MARKOV PROCESSES FOR MODELING AND ANALYZING A NEW GENETIC MAPPING METHOD

BY

Eleanor Feingold

TECHNICAL REPORT NO. 5

July 1992

PREPARED UNDER THE AUSPICES OF

NATIONAL SCIENCE FOUNDATION GRANT DMS91-04432

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Markov Processes for Modeling and Analyzing a New
Genetic Mapping Method

By

Eleanor Feingold

Technical Report No. 5
July 1992

Prepared Under the Auspices
of
National Science Foundation Grant DMS91-04432

Department of Statistics
Stanford University
Stanford, California
Markov Processes for Modeling and Analyzing A New Genetic Mapping Method

Eleanor Feingold, Stanford University

ABSTRACT

This paper describes a set of stochastic processes that is useful for modeling and analyzing a new genetic mapping method. Some of the processes are Markov chains, and some are best described as functions of Markov chains. The central issue is boundary crossing probabilities, which correspond to p-values for the existence of genes for particular traits. The methods elaborated by Aldous (1989) provide very accurate approximate p-values, as spot-checked against simulations.

Keywords: Markov chain, genetics, genetic mapping, boundary crossing probability

1. INTRODUCTION

Genetic mapping means finding the location on the genome of the gene responsible for some trait of interest, for example susceptibility to breast cancer in humans. A method that is very useful for mapping complex traits such as disease susceptibility involves comparing the genetic material of two relatives who both have the trait (Risch (1990)). If a gene for the trait exists, it should be in a region of the genome where the genetic material of the two relatives matches (more precisely, a region where they share "identity by descent"). Currently, the comparison of the genetic material to see if it matches can only be done at a small number of "markers" on the genome. Thus a great deal of research effort is focused on either expanding the set of available markers or finding some other method of getting a continuous specification of the regions of the genome where two (or more) relatives share identity by descent.

This paper looks at how to analyze data from pairs of affected relatives once a continuous specification of the regions of identity by descent is available. I describe a set of stochastic processes that models the regions of identity by descent for many pairs at
once. Different processes correspond to different types of relative pairs, e.g. siblings, cousins, etc. Some of the processes are Markov chains, and some are best described as functions of Markov chains. The central issue is boundary crossing probabilities for the processes, which correspond to p-values for the existence of genes for particular traits. I approximate the boundary crossing probabilities using techniques of the type described by David Aldous in *Probability Approximations via the Poisson Clumping Heuristic*. Section 2 of this paper describes the mapping experiment and the basic statistical question. Sections 3 through 7 give detailed descriptions of the stochastic processes of interest and the solutions to the boundary crossing problems. Section 8 discusses related problems, some of which remain to be solved. Since the laboratory methods are still in development, my work is oriented towards developing models for understanding the genetics rather than towards developing detailed procedures for handling experimental data. The particular impetus for my work is the research of Patrick Brown at Stanford University (Brown (1992)), who has made significant progress on a method for getting continuous specification of regions of identity by descent.

2. THE GENETIC MAPPING EXPERIMENT

The basic principle behind affected-pair mapping methods is as follows. If you chose a pair of relatives who share a trait and find the region where they share identity by descent, the gene for the trait should be somewhere in that region. If you then take another affected pair from another family and find their matching region, you expect the gene to fall somewhere into the overlap of the two matching regions. So by finding many such pairs from many different families, you can narrow down the area into which the gene must fall.

Figure 1 shows the regions of identity by descent for a grandchild and maternal grandmother, under the simplifying assumption of one pair of chromosomes per person. The grandchild gets one chromosome from each parent, and it is a mixture of that parent's
two chromosomes. (The diagram does not distinguish the chromosomes on the paternal side). There is a standard Poisson model due to Haldane (Ott (1985), p.8) that is used to describe the crossovers during meiosis which create that mixture. It tells us that the grandchild's maternal chromosome should match the grandmother half of the time and the grandfather half of the time, with the switches coming at independent exponential distances along the chromosome. Coding a "1" for matching and a "0" for non-matching, this gives a continuous-time Markov process for the grandmother/grandchild comparison on the state space 0 and 1, with the "time" axis being distance from the end of the chromosome.

