

Markers and physical mapping

1 Introduction

There have been tremendous advances in the mapping of genomes, the most notable being the recent completion of the human genome project (i.e. the sequencing of the human genome). 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 the use of genes. In addition, by knowing how genes work together to give rise to observed phenotypes, we can potentially produce medicines that alter protein expression in specific tissues so as to treat disease. Finally, 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.

To determine the sequence of nucleotides present in a DNA molecule, one starts with a large number of copies of a single stranded DNA molecule. These DNA molecules are stored in *clones* which are either bacteria, yeast or rodent cells that contain a segment of human DNA. A clone library is a collection of clones that cover a genome or chromosome. One then fluorescence labels a collection of nucleotides using 4 different fluorophores so that the identity of a nucleotide can be determined by a machine designed to detect each of the fluorophores used. In addition to this collection of labeled nucleotides, one creates a collection of molecules similar to the 4 nucleotides called dideoxynucleotides (ddNTPs) that differ from normal nucleotides in that as soon as a ddNTP is incorporated in a sequence of nucleotides, the sequence of nucleotides is terminated (there is a ddNTP that corresponds to each of the 4 usual nucleotides). One then combines the single stranded DNA molecules, the labeled nucleotides and the ddNTPs with compounds that promote copying the single stranded DNA molecule. When this happens, chains of labeled nucleotides are built up that terminate whenever a ddNTP is incorporated into the sequence. The result is a set of nucleotide sequences that differ in length and are copies of a portion of the original single stranded DNA sequences. The lengths of the fragments can be controlled by varying the concentration of the ddNTPs. The fragments are then separated using a denaturing polyacrylamide gel (i.e. a viscous material that has a denaturing compound embedded in it so that the DNA stays in its single stranded form). The larger fragments will not move as far in the gel as the lighter fragments, hence the set of fragments can be separated. One can then use a machine that reads the fluorophores embedded in the gel to determine the sequence of the starting DNA sequence. This process has been further automated using capillary-based DNA sequencing in which thin tubes are used in place of gels. A problem with this approach is that one can only sequence DNA fragments of length up to 750-1000 bases. To understand the extent that this is a limitation, note that the long arm of chromosome 22 (the shortest chromosome) is about 33.4 million basepairs (the long arm of chromosome 22 is the only part of that chromosome with DNA that is known to code for proteins).

For small genomes the *shotgun method* has provided a means of determining the sequence of a genome. This approach attempts to overcome the discrepancy between the length of a DNA molecule that can be sequenced and the typical size of a DNA molecule by first breaking up the molecule into smaller, overlapping segments, sequence these, then put it all back together-this is the shotgun method. As an example, suppose that an organism has the following DNA sequence on one **strand**

```
AAACATGTGTGTCAATATACGGCGCCTTTATATATATGTGCGTGC
```

Suppose one has a large collection of these sequences and then fragments them to obtain many subsequences, of which suppose 5 are

```
AACATGTGTGTCAATATA
  GTCAATATACGCGGCC
    TATACGCGGCCTTTAT
      GGCCTTTATATATATGT
        ATATATATGTGCGTG
          TGTGCGTGCGTGC
            GTGCGTGC
```

If one didn't know the original sequence, but knew the fragmented sequences, one can deduce the original sequence. To do so, one could label all of the fragments, then consider every possible permutation of these fragments and choose the shortest sequence that is consistent with all of the fragments. Unfortunately, for complex organisms, putting all the segments back together correctly is very difficult since the sequences are very long relative to the size of the DNA sequences one can insert into a vector for use in a clone library (recall the human genome is about 3 billion bp, or 3,000 megabase pairs, Mb). The largest clones one can use for constructing a clone library for humans (namely, yeast artificial chromosomes) require thousands of clones, hence the number of possible permutations of the fragments is too large to consider every possible permutation. In addition, there is a lot of repetitive subsequences in DNA. For example,

```
TGTTTGTTG TGTGTGTG TGTGTGTTG TGTCTGTTT
```

is a short segment of locus CAAK04001064 from the zebra fish. 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 either of the latter 2 approaches to be viable methods, one needs to know the location of some features of the DNA molecule to serve as landmarks. This is much like trying to look at a map of a large region: to determine the fine scale aspects of land that the map is intended to describe, one needs some feature that is detectable on the low resolution map (e.g. a mountain) and is present on all of the maps that are of finer resolution. The directed shotgun would try to determine from the high resolution map that there is a mountain, then use the mountain to make finer scale maps fit together, whereas the clone contig approach would use the fact that the mountain is visible to the low resolution map to then understand how higher resolution maps would fit together. We call these landmarks *markers*, a term we also use to refer to any known feature of a genome. The simplest markers indicate which chromosome is under consideration. There are 2 broad categories for methods aimed at finding the location of genes: *genetic mapping* and *physical mapping*. After more discussion of markers, we will first consider some of the approaches that have been used to construct physical maps, then we will consider genetic mapping.

