Mapping Heredity: Using Probabilistic Models and Algorithms to Map Genes and Genomes

Eric S. Lander

The human genome is a vast biochemical jungle in which scientists have begun hunting for the genetic basis of inherited diseases. Even a one-letter error in the $3 \times 10^9$ base pairs (bp) of deoxyribonucleic acid (DNA) inherited from either parent may be sufficient to cause a disease. Thus, to detect inherited diseases, one must be able to detect mistakes present at just over one part in $10^{10}$. The task is sometimes likened to finding a needle in a haystack, but this analogy actually understates the problem: the typical 2-gram needle in a 6,000-kilogram haystack represents a 3,000-fold larger target. In certain respects, the gene hunter’s task is harder still, because it may be difficult to recognize the target even if one stumbles upon it. Although molecular biologists refer to the human genome as if it were well defined in mathematicians’ terms, it is recognized that, except for identical twins, no two humans have identical DNA sequences. Two genomes chosen from the human population are about 99.8 percent identical, affirming our common heritage as a species. But the 0.2 percent variation translates into some six million sequence differences. Common sites of sequence variations are called DNA polymorphisms. Most polymorphisms are thought to be nonfunctional variations, arising by mutation, having no deleterious consequence, and increasing (and decreasing) in frequency by stochastic drift. The presence of considerable DNA polymorphism in the population has sobering consequences for disease hunting. Even if it were straightforward to determine the entire DNA sequence of individuals, one could not find the gene for cystic fibrosis (CF) simply by comparing the sequences of a CF patient and an unaffected person: there would be too many polymorphisms.

How does a geneticist find the genes responsible for cystic fibrosis, diabetes, or heart disease? The answer is to proceed hierarchi-
The first step is to use a technique called \textit{genetic mapping} to narrow down the location of the gene to about 1/1,000 of the human genome. The second step is to use a technique called \textit{physical mapping} to clone the DNA from this region and to use molecular biological tools to identify all the genes. The third step is to identify candidate genes (based on the pattern of gene expression in different tissues and at different times) and to look for functional sequence differences in DNA (for example, mutations that introduce stop codons or that change crucial amino acids in a protein sequence) in affected patients.

We focus on genetic and physical mapping because they essentially involve mathematical analysis.

\textbf{Genetic Mapping}

\textbf{The Concept of Genetic Maps}

Genetic mapping is based on the perhaps counterintuitive notion that it is possible to find where a gene is without knowing what it is. Specifically, it is possible to identify the location of an unknown disease-causing gene by correlating the inheritance pattern of the disease in families with the inheritance pattern of known genetic markers. It is useful to return to Mendel’s Laws of Inheritance:

\textbullet{} \textit{First Law.} For any gene, each parent transmits one allele chosen at random to its offspring.

\textbullet{} \textit{Second Law.} For any two genes, the alleles transmitted by a parent are independent (that is, there is no correlation in the alleles transmitted).

Although Mendel’s First Law has held up well over the past 130 years, the Second Law turned out to be false in general. Two genes on different chromosomes show no correlation in their inheritance pattern, but genes on the same chromosome typically show correlation.

Consider the simple backcross in Figure 1, showing the inheritance of two genes \textit{A} and \textit{B} on the same chromosome. The \textit{F}_1 individual carries one chromosome with alleles \textit{a}_1 and \textit{b}_1 at the two genes and another chromosome with alleles \textit{a}_2 and \textit{b}_2. Often, one or the other chromosome is transmitted completely intact to the offspring. If this always happened, the inheritance pattern at the two genes would be completely dependent: \textit{a}_1 would always be co-inherited with \textit{b}_1. But the situation is more interesting. \textit{Crossing over} can occur at random points along the chromosomes, involving an even swap of DNA material. If a crossover occurs between genes \textit{A} and \textit{B}, it results in \textit{recombination} between the genes, producing a chromosome carrying a new combination of alleles: \textit{a}_1\textit{b}_2 or \textit{a}_2\textit{b}_1. In fact, multiple crossovers can occur along a chromosome; recombination between two loci will result whenever an odd number of crossovers occur.

Genetic mapping is based on the recognition that the \textit{recombination frequency} \(\theta\) between two genes (or loci) provides a measure of the distance between them. If two genes are close together, \(\theta\) will be small. If the recombination frequency is clearly less than 0.50, the genes are said to be \textit{linked}.

The \textit{genetic distance} \(d_{A,B}\) between two genes \textit{A} and \textit{B} is defined as the expected number of crossovers between the genes. If one assumes that crossovers are distributed independently with respect to one another (this assumption is not quite right but is adequate for many purposes), genetic distances can easily be converted into recombination frequency, for the number of crossovers between genes \textit{A} and \textit{B} will then be Poisson distributed with mean \(d = d_{A,B}\); and so the probability of an odd number of crossovers can be shown to be

\[\theta = \left(1 - e^{-2d}\right)/2.\]

For small distances, the formula is \(\theta \approx d\), which reflects the fact that the possibility of more than one crossover can be neglected. For large distances \(d\), the recombination frequency \(\theta\) approaches 0.50, that is, independent assortment.