![Diagram of a family tree showing genetic matching between generations.](image)

**Figure 1.**

With \( N \) independent grandparent/grandchild pairs, one can add up the component Markov processes and get a process on the state space \( \{0, 1, \ldots, N\} \), where the state is the
number of pairs that match at that point on the chromosome. This might result in a sample path like Figure 2. The peak in the sample path is the logical place to look for the gene, since it is the place on the genome where most pairs share identity by descent. If the trait of interest is absolutely controlled by a single gene, one would expect the peak value to be equal to $N$. More commonly, for example if the trait is a disease, people may have the trait without having the gene or may have the gene without having the trait. So the peak would normally be at something less than $N$. This leads to the statistical testing problem of deciding when a peak is high enough to be significant.

![Graph showing a sample path with a peak at N](image)

Figure 2.

The formal test I propose is as follows. (This is the likelihood ratio test in the cases where the stochastic process is a Markov chain, but not in the cases where it is a function of a Markov chain.) The null hypothesis is that there is no gene, so that the comparison for each pair is just the Markov process implied by the meiosis model. The alternative hypothesis is that at least some of the pairs share a gene for the disease, i.e., that some of the pairs follow the same process conditional on being equal to 1 at some (unknown) point.
on the genome. If we have observed data with a peak of height \( b \), we want to calculate the probability (p-value) of seeing a peak of height \( b \) or higher under the null hypothesis. If the probability is sufficiently small, we reject the null hypothesis and conclude that there is a shared gene. The p-value calculation is the boundary-crossing problem: what is the probability that the null Markov process reaches height \( b \) within the fixed length of the genome?

3. GRANDPARENT/GRANDCHILD PAIRS

The stochastic process representing the comparison of a single grandparent/grandchild pair is a continuous-time Markov chain on the states 0 and 1. The \( Q \)-matrix is

\[
\begin{pmatrix}
-\lambda & \lambda \\
\lambda & -\lambda
\end{pmatrix},
\]

where \( \lambda \) is the crossover rate per unit length, usually \( .01/\text{centimorgan}(\text{cM}) \). Under the null hypothesis, the process starts in its stationary distribution, which is \( P(0) = 1/2 \) and \( P(1) = 1/2 \). Adding up \( N \) pairs gives a Markov process on the state space \{0,1, ..., N\}, with transition rates \( Q(i,i+1) = \lambda(N-i) \) and \( Q(i,i-1) = \lambda i \) and stationary distribution binomial\((N,1/2)\). For large \( N \), this can be approximated by the Ornstein-Uhlenbeck process satisfying the stochastic differential equation \( dZ_t = -2\lambda Z_t dt + 2\sqrt{\lambda} dW_t \), where \( W_t \) is standard Brownian motion.

The p-value is the probability that the comparison process for \( N \) pairs, \( X_t \), exceeds the observed peak level \( b \) somewhere in the genome. It suffices to solve the problem for a single chromosome, because

\[
P\{\text{peak of height } b \text{ somewhere on 23 chromosomes}\} = 1 - \prod_{i=1}^{23} P\{\text{no peak on chromosome } i\}
\]
since the crossovers during meiosis occur independently in each of the 23 chromosomes. So, restricting attention to the single-chromosome problem, we want to approximate

$$\alpha = P\left\{ \max_{0 \leq t \leq l} X_t \geq b \right\},$$

where $l$ is the length of the chromosome in centimorgans. The most useful method for approximating this probability (useful is defined as accurate, tractable to compute, and applicable to all types of relative pairs) is to assume that the time $T_b$ until the level $b$ is first reached is exponentially distributed. Good references for this approach are Keilson (1979), Aldous (1989), and Leadbetter et al. (1983). I use methods of the type described by Aldous to approximate the mean, $ET_b$. I refer the reader to Aldous for a full description of his methods and terminology. These methods give accurate results for both small and large $N$, as long as $b$ is large enough that $\alpha$ is small.