2 Types of markers

At first, other than aspects of chromosomal organization and structure (such as the location of the centromere), genes were the only known markers. Genes with known locations 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. We will see that markers that take

different values (or have multiple alleles) are crucial for genetic mapping. A marker is said to be polymorphic if this is the case.

2.1 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. In addition, the locations of these markers are too spread out to use for fine scale mapping. However, these markers are used for physical mapping in a manner that will be discussed below.

2.2 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. The biological source of polymorphism at these loci is thought to have arisen due to a phenomenon called replication slippage that can occur when single stranded DNA is copied during cell division. Due to this variation in DNA sequence length among people, one can not just think of physical distance on a chromosome in terms of base pairs. Hence simple models for DNA in which the sequence is of some fixed length with different base pairs appearing in the various linearly arranged positions are inadequate for studying many aspects of DNA. There are 2 basic sorts of SSLPs: 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 that are repeated a number of times that varies across people). The alleles that someone has at a loci where a SSLPs is located are determined via a southern blot. Micro-satellites are preferred because they are more evenly distributed throughout the genome and the genotyping is faster and more accurate. However, a very small fraction of these markers are located in the coding portion of some gene. This last feature can be an undesirable feature for a marker since we generally want to use markers that are unassociated with a phenotype (as this will violate the assumptions that lead to Hardy-Weinberg equilibrium, as we will discuss in Section ??).

2.3 Single nucleotide polymorphisms (SNPs)

There are many (over 1,000,000) 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 RFLPs, 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, large scale genotyping can be automated (in contrast to the other 2 types of markers). Since a SNP can only take 2 values, the extent of polymorphism at a SNP is summarized by the frequency with which one observes the less frequent allele, which is called the minor allele frequency (the other allele is referred to as the major allele). Typically a SNP must have a minor allele frequency of at least 1% to qualify as a SNP. Automatic genotyping can be conducted using microarrays, a technology discussed in the final 4 chapters. Since the goal of using these markers is to determine the alleles for a subject at every location in the genome at which humans are known to differ, one can obtain the genome of a subject by determining the alleles someone has at this much smaller number of locations. These are rapidly becoming the most

common type of marker used to conduct genetic mapping due to automatic genotyping, however many data sets exist that used SSLPs. Researchers frequently distinguish between a coding SNP, i.e. a SNP in a coding region of a gene from a non-coding SNP. While both types play a role in genetic mapping, coding SNPs are of great interest since variation in that SNP can lead to differences in the gene product (recall that some codons map to the same amino acid so not all variations in the nucleotide sequence lead to changes in the amino acid sequence).

3 Physical mapping of genomes

Many methods for physical mapping 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. Here 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 whose distinctive feature is that clones are assayed to determine if they have a certain set of loci.

3.1 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 both enzymes at the same time. After each digest, one separates the fragments using gel electrophoresis (a method that exploits the fact that smaller fragments will move further through a gel when exposed to some force to make the fragments move) to determine the length of the products after the digest takes place. After conducting each of these digests, one attempts to determine where the locations of the restriction sites for each enzyme are located. For example, if using restriction enzyme a yields fragments of length a_1, \dots, a_n and use of restriction fragment b yields fragments of length b_1, \dots, b_m then there must be $n - 1$ restriction sites for enzyme a and $m - 1$ restriction sites for enzyme b , but that is all that is known (note $\sum_i a_i = \sum_i b_i$ since we are using the same sequence for both digests). If we can find the relative locations of these $m + n - 2$ sites then we will have a physical map that indicates the locations of certain known features (here the restriction sites). Thus the goal is to use the information from the experiment where both restriction enzymes are used to find these relative locations. Suppose there are only 2 restriction sites (one for each enzyme), so that each digest using a single enzyme produces 2 fragments. Each of these indicate how close to one end of the sequence each restriction site is, yet does not indicate the relative location of these 2 sites. If we assume that a_1 is the distance from the end of the sequence we will call 0 to the location of the restriction site for enzyme a , then if $a_1 < b_1$ and $a_1 < b_2$ then the use of both enzymes will either result in the set of fragments of length $a_1, b_1 - a_1$ and b_2 or $a_1, b_2 - a_1$ and b_1 . If we observe the first set of fragments then the restriction site for enzyme b must be a distance of b_1 from the end of the sequence that we have chosen to be zero. If we observe the second set of restriction fragments then the the restriction site for enzyme b must be at location b_2 from the end of the sequence taken to be zero. While for this simple example, one can determine the distance between the restriction sites, for even slightly more complex examples it is impossible to determine the order from the restriction using both (or, more generally, all) of the restriction enzymes (see exercise 1 for an example).

