polymorphic (moreover, there are often multi-allelic forms for these markers). The polymorphism in these markers lies in the number of units which are repeated. There are 2 basic sorts: mini-satellites (or variable number of tandem repeats) and micro-satellites. The difference lies in the length of the repeat unit, with micro-satellites having the shorter unit (usually just 2, 3 or 4 nucleotide units). Micro-satellites are preferred because they are more evenly distributed throughout the genome and the genotyping is faster and more accurate.

1.1.3 Single nucleotide polymorphisms (SNPs)

There are many (over 200,000) locations on the genome in which there is a polymorphism at a single nucleotide. These locations are known as SNPs. While these have the same weakness as RLFPs, namely there are only 2 alleles, there are advantages to the use of these markers. First is the huge number of them, but more importantly, the genotyping can be automated (in contrast to the other 2 types of markers).

2 Physical mapping of genomes

Many methods have been proposed, but there are 3 basic varieties: 1.) restriction mapping 2.) fluorescent *in situ* hybridization (FISH), and 3.) sequence tagged site (STS) mapping. We can't treat all of these so we give the idea behind each, then look at the details of one method, namely, radiation hybrid mapping. This method falls under the category of STS mapping.

2.1 Restriction mapping

This is a method for finding RFLPs. The simplest example is to first digest a DNA molecule with one restriction enzyme, then digest the molecule with another enzyme, then finally digest the molecule with a both enzymes at the same time. After each digest, one separates the fragments using gel electrophoresis to determine the length of the products after the digest takes place. One can resolve any uncertainties by conducting a partial restriction, i.e., stop the restriction before it is complete. As simple examples illustrate, this analysis is very difficult if there are many restriction sites, so this method has limited usefulness (this method has been used to map virus genomes). Progress has been made by using restriction enzymes which cleave the DNA molecule at only a few sites (so called *rare cutters*). The computational problem of reassembling the sequence of the genome from DNA that has been digested by 2 restriction enzymes is known as the *double digest problem* (see Waterman (1995) for a full treatment).

2.2 FISH mapping

This method is based on using a fluorescent dye to label a sequence of DNA, then creating conditions whereby the labeled sequence will bind to its complementary version in the genome under study. By using multiple probes with different dyes, one can determine the relative position of a number of DNA sequences. Originally, one would measure under the microscope how far the fluorescent label is from the end of the short arm of a chromosome (with the chromosome in metaphase). This procedure gave very poor resolution (although one could at least tell what chromosome a segment was on), but recent variations have improved resolution (by "stretching the chromosomes out", not using metaphase chromosomes or using purified DNA). Besides difficulties with resolution, FISH can only locate a few markers at once.

2.3 STS mapping

STS mapping is the most powerful physical mapping procedure. A *sequence tagged site* is a short segment of DNA (100 to 500 bp in length) that only occurs once in a chromosome or genome and is easily identified. In this procedure, one first fragments a section of DNA into a set of overlapping fragments. Given a set of fragments, one then determines how close 2 STSs are by noting how often they are on the same fragment. SSLPs are often used as STSs as are random sequences of DNA and expressed sequence tags (ESTs). An EST is a segment of DNA which codes for a protein. There are 2 basic methods for producing the collection of DNA fragments: radiation hybrids and clone libraries. Below we discuss radiation hybrids in detail. A *clone library* is a collection DNA fragments which have been put on a vector (i.e. a living host, like bacteriophage λ, or an artificial chromosomes) and stored. An advantage of using clone libraries is that one can determine which clones share an STS, and in this way one can assemble the overlapping clones (the result is a *clone contig*).

3 An example of statistical analysis in physical mapping: Radiation hybrid mapping

3.1 Experimental technique

The basic idea of radiation hybrid mapping is to bind human chromosome fragments into a rodent (typically hamster) genome, then determine what markers are present in the same rodent cells. If 2 markers are present in the same rodent cell, they are most likely close on the human chromosome. In more detail, one starts with human cells and subjects them to X-ray radiation. This causes the chromosomes to break apart (3000-8000 rads of radiation result in fragments of length 5-10 Mb). Then one stimulates the cells containing the fragmented human DNA to fuse with hamster cells. Since not all of the hamster cells take up the human DNA we get rid of these by growing the fused cells in a medium which selectively kills the cells without human DNA. In this way we can generate a whole genome radiation hybrid panel. Originally the method was used to map just a single chromosome at a time, but the details of that method imply the work required to map a whole genome is about the same as the work necessary to map a single chromosome, so the whole-genome method is what is used currently. Standard whole-genome radiation hybrid panels are now available. These are used all over the world and aid in the coordination of genomes projects.