For the grandparent/grandchild pair, Aldous' example B4 is directly applicable. The method is to approximate the Markov chain at $b$ by a simple random walk with up transition rate $\lambda(N-b)$ and down rate $\lambda b$. Then the expected clump size is

$$[\lambda b - \lambda(N-b)]^{-1} = [\lambda(2b-N)]^{-1}. \quad \text{And}$$

$$\pi(b) = \binom{N}{b} \frac{1}{2^N},$$

so

$$ET_b = \left[ \binom{N}{b} \frac{1}{2^N} (2b-N)\lambda \right]^{-1},$$

and the p-value is

$$\alpha = 1 - \exp\left\{ -t \left[ \binom{N}{b} \frac{1}{2^N} (2b-N)\lambda \right] \right\}. $$
Simulations show that this approximation is extremely accurate. Table 1 shows two sets of results that represent "typical" numbers for \( N \) and \( l \). If we are interested in p-values around .05 over twenty-three chromosomes, we should look at values around .002 for a single chromosome.

\[
\begin{array}{ccc}
\text{100,000 repetitions} & \text{100,000 repetitions} \\
N = 100 & N = 20 \\
\lambda = .01 & \lambda = .01 \\
l = 100 & l = 100 \\
\end{array}
\]

\[
\begin{array}{ccc}
p-values via & p-values via \\
\hline
b & simulation & approximation & b & simulation & approximation \\
67 & .00827 & .00787 & 17 & .01520 & .01511 \\
68 & .00425 & .00405 & 18 & .00281 & .00289 \\
69 & .00201 & .00199 & 19 & .00027 & .00034 \\
70 & .00099 & .00093 &
\end{array}
\]

Table 1.

To get a feel for how good the approximations are over the whole genome, assume 23 chromosomes of length 100. Then the p-values over 23 chromosomes are as shown in Table 2.

\[
\begin{array}{ccc}
p-values via & p-values via \\
\hline
b & simulation & approximation & b & simulation & approximation \\
67 & .174 & .166 & 17 & .299 & .295 \\
68 & .093 & .089 & 18 & .063 & .064 \\
69 & .045 & .045 & 19 & .006 & .008 \\
70 & .023 & .023 &
\end{array}
\]

Table 2.
4. **HALF-SIBLINGS**

Figure 3 shows the regions of identity by descent for a pair of half siblings. The stochastic process for this comparison has exactly the same structure as the grandparent/grandchild case, but the transition rate is $2\lambda$, because there is a transition from matching to non-matching or vice versa whenever either half-sibling switches from one grandparent to the other. The $Q$-matrix for a single pair is

\[
\begin{pmatrix}
-2\lambda & 2\lambda \\
2\lambda & -2\lambda
\end{pmatrix}
\]
Thus the rest of the development for the half-sibling case follows the
grandparent/grandchild case exactly, and p-values can be calculated by the formula

\[ \alpha = 1 - \exp\left\{ -L\left[ \left. N \right|_{b} \frac{1}{2^N} (2b-N)2\lambda \right] \right\} . \]

5. SIBLINGS

As Figure 4 shows, in the sibling case the process for a single pair is no longer a
chain on the state space 0 and 1. It can take values 0, 1, or 2, because siblings can match
on both chromosomes. To further complicate matters, laboratory techniques do not allow
us to distinguish readily between one match and two. So the process we actually observe
for a single sibling pair is a function of a 3-state Markov chain. Call the observed process
\( Y_t \) and the underlying process \( X_t \). \( Y_t = 1 \) whenever \( X_t = 1 \) or 2.

The \( Q \)-matrix for \( X_t \) with one pair of siblings is

\[
\begin{pmatrix}
-4\lambda & 4\lambda & 0 \\
2\lambda & -4\lambda & 2\lambda \\
0 & 4\lambda & -4\lambda \\
\end{pmatrix}
\]