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*).

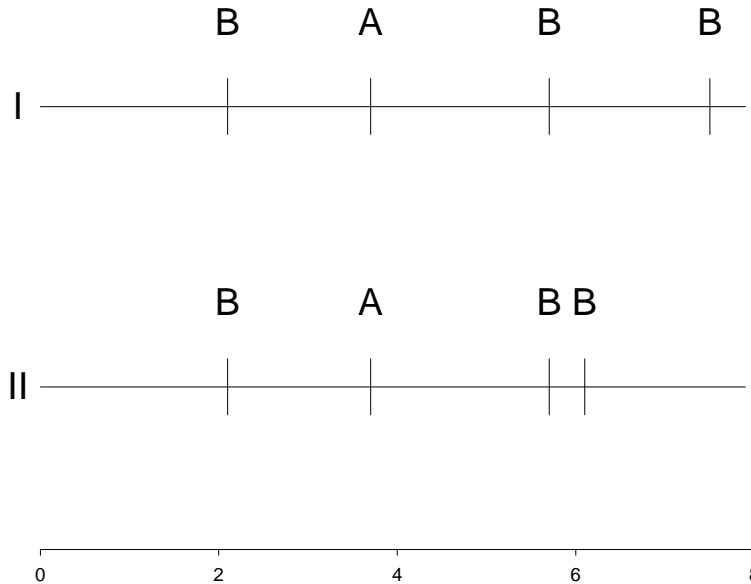


Figure 1: The 2 possible maps of locations of restriction fragment sites that are consistent with the example discussed in the text.

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). This problem can be viewed as attempting to find the permutation of the indices for the restriction fragments that is most consistent with the data.

Consider the following simple example. Suppose that we use 2 restriction enzymes, call them A and B , and suppose that after applying A to a 7.9 kb (i.e. kilobases, or 1000 bases) sequence of DNA we obtain fragments of the length 3.7 and 4.2, while after applying B we get fragments of length 0.4, 1.8, 2.1 and 3.6. This indicates that the restriction site for enzyme A is 3.7 kb from one of the ends of the original fragment, but not much else. Now suppose we use both enzymes simultaneously and obtain fragments of length 0.4, 1.6, 1.8, 2.0 and 2.1. Now much more is known about the relative locations of these restriction sites. After some thought, we can discern 2 possible maps for the relative positions of the restriction sites based on these sets of fragmentation patterns, and these are shown in Figure 1. Now suppose we have many copies of this 7.9 kb fragment and we run a partial restriction with enzyme A . This means that we stop the reaction before all of the fragments are cut by the restriction enzyme. If p_A is the probability that enzyme A cuts a fragment in this partial restriction (with p_B defined analogously), then we expect to observe a set of fragments where about $1 - p_A$ of the fragments are of length 7.9, about $p_A/2$ are of length 4.2 and $p_A/2$ are of length 3.7. This is because, for each fragment that gets cut, 2 fragments result and these 2 fragments will be found in

Table 1: Probabilities for observing lengths of fragments from a partial restriction.