3.2 The Poisson process as a model for breaks

Consider a segment of DNA of length t as a collection of a large number, say N, of individual intervals over which a break can take place (there is not a break at $t = 0$). Furthermore, suppose the probability of a break in each interval is the same and is very small, say $p(N, t)$, and breaks occur in the intervals independently of one another. Here, we need p to be a function of N and t because if there are more locations (i.e. N is increased for fixed t or vice versa) we need to make p smaller if we want to have the same model (since there are more opportunities for breaks). Since the number of breaks is a cumulative sum of independent and identically distributed Bernoulli variables, the distribution of the number of breaks up until time t, $X(t)$, is Binomially distributed with parameters $(N, p(N, t))$, that is

$$P(X(t) = k) = \binom{N}{k} p(N, t)^k (1 - p(N, t))^{N-k}.$$

If we suppose $p(N,t) = \lambda t/N$, for some λ, and we let N become arbitrarily large, then we find

$$P(X(t) = k) = (\lambda t)^k \mathrm{e}^{-\lambda t}/k!,$$

that is, the number of breaks up to and including time t is distributed according to the Poisson distribution with parameter λt. Here $X(t)$ represents a collection of random variables depending on t (a parameterized collection of random variables is known as a *stochastic process*). This stochastic process is known as the Poisson process. A very important aspect of the Poisson process is that it has the *Markov property*. A stochastic process (where the parameter is unidimensional, like distance along a chromosome) has the Markov property if given the present, the past doesn't help us predict the future. Since we think of the Poisson process as a cumulative sum of many *independent* Bernoulli trials, the Markov property clearly holds for Poisson processes.

3.3 Data from a radiation hybrid panel

The data for producing an estimate of locus order or marker location using radiation hybrid methods is an m-vector of zeros and ones. In these data vectors, a one indicates the human marker is present and a zero indicates otherwise. We will use x_i to denote the data vector for subject i and x to denote the collection of these vectors. There is often missing data due to ambiguities in the typing (but our introductory treatment will ignore this).

3.4 Minimum number of obligate breaks

One criterion used to estimate the order of markers along a chromosome is to use the *minimum number of obligate breaks* method. To understand the method, suppose we observe a data vector like (0,0,1,1,0,0). There are a variety of breaking and retention patterns that could give rise to such data. For example, there could be a break between every marker (5 breaks) and sometimes the human marker is retained whereas other times the rodent DNA is retained. Another explanation is there are only two breaks: one between markers 2 and 3, and one between markers 4 and 5. If as low of doses of radiation are administered as is possible to break up the DNA, we would expect that the true number of breaks is as small as the data would allow. In the previous example, we would think the scenario involving 2 breaks is more likely than the scenario involving 5 breaks. Given a data vector for subject i with element $x_{i,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, we can compute the number of obligate breaks for this subject by counting how many times $x_{i,j} \neq x_{i,j+1}$. If 1_A is an indicator variable for the event A, then the number of obligate breaks for the i^{th} hybrid is $B_i = \sum_j 1_{\{x_{i,j} \neq x_{i,j+1}\}}$. Here we have computed the number of obligate breaks using the ordering of the markers in some arbitrary order. To find the most likely order, we should compute this variable for each subject for every possible permutation of the markers. We denote permutations with σ. In the previous example, (0,0,1,1,0,0), if we use the permutation $\sigma = (1, 2, 5, 6, 3, 4)$ then we get the permuted data (0,0,0,0,1,1), and this vector has only one obligate break.

The idea of the method is to find the permutation, σ, which makes the average number of obligate breaks $\frac{1}{n} \sum B_i(\sigma)$ as small as possible. For just a few markers, one can compute this average for every permutation and select the permutation giving the smallest average. Optimizing a function over a set of permutations is a well studied problem, and is of the type generally referred to as the "traveling salesman problem". While exact solutions to these sorts of problems can be obtained through enumeration, this is often takes too much time. As an alternative, approximate solutions can be obtained using randomized search procedures, such as simulated annealing (see Press et al. (1992)).

3.4.1 Consistency of the order

This method provides a consistent method for ordering markers. To see this, first note that by the law of large numbers, our estimate converges to $\mathrm{E}B_1(\sigma)$. Assume that the identity permutation $I = (1, 2, \ldots, m)$ is the true order of the markers (so our numerical ordering is the true order of the markers). Since the parameter space is finite (the set of permutations of a number of integers) if we just show

$$\mathrm{E}B_1(\sigma) \geq \mathrm{E}B_1(I),$$

then we will know our estimate is consistent. Since $B_1(\sigma) = \sum_j 1_{\{x_{i,\sigma(j)} \neq x_{i,\sigma(j+1)}\}}$, $\mathrm{E}B_1(\sigma) = \sum_j P\{x_{1,\sigma(j)} \neq x_{1,\sigma(j+1)}\}$. To see why the identity permutation minimizes this expression consider the permutation $\sigma = (1, 2, 4, 3, 5)$. For this permutation

$$\mathrm{E}B_1(\sigma) - \mathrm{E}B_1(I) = P(x_{1,2} \neq x_{1,4}) + P(x_{1,3} \neq x_{1,5}) - P(x_{1,2} \neq x_{1,3}) - P(x_{1,4} \neq x_{1,5}).$$