For \( N \) pairs, \( X_t \) can be described as a chain on a two-dimensional state space. The transition rates are

\[
Q_X \begin{pmatrix} i \\ j \end{pmatrix}, \begin{pmatrix} i-1 \\ j+1 \end{pmatrix} = 4\lambda j
\]

\[
Q_X \begin{pmatrix} i \\ j \end{pmatrix}, \begin{pmatrix} i+1 \\ j-1 \end{pmatrix} = 2\lambda j
\]

\[
Q_X \begin{pmatrix} i \\ j \end{pmatrix}, \begin{pmatrix} i \\ j-1 \end{pmatrix} = 2\lambda j
\]

\[
Q_X \begin{pmatrix} i \\ j \end{pmatrix}, \begin{pmatrix} i \\ j+1 \end{pmatrix} = 4\lambda k
\]

where \( i \) = number of single-pair chains in state 0, \( j \) = number in state 1, and \( k \) = number in state 2 (\( i + j + k \) is constrained to equal \( N \), so the process is two-dimensional). The stationary distribution of \( X_t \) is multinomial\( \{N, \frac{1}{4}, \frac{1}{2}, \frac{1}{4} \} \). \( Y_t \) for \( N \) pairs is just \( j + k \).

To find the p-value, I consider \( Y_t \). This is similar to Aldous' example B19, where he finds a clump size for a process that is the sum of two dimensions of a vector-valued Markov chain by making a Markov version of the process. Let \( b \) be the observed maximum of \( Y_t \) for \( N \) pairs. Then I can calculate the stationary probability, \( \pi_Y(b) \), exactly as

\[
\pi_Y(b) = \sum_{i=0}^{b} \pi_X(i \text{ ones and } b-i \text{ twos in } X_t)
\]

\[
= \sum_{i=0}^{b} \pi_X \left( \begin{pmatrix} N-b \\ i \\ b-i \end{pmatrix} \right)
\]
\[
\begin{align*}
&= \sum_{i=0}^{b} \binom{N}{N-b, i, b-i} \frac{1}{4^{N-b-i}} \frac{1}{2^i} \frac{1}{4^{b-i}} \\
&= \binom{N}{b} \left( \frac{3}{4} \right)^b \left( \frac{1}{4} \right)^{N-b}.
\end{align*}
\]

To create the Markov version of \( Y_t \), we must compute what I'll call \( Q_Y(b, b+1) \) (not really a transition rate since \( Y_t \) is not a Markov chain). This is done by averaging over all states of \( X_t \) for which \( Y_t = Y(X_t) \) equals \( b \). So

\[
Q_Y(b, b+1) = \sum_{x \in X: Y(x) = b} \frac{\pi_X(x)}{\pi_Y(b)} \left( \sum_{w \in X: Y(w) = b+1} Q_X(x, w) \right)
\]

\[
= \sum_{j=0}^{b} \frac{1}{\pi_Y(b)} \pi_X \left( \binom{N-b}{j} \binom{N-b-1}{b-j} \right) Q_X \left( \binom{N-b}{j}, \binom{N-b}{b-j+1}, \binom{N-b}{b-j} \right)
\]

\[
= 4\lambda(N-b),
\]

where \( X \) is the state space of \( X_t \). Similarly, \( Q_Y(b, b-1) = 4\lambda b/3 \). This gives

\[
ET_b = \left[ \binom{N}{b} \left( \frac{3}{4} \right)^b \left( \frac{1}{4} \right)^{N-b} \left( \frac{4\lambda b}{3} - 4\lambda(N-b) \right) \right]^{-1},
\]

and thus

\[
\alpha = 1 - \exp \left[ -1 \left[ \binom{N}{b} \left( \frac{3}{4} \right)^b \left( \frac{1}{4} \right)^{N-b} \left( 4\lambda b - 3N \right) \frac{4\lambda}{3} \right] \right].
\]

Because the "Markovization" introduces error, we do not expect this approximation to be as good as the grandparent/grandchild approximation. Simulations bear that out, but the accuracy is still excellent, as Table 3 shows.
one chromosome (100,000 repetitions)
$N = 100 \quad \lambda = .01 \quad l = 100$

23 chromosomes

p-values via

<table>
<thead>
<tr>
<th>b</th>
<th>simulation</th>
<th>approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>.01723</td>
<td>.01896</td>
</tr>
<tr>
<td>90</td>
<td>.00691</td>
<td>.00749</td>
</tr>
<tr>
<td>91</td>
<td>.00248</td>
<td>.00264</td>
</tr>
<tr>
<td>92</td>
<td>.00081</td>
<td>.00082</td>
</tr>
</tbody>
</table>

p-values via

<table>
<thead>
<tr>
<th>b</th>
<th>simulation</th>
<th>approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>.330</td>
<td>.356</td>
</tr>
<tr>
<td>90</td>
<td>.147</td>
<td>.159</td>
</tr>
<tr>
<td>91</td>
<td>.056</td>
<td>.059</td>
</tr>
<tr>
<td>92</td>
<td>.018</td>
<td>.019</td>
</tr>
</tbody>
</table>

Table 3.