| Map I | | Map II | |
|----------------|--------------------|----------------|--------------------|
| probability | fragment lengths | probability | fragment lengths |
| p_B^3 | 0.4, 1.8, 2.1, 3.6 | p_B^3 | 0.4, 1.8, 2.1, 3.6 |
| $p_B^2(1-p_B)$ | 0.4, 1.8, 5.7 | $p_B^2(1-p_B)$ | 0.4, 1.8, 5.7 |
| $p_B^2(1-p_B)$ | 0.4, 2.1, 5.4 | $p_B^2(1-p_B)$ | 1.8, 2.1, 4.0 |
| $p_B^2(1-p_B)$ | 2.1, 2.2, 3.6 | $p_B^2(1-p_B)$ | 2.1, 2.2, 3.6 |
| $p_B(1-p_B)^2$ | 2.1, 5.8 | $p_B(1-p_B)^2$ | 2.1, 5.8 |
| $p_B(1-p_B)^2$ | 2.2, 5.7 | $p_B(1-p_B)^2$ | 2.2, 5.7 |
| $p_B(1-p_B)^2$ | 0.4, 7.5 | $p_B(1-p_B)^2$ | 1.8, 6.1 |
| $(1-p_B)^3$ | 7.9 | $(1-p_B)^3$ | 7.9 |

exactly equal frequency. However the result is quite different if we use restriction enzyme B . To see this, we will suppose that each of the maps in Figure 1 is the true map, then examine the probability distribution of the length of the fragments. The result is shown in Table 1. We can see that each map has unique fragments that would not be observed if the other map was the true map.

Next note that if we let x represent the length of a random fragment that is sampled after the the partial restriction then we have

$$\begin{aligned}
 p(x) &= p(x|0 \text{ breaks})p(0 \text{ breaks}) + p(x|1 \text{ break})p(1 \text{ break}) \\
 &\quad + p(x|2 \text{ breaks})p(2 \text{ breaks}) + p(x|3 \text{ breaks})p(3 \text{ breaks}) \\
 &= p(x|0 \text{ breaks})(1-p_B)^3 + p(x|1 \text{ break})p(1 \text{ break}) \\
 &\quad + p(x|2 \text{ breaks})p(2 \text{ breaks}) + p(x|3 \text{ breaks})p_B^3.
 \end{aligned}$$

If we further suppose that the individual fragments lengths are measured with normally distributed error with variance σ^2 then we can get a more explicit expression for $p(x)$. First, under this condition we clearly have

$$p(x|0 \text{ breaks}) = N(x|7.9, \sigma^2).$$

With a little more work we find that

$$\begin{aligned}
 p(x|3 \text{ breaks}) &= (1/4) \left(N(x|0.4, \sigma^2) + N(x|1.8, \sigma^2) + N(x|2.1, \sigma^2) \right. \\
 &\quad \left. + N(x|3.6, \sigma^2) \right).
 \end{aligned}$$

This is because for each of the original fragments, we obtain 4 fragments and we measure the lengths of these fragments subject to normal errors. To determine $p(x|1 \text{ break})p(1 \text{ break})$, note that there are 3 ways that a break can occur, and for each of these breaks 2 fragments will be produced. Each of these 2 fragments will occur with equal frequency and all breaks occur with the same probability, $p_B(1-p_B)^2$. For each fragment that occurs from a single break, we will obtain a normal density, so we find that $p(x)$ is a linear combination of normal densities. The densities for both maps are plotted in Figure 2 for 4 choices of p_B . Note that when $p_B = 0.2$ almost all of the fragments are the length of the original fragment, whereas when $p_B = 0.8$ one can hardly distinguish between the 2 maps (the 4 peaks in that plot correspond to the lengths of fragments one would observe from the complete restriction). For intermediate cases, the 2 maps can be distinguished based on the observed fragment lengths.