Suppose that every fragment is retained in the hybrid cells with probability ρ and there is a break between the i^{th} and $(i+1)^{\text{th}}$ loci with probability θ_i. Then $P(x_{1,2} \neq x_{1,4}) = 2\rho(1-\rho)[\theta_2 + \theta_3] - 2\rho(1-\rho)\theta_2\theta_3$ whereas $P(x_{1,2} \neq x_{1,3}) = 2\rho(1-\rho)\theta_2$, hence $P(x_{1,2} \neq x_{1,4}) \geq P(x_{1,2} \neq x_{1,3})$ (since $\theta_2 \leq 1$). Clearly the same sort of reasoning will apply to the other difference for our example permutation. The basic idea expressed with this example is that if the true ordering follows our numerical ordering, then the expected value of $B_1(\sigma)$ for any permutation that deviates from the identity will entail the probability of an event like $x_{1,j} \neq x_{1,k}$ for $j - k > 1$. Such events have higher probability than $x_{1,j} \neq x_{1,j+1}$ since there is more than one opportunity for a break to occur. These higher probabilities make the expected value higher than it is for the identity permutation.

3.5 Maximum likelihood and Bayesian Methods

The method presented above has two shortcomings. First, we do not obtain estimates of the distances between the the loci. This is an important shortcoming from the perspective of building maps of genomes or estimating risks. Secondly, there is no way to assess the strength of evidence in favor of a given order. Maximum likelihood provides a way to overcome the first, and Bayesian methods can most easily overcome the second (and the first).

It suffices to consider how to compute the likelihood for a single hybrid (since all observations are assumed independently and identically distributed). For this, we will take advantage of the Markov property for Poisson processes. As mentioned above, we think of the breaks as occurring along the length of a chromosome according to a Poisson process. But this implies we can write the likelihood for a single observation as

$$L(x_i; \theta, \rho) = L(x_{i,1}) \prod_j L(x_{i,j+1} | x_{i,j}),$$

see Exercise 2 below. Now $L(x_{i,1})$ is just the likelihood for a Binomial with likelihood of success given by the retention probability ρ. To find $L(x_{i,j+1}|x_{i,j})$ we note $x_{i,j+1}$ is also Bernoulli, and if, for example, $x_{i,j} = 0$ then $x_{i,j+1}$ is a success with probability $\rho\theta_j$ (i.e. a break in a fragment at that location and retention of the human fragment). One can then maximize the likelihood (numerically) with respect to the parameters θ_j and ρ.

While we can look at ratios of log-likelihoods to compare the strength of evidence in favor of certain orderings, the Bayesian approach allows one to compute ratios between posterior probabilities in favor of certain orderings, and these are easily interpreted. One can more simply compute probabilities for many orders and retain those which are sufficiently high. In practice most of the posterior probability is confined to a short list of orders with appreciable probabilities (see Lange and Boehnke (1992)).

Exercises

1. Suppose we treat a 5.7 kb length sequence on which we would like to know the locations of the restriction sites for 2 restriction endonucleases, A and B. When we treat the sample with A we get products of length 2.6, 1.7 and 1.4. When we treat the sample with B we get 2.7, 2.0, and 1.0. When we use both, we get 2.0, 1.7, 1.0, 0.6, and 0.4.

   (a) List every set of loci for the restriction sites which are consistent with the fragmentation patterns.

   (b) To clear up ambiguities in the order we should run a partial restriction. Suppose the probability of breakage, p, is the same for all restriction sites. If there length of the measured fragments is measured with error so that we can approximate the lengths with a normal random variable with mean given by the true length and standard deviation of 0.2, what is the distribution of the fragment lengths for each of the orderings you listed in part A? Sketch histograms of these distributions.

   (c) Use your answer from part B. to explain which partial restriction you should run, A or B.

2. Use the Markov property to show we can write the likelihood for radiation hybrid data as

$$L(x;\rho,\theta_1,\ldots,\theta_{j-1}) = \prod_{i=1}^{n} L(x_{i,1}) \prod_{j=1}^{m-1} L(x_{i,j+1}|x_{i,j}).$$

3. Show that the likelihood for the radiation hybrid data can be expressed as

$$L(x;\rho,\theta_1,\ldots,\theta_{j-1}) = \prod_{i=1}^{n} \rho^{x_{i,1}}(1-\rho)^{1-x_{i,1}} \prod_{j=1}^{m-1} \Big(\rho\theta_j + x_{i,j}(1-\theta_j)\Big)^{x_{i,j+1}} \Big(1-\rho\theta_j - x_{i,j}(1-\theta_j)\Big)^{1-x_{i,j+1}}.$$

References

Lange, K., Boehnke, M. (1992), “Bayesian methods and optimal experimental design for gene mapping by radiation hybrids”, *Annals of Human Genetics*, 56:119–144.

Press, W., Teukolsky, S., Vetterling, W., and B. Flannery (1992), *Numerical Recipes in C*, Cambridge University Press, UK.