6. AUNT/NIECE PAIRS

Figure 5.
Figure 5 shows the aunt/niece comparison (the name "aunt/niece" is of course arbitrary). This process is not a Markov chain. It can, however, be described as a function of an eight-state Markov chain, where the underlying chain reflects the outcomes of all the meioses that determine the aunt/niece comparison. The eight-state chain is defined as follows. Let $X_t^{(1)}$ be the indicator of whether the niece's maternal chromosome matches the grandmother at point $t$, i.e., the grandparent/grandchild process for the grandmother and the niece. Let $X_t^{(2)}$ be the indicator of whether the aunt and the mother match in their maternal chromosome. And let $X_t^{(3)}$ be the indicator of whether the aunt and the mother match in their paternal chromosome. $X_t^{(2)}$ and $X_t^{(3)}$ are both equivalent to half-sibling processes. The three processes are independent, because they each indicate the outcome of a different meiosis. Together they form a Markov process on a state space of size $2^3=8$. The stationary distribution is $\left(\begin{array}{cccccccc} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{array}\right)$. As in all the other cases, this chain starts in its stationary distribution under the null hypothesis.

For a single pair, the $Q$-matrix for $(X_t^{(1)}, X_t^{(2)}, X_t^{(3)})$ is

<table>
<thead>
<tr>
<th>state #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>state:</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\begin{pmatrix}
-4\lambda & 2\lambda & 2\lambda & 0 & \lambda & 0 & 0 & 0 \\
2\lambda & -4\lambda & 0 & 2\lambda & 0 & \lambda & 0 & 0 \\
2\lambda & 0 & -4\lambda & 2\lambda & 0 & 0 & \lambda & 0 \\
0 & 2\lambda & 2\lambda & -4\lambda & 0 & 0 & 0 & \lambda \\
\lambda & 0 & 0 & 0 & -4\lambda & 2\lambda & 2\lambda & 0 \\
0 & \lambda & 0 & 0 & 2\lambda & -4\lambda & 0 & 2\lambda \\
0 & 0 & \lambda & 0 & 2\lambda & 0 & -4\lambda & 2\lambda \\
0 & 0 & 0 & \lambda & 0 & 2\lambda & 2\lambda & -4\lambda \\
\end{pmatrix}
\]

Let $Y_t$ be the aunt/niece process that we observe (the indicator of whether the aunt and the
niece match). Then for a single pair, \( Y_t = 1 \) whenever \( (X_t^{(1)}, X_t^{(2)}, X_t^{(3)}) \) is in state 3, 4, 6, or 8. For \( N \) pairs, the underlying process can be constructed just as it was for siblings. In this case, it is on a seven-dimensional state space, with stationary distribution multinomial \( \left( \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8} \right) \). Just as in the sibling case, the observed process is the sum of several dimensions of a vector-valued underlying chain. In this case, \( Y_t \) for \( N \) pairs is the sum of the numbers of pairs in states 3, 4, 6, and 8.

The method that I used to find \( p \)-values in the sibling case can be applied here, but with eight states instead of three it is cumbersome. A computationally easier approach is to apply the sibling-type "Markovization" to the process for a single pair and then assume that the sum of the Markov approximations to the one-pair processes behaves similarly to the Markov approximation to the sum of the processes. In fact, it does more than behave similarly; it is identical.