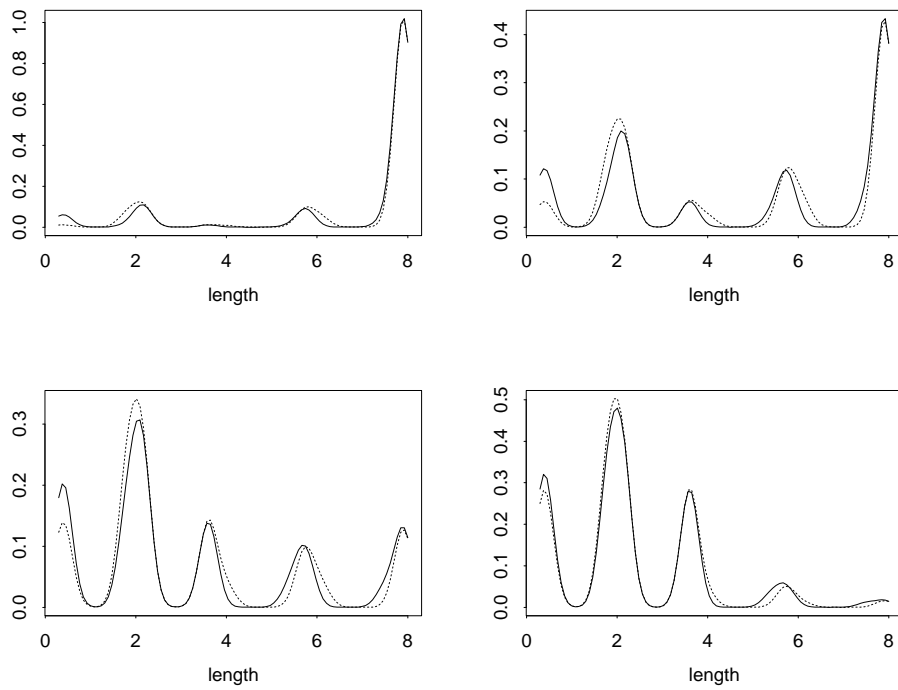


Figure 2: An example of the densities of the lengths of fragments observed with error from a partial restriction for 2 different marker maps. Four different values of p_B shown: 0.2 (upper left), 0.4 (upper right), 0.6 (lower left), and 0.8 (lower right). Map 1 is represented by the solid line.

3.2 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 a portion of the cell cycle known as 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.

3.3 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 and occurs once in a genome. ESTs are typically 500-800 nucleotides in length. There are 2 basic methods for producing the collection of DNA fragments: radiation hybrids and clone libraries. Below we discuss radiation hybrids in detail.

4 Radiation hybrid mapping

4.1 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.

4.2 Data from a radiation hybrid panel

The data for producing an estimate of locus order or marker location using radiation hybrid methods with m markers is an m -vector of zeros and ones for each hybrid cell. 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 determining if the human marker is present or not.

4.3 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 low 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 cell i with element $x_{i,j}$ for $i = 1, \dots, n$ and $j = 1, \dots, m$, 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)).

4.3.1 Consistency of the order

The permutation of the marker order that minimizes the number of obligate breaks provides a consistent method for ordering markers. An estimate is said to be consistent for a parameter when the accumulation of more observations makes the estimate progressively closer to the true value of that parameter. To see that this estimate is consistent, first note that by the law of large numbers, our estimate converges to $EB_1(\sigma)$. Assume that the identity permutation $I = (1, 2, \dots, m)$ is the true order of the markers (so our numerical ordering is the true order of the markers). This assumption is much like saying “suppose θ_0 is the true value of the unknown parameter” in cases where the parameter is a real number. Since the parameter space is finite (the set of permutations of a finite number of integers) if we just show

$$EB_1(\sigma) \geq EB_1(I),$$

then we will know our estimate is consistent. Since $B_1(\sigma) = \sum_j 1_{\{x_{1,\sigma(j)} \neq x_{1,\sigma(j+1)}\}}$, $EB_1(\sigma) = \sum_j P\{x_{1,\sigma(j)} \neq x_{1,\sigma(j+1)}\}$. To see why the identity permutation minimizes this expression consider the permutation $\sigma = (1, 2, 4, 3, 5)$. For this permutation

$$\begin{aligned} EB_1(\sigma) - EB_1(I) &= P(x_{1,2} \neq x_{1,4}) + P(x_{1,3} \neq x_{1,5}) \\ &\quad - P(x_{1,2} \neq x_{1,3}) - P(x_{1,4} \neq x_{1,5}). \end{aligned}$$

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 for any permutation other than the identity higher than it is for the identity permutation.

4.4 Maximum likelihood and Bayesian Methods

The method presented above has two shortcomings. First, we do not obtain estimates of the distances between 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. Likelihood based approaches can overcome both of these problems.

It suffices to consider how to compute the likelihood for a single hybrid (since all observations are assumed iid). 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 that breaks between the fixed locations given by the markers occur according to a Markov Chain. This implies that we can write the likelihood for a single observation as

$$\begin{aligned} L(x_i; \theta, \rho) &= L(x_{i,1}) \prod_j L(x_{i,j} | x_{i,j-1}, \dots, x_{i,1}) \\ &= L(x_{i,1}) \prod_j L(x_{i,j} | x_{i,j-1}) \end{aligned}$$