Genetic mapping is an essential first step in characterizing a new mutation. Consider first the

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Schematic drawing of genetic recombination in an \textit{F}_1 heterozygote with distinct alleles at two loci (marked as \textit{A} and \textit{B}) on a chromosome. When no recombination occurs between \textit{A} and \textit{B} in meiosis, chromosomes carrying the original pair of alleles result. When recombination occurs, the resulting chromosomes carry a new combination of alleles.}
\end{figure}
but is tightly linked to locus C. The proportion of recombinant chromosomes provides a straightforward statistical estimator of the recombination frequency. In this case, the recombination frequency between A and B is about \( \frac{20}{200} = 10\% \). The gene A can be positioned more precisely by using a three-point cross shown in Figure 2b, in which two nearby genetic markers are segregating. Here, it is clear that A maps about midway between genes C and D (see figure caption).

For experimental organisms and simple traits, genetic mapping provides a straightforward way to locate the trait-causing gene to a small inter-
val. *Drosophila* geneticists rarely need to appeal to statistical or mathematical concepts. For geneticists studying human families or complex traits, however, the situation is quite different.

**Challenges of Genetic Mapping: Human Families and Complex Traits**

Medical geneticists studying diseases face two major problems: (1) for human diseases, one cannot arrange matings at will but must rather retrospectively interpret existing families; and (2) for both human diseases and animal models of these diseases, the trait may not be simply related to the genotype at a single gene. Owing to these complications, genetic mapping of disease genes often requires sophisticated mathematical analysis.

The first problem is the inability to arrange matings. To offset this limitation, human geneticists need to have a huge collection of frequent, naturally occurring genetic markers so that the inheritance pattern of each chromosomal region can be followed just as if one had deliberately set up a cross incorporating specific genetic markers. In 1980 David Botstein set off a revolution by recognizing that the naturally occurring DNA polymorphisms in the human population filled the need [4]. By 1994, over 4,000 DNA polymorphisms had been identified and mapped relative to one another.

Even with a dense genetic map of DNA polymorphisms, human genetic mapping confronts several special problems of incomplete information: (1) For individuals homozygous \((a_1/a_1)\) at a gene, one cannot distinguish between the two homologous chromosomes at this location. (2) For individuals heterozygous \((a_1/a_2)\) at a gene, one cannot tell which allele is on the paternal chromosome and which is on the maternal chromosome unless one can study the individual's parents. (3) Information for deceased individuals (or for those who choose not to participate in a genetic study) is completely missing from the pedigree.

Another problem is that many traits and diseases do not follow simple Mendelian rules of inheritance. This problem has several aspects:

- **Incomplete penetrance.** For some "disease genes", the probability that an individual inheriting the disease gene will have the disease phenotype may be less than 1. This probability is called the **penetrance** of the disease genotype. Penetrance may depend on other unknown genes, age, environmental exposure, or random chance. For example, a gene called *BRCA1* on chromosome 17 predisposes to early onset of breast cancer in some women, but the penetrance is estimated to be about 60 percent by age 50 and 85 percent by age 80.

- **Phenocopy.** Some diseases can be due to non-genetic causes. For example, colon cancer can be caused by mutations in the APC gene on human chromosome 5, but most cases of colon cancer are thought to be nongenetic in origin (and are often attributed to diet).

- **Genetic heterogeneity.** Some diseases may be caused by mutations in any one of several different genes.

- **Polygenic inheritance.** Some diseases may involve the interaction of mutations at several different genes simultaneously.

**Maximum Likelihood Estimation**

To handle the problem of incomplete information, geneticists have adopted the statistical approach of **maximum likelihood estimation** (MLE).

The geneticist would ideally like to have complete genotypic data \(X\)—for example, the genotype for every family member, including the precise parental chromosome from which each allele was inherited. Given complete information, it is usually easy to estimate the required parameters: for example, the recombination frequency can be estimated by counting recombinant chromosomes, and the penetrance can be estimated by finding the proportion of individuals with a disease-predisposing genotype who manifest the disease. Unfortunately, one typically has only incomplete data \(Y\), from which it is difficult to estimate \(\theta\) directly.

The maximum likelihood estimate \(\hat{\theta}\) is the value that makes the observed data \(Y\) most likely to have occurred, that is, the value that maximizes the likelihood function \(L(\theta) = \text{prob}(Y|\theta)\). Using Bayes' Theorem, one can calculate \(L(\theta)\).