To prove that the two approaches are identical, it suffices to prove that they give the same \( Q_y(b, b+1) \), since \( \pi(b) \) is the same for both by definition. The most general formulation of the problem is to assume we are given \( N \) independent copies of some Markov process on \( m \) states. This process has transition rate matrix \( Q_1 \) and stationary distribution \( (p_1, p_2, \ldots, p_m) \). \( X_t \) is the vector-valued process that tells us how many of the copies are in each state, and it can be written as \( (X_1, X_2, \ldots, X_m) \), (suppressing the time dependence). \( X_t \) has transition rate matrix \( Q_X \) and stationary distribution \( \pi_X = \text{multinomial} (N, (p_1, p_2, \ldots, p_m)) \). There is a process \( Y_t \) that is the sum of several dimensions of \( X_t \). Without loss of generality, let \( Y_t = \sum_{i=j+1}^{m} X_i \).

Then Aldous' version of \( Q_y(b, b+1) \) can be written as

\[
Q_y(b, b+1) = \sum_{x \in X: x_{j+1} + \ldots + x_m = b} \frac{\pi_X(x)}{\pi_Y(b)} Q_X(x, "b+1") ,
\]

where \( Q_x(x, "b+1") \) is shorthand for
\[
\sum_{w \in X: w_{j+1} + \ldots + w_m = b+1} Q_X(x, w).
\]

And my version of \( Q_Y(b, b+1) \) can be written as

\[
Q_Y^*(b, b+1) = (N-b) \sum_{i=1}^{j} \frac{p_i}{\sum_k p_k} Q_1(i, "1") ,
\]

where, similarly, \( Q_1(i, "1") \) is shorthand for

\[
\sum_{k=j+1}^{m} Q_1(i, k).
\]

**Theorem.** \( Q_Y^*(b, b+1) = Q_Y(b, b+1) \).

**Proof.**

\[
Q_Y(b, b+1) = \sum_{x \in X: x_{j+1} + \ldots + x_m = b} \frac{\pi_X(x)}{\pi_X(b)} Q_X(x, "b+1")
\]

\[
= \sum_{x \in X: x_{j+1} + \ldots + x_m = b} \frac{\pi_X(x)}{\pi_X(b)} \left( \sum_{i=1}^{j} x_i Q_1(i, "1") \right)
\]

\[
= \frac{1}{\pi_X(b)} \sum_{i=1}^{j} Q_1(i, "1") \left( \sum_{x \in X: x_{j+1} + \ldots + x_m = b} x_i \pi_X(x) \right)
\]

\[
= \frac{1}{\pi_X(b)} \sum_{i=1}^{j} Q_1(i, "1") E \left( X_i | x_{j+1} + \ldots + x_m = b \right)
\]

\[
= \frac{1}{\pi_X(b)} \sum_{i=1}^{j} Q_1(i, "1") P(X_{j+1} + \ldots + X_m = b) E \left( X_i | X_{j+1} + \ldots + X_m = b \right)
\]

15
\[
\begin{align*}
&= \frac{1}{\pi_X(b)} \sum_{i=1}^{j} Q_1(i, "1") \left( \binom{N}{b} \left( \sum_{k=1}^{j} p_k \right)^{N-b} \left( \sum_{k=j+1}^{m} p_k \right)^{b} E\{X_i|X_{j+1} + \ldots + X_m = b\} \\
&= \frac{1}{\pi_X(b)} \sum_{i=1}^{j} Q_1(i, "1") \left( \binom{N}{b} \left( \sum_{k=1}^{j} p_k \right)^{N-b} \left( \sum_{k=j+1}^{m} p_k \right)^{b} E\{X_i|X_1 + \ldots + X_j = N-b\} \\
&= \frac{1}{\pi_X(b)} \sum_{i=1}^{j} Q_1(i, "1") \left( \binom{N}{b} \left( \sum_{k=1}^{j} p_k \right)^{N-b} \left( \sum_{k=j+1}^{m} p_k \right)^{b} \frac{p_i}{\sum_{k=1}^{j} p_k} (N-b) \\
&= (N-b) \sum_{i=1}^{j} \frac{p_i}{\sum_{k=1}^{j} p_k} Q_1(i, "1") \\
&= Q^{*}(b,b+1)
\end{align*}
\]
Thus the two approaches are identical.