Now $L(x_{i,1})$ is just the likelihood for a Binomial with likelihood of success given by the retention probability ρ . For now, suppose the order of the loci is known and is the same as the numerical ordering we are using. 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). Similarly, if $x_{i,j} = 1$ then $x_{i,j+1}$ is a success with probability $\rho\theta_j + 1 - \theta_j$ (i.e. a break in a fragment at that location and retention of the human fragment or no break). We then find that we can write the likelihood for a single cell as

$$\begin{aligned} L(x; \rho, \theta_1, \dots, \theta_{j-1}) &= \prod_{i=1}^n \rho^{x_{i,1}} (1-\rho)^{1-x_{i,1}} \prod_{j=1}^{m-1} \left(\rho\theta_j + x_{i,j}(1-\theta_j) \right)^{x_{i,j+1}} \\ &\quad \times \left(1 - \rho\theta_j - x_{i,j}(1-\theta_j) \right)^{1-x_{i,j+1}}. \end{aligned}$$

One can then maximize the likelihood with respect to the parameters θ_j and ρ . To do so, note that if one differentiates the log-likelihood with respect to θ_j the result is

$$\frac{\partial \ell}{\partial \theta_j} = \sum_i \left[x_{i,j+1} \frac{\rho - x_{i,j}}{\rho\theta_j + x_{i,j}(1-\theta_j)} + \left(1 - x_{i,j+1} \frac{x_{i,j} - \rho}{1 - \rho\theta_j - x_{i,j}(1-\theta_j)} \right) \right].$$

If we then let n_{jk} represent the number of times that $x_{i,l} = j$ and $x_{i,l+1} = k$ we find that

$$\frac{\partial \ell}{\partial \theta_j} = n_{00} \rho / (\rho \theta_j) + n_{01} / \theta_j + n_{10} (1 - \rho) / (\theta_j (1 - \rho)) + n_{11} (\rho - 1) / (\rho \theta_j + 1 - \theta_j).$$

If we then set this partial derivative to zero and solve for θ_j we find that we have a quadratic equation in θ_j that only involves ρ , and n_{jk} for $j = 0, 1$ and $k = 0, 1$. Hence one can obtain an explicit expression for the MLE of θ_j , denoted $\hat{\theta}_j$, that just depends on ρ since it is not hard to show that the lower root of the equation will be negative. Thus one can find the MLEs here by substituting $\hat{\theta}_j(\rho)$ into the log-likelihood and conducting a univariate maximization to obtain $\hat{\rho}$ (which can be done by simply plotting the log-likelihood as a function of ρ). Once $\hat{\rho}$ is determined we can substitute this value into $\hat{\theta}_j(\rho)$ to obtain the MLEs of θ_j .

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)). For this reason, some first use the minimum number of obligate breaks to determine the order of the loci, then use likelihood based methods to determine the relative distance between the loci.

A characteristic of these data sets is that there are frequently missing values that arise due to ambiguity in the measured data. To account for this missing data, we use the EM algorithm treating the missing data as missing. We can use the previous expression for the likelihood to obtain an expression for the complete data log-likelihood. The E step then consists of taking the expectation of this expression conditional on current values for the retention probabilities and ρ . The M step is then just the same as finding the MLEs when there is no missing data.

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? Produce 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

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

3. Show that the lower root of the quadratic equation that one encounters in finding the MLE of θ_j for the radiation hybrid data experiments is always less than or equal to zero.
4. First show that if x_{ij} is Markov in the parameter j then

$$p(x_{ij}|x_{i,j-1}, x_{i,j+1}) \propto p(x_{i,j+1}|x_{ij})p(x_{ij}|x_{i,j-1}).$$

Then use this to show how to implement the EM algorithm for radiation hybrid panel data when information for one marker is missing, but we observe the outcome for the flanking markers.

5. Suppose that we treat the order of the markers, σ , as a parameter. Describe how one could find the MLE of σ . If we suppose that all orders are equally likely, describe how one could sample from the posterior distribution of σ (assume that all probabilities that appear in the model are uniformly distributed on $(0,1)$).
6. In the context of the model proposed for radiation hybrid panel data, suppose that ρ and θ_j have uniform priors. Also, treat the permutation σ as a parameter. Show that the marginal posterior of ρ and σ satisfies

$$p(\rho, \sigma) \propto \rho^{\sum_i x_{i,1}} (1 - \rho)^{n - \sum_i x_{i,1}} \prod_{ij} \eta_{i,j}^{x_{i,j+1}} (1 - \eta_{i,j})^{1 - x_{i,j+1}}$$

where

$$\eta_{i,j} = \frac{1}{2}(\rho + x_{i,j}).$$