To determine whether \(\hat{\theta}\) is significantly different from a null value \(\theta_0\) (for example, to see whether an estimated recombination frequency is significantly less than 50 percent), one examines the likelihood ratio \(Z = L(\hat{\theta})/L(\theta_0)\). If \(Z\) exceeds some appropriate threshold \(T\), a statistically significant effect has been found.

In principle, virtually any genetic problem can be treated by this approach. In practice, two important issues arise:

**Efficient Algorithms.** The number of terms in the Bayes sum scales as roughly \(O(c^{mn})\), where \(m\) is the number of people in the family, \(n\) is the number of genetic markers studied, and \(c\) is a constant. Except in the case of the smallest problems, it is infeasible to enumerate all the terms in the sum. Thus, it is a challenge even to calculate the likelihood \(L(\theta)\) at a single point, let alone to find the value \(\theta\) that maximizes the function. Considerable mathematical attention has been devoted to finding efficient ways to calculate \(L(\theta)\).

Recently, mathematical geneticists have explored ways to approximate \(L(\theta)\) by sampling.
Dove and his colleagues showed that Min was in fact a mutation in the mouse version of the APC gene [20]. The Min mouse thus provided a model of human colon cancer and, in particular, a way to look for other genes that might suppress the development of colon tumors. When Dove and colleagues crossed this mouse to another mouse strain called AKR, they got a surprising result: progeny develop many fewer colon tumors.

A backcross was arranged in which the F1 progeny were mated back to the more susceptible strain (Figure 3). For any modifier locus, 50 percent of the progeny should inherit one copy of the suppressing allele from the AKR strain (that is, have genotype AB) and 50 percent should be homozygous for the nonsuppressing allele (that is, have genotype BB). Each animal inheriting the Min mutation was scored for its phenotype by dissecting the intestine and counting the number of tumors and for its genotype by typing the mice for a dense map of DNA polymorphisms that had been constructed in our laboratory [6].

The data for animal i can be thought of as a phenotype \( \phi_i \) and a continuous function \( g_r(x) \) indicating the genotype, which is either AB or BB at each position along the chromosome (Figure 4).
At every position \( x \) along the chromosome, the animals can be divided into two sets according to their genotype:

\[
AB(x) = \{ \text{animal } i \mid g_1(x) = AA \} \\
and BB(x) = \{ \text{animal } i \mid g_1(x) = BA \}.
\]

If a major modifier gene occurs at location \( x \), then the animals in \( AB(x^+) \) should have many fewer tumors than the animals in \( BB(x) \). One could thus perform a \( t \)-test at every position along the chromosome to find a region where the \( t \)-statistic \( Z(x) \) exceeds some critical threshold \( T \).

How high a threshold is needed to ensure statistical significance, if one scans the entire genome? We will focus on a single chromosome. If there is no modifying gene along the chromosome, the \( t \)-statistic \( Z(x) \) at any given point \( x \) should be normally distributed with mean 0. It is thus easy to determine the appropriate significance level for the single test at \( x \). But we need to know about the distribution of \( \max_{x \in G} Z(x) \), where the maximum is taken over the entire chromosome.

It is not hard to show that the statistics \( Z(x) \) in our genetic example follow an Ornstein-Uhlenbeck process with \( \beta = 2 \). Using recent mathematical results [7, 16], one can thus show that, for large \( t \),

\[
\text{Prob} \{ \max_{0 \leq x \leq G} Z(x) \geq t \} \sim 2Gt^2(1 - \Phi(t)),
\]

where \( \Phi(t) \) is the standard normal distribution function and \( G \) is the length of the chromosome measured in expected numbers of crossovers (a unit called the morgan). In short, the probability of exceeding threshold \( t \) somewhere along a genome of length \( LG \) is larger by a factor of \( 2Gt^2 \) than the probability of exceeding it at a single point.

Returning to the problem of colon cancer, we applied this analysis to the entire mouse genome (the genetic length \( G = 16 \) morgans). By genetic mapping, we found a striking region on mouse chromosome 4 for which \( Z_{\max} = 4.5 \). The nominal significance level of the statistic is \( p = 3.5 \times 10^{-6} \). After correcting for searching over an entire genome (by multiplying by \( 2G(Z_{\max})^2 \)), the genome-wide significance level is \( p \approx 0.002 \). This suggests that there is indeed a modifying gene in this region of chromosome 4.

On the strength of this analysis, several additional crosses were arranged to confirm this result. With more than 300 animals analyzed, the results are now unambiguous: the corrected significance level is now \( < 10^{-10} \), and it appears that a single copy of the suppressing form of the gene can decrease tumor numbers at least twofold. Physical mapping is now underway to clone the gene in order to learn its role of reducing colon cancer in genetically predisposed mice. With luck, it may suggest ways to do the same in humans.

References