Calculating \( Q_\gamma(b,b+1) \) for the aunt/niece case by the simpler formula yields

\[
Q_\gamma(b,b+1) = (N-b) \sum_{i=1}^{j} \frac{p_i}{\sum_{k=1}^{j} p_k} Q_1(i, "1")
\]

\[
= (N-b) \frac{1/8}{1/2} (2\lambda + 3\lambda + 2\lambda + 3\lambda)
\]

\[
= (N-b) \frac{5\lambda}{2}.
\]

Similarly, \( Q_\gamma(b,b-1) = 5\lambda/b/2. \) And \( \pi(1) \) for one pair equals 1/2, so \( \pi_\gamma(b) \) for \( N \) pairs is the binomial probability \( \left( \binom{N}{b} \right) \frac{1}{2^N} \). Then \( \alpha = 1 - \exp\left\{-\left[\left( \binom{N}{b} \right) \left( \frac{1}{2^N} \right) (2b-N) 5\lambda \right] \right\} \frac{1}{2} \).

The simulation results in Table 4 show that this approximation is again very good.
one chromosome (20,000 repetitions)  
$N = 100$  $\lambda = .01$  $l = 100$  

23 chromosomes  

p-values via  

<table>
<thead>
<tr>
<th>b</th>
<th>simulation</th>
<th>approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>.01840</td>
<td>.01957</td>
</tr>
<tr>
<td>68</td>
<td>.00865</td>
<td>.01010</td>
</tr>
<tr>
<td>69</td>
<td>.00415</td>
<td>.00496</td>
</tr>
<tr>
<td>70</td>
<td>.00215</td>
<td>.00231</td>
</tr>
<tr>
<td>71</td>
<td>.00090</td>
<td>.00103</td>
</tr>
</tbody>
</table>

p-values via  

<table>
<thead>
<tr>
<th>b</th>
<th>simulation</th>
<th>approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>.348</td>
<td>.365</td>
</tr>
<tr>
<td>68</td>
<td>.181</td>
<td>.208</td>
</tr>
<tr>
<td>69</td>
<td>.091</td>
<td>.108</td>
</tr>
<tr>
<td>70</td>
<td>.048</td>
<td>.052</td>
</tr>
<tr>
<td>71</td>
<td>.020</td>
<td>.023</td>
</tr>
</tbody>
</table>

Table 4.

7. **Cousins**

![Cousin Diagram]

Figure 6.
Figure 6 shows the cousins case, which is very similar to the aunt/niece case. The direct comparison process is not a Markov chain, but it can be described as a function of a Markov chain. The underlying chain has sixteen states, and is defined as follows. Let $X_t^{(1)}$ be the indicator of whether the two mothers match on their maternal chromosome (equivalent to a half-sibling process). Let $X_t^{(2)}$ be the indicator of whether they match on their paternal chromosome (another half-sibling process). Let $X_t^{(3)}$ be the grandparent/grandchild process for cousin #1 and the grandmother. And let $X_t^{(4)}$ be the grandparent/grandchild process for cousin #2 and the grandmother. The $Q$-matrix for $(X_t^{(1)}, X_t^{(3)}, X_t^{(3)}, X_t^{(4)})$ is

<table>
<thead>
<tr>
<th>state #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2λ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

The observed process, $Y_t$, is 1 whenever $(X_t^{(1)}, X_t^{(2)}, X_t^{(3)}, X_t^{(4)})$ is in state 1, 4, 5, or 12. So for $N$ pairs, $Y_t$ is the sum of the numbers of pairs in states 1, 4, 5, and 12.

Since this chain has sixteen states, it is only practical to find p-values by the simpler method that I introduced in the aunt/niece case. $\pi(1)$ for one pair equals 1/4, so $\pi(1)$ for $N$ pairs is $\binom{N}{b} \left( \frac{1}{4} \right)^b \left( \frac{3}{4} \right)^{N-b}$. And
\[ Q_Y(b, b+1) = (N-b) \sum_{i=1}^{j} \frac{p_i}{\sum p_k} Q_1(i,1) \]

\[ = (N-b) \frac{1/16}{3/4} (2\lambda + 2\lambda + \lambda + \lambda + 2\lambda + 2\lambda + \lambda + 2\lambda + 0\lambda + 0\lambda + 2\lambda) \]

\[ = (N-b) \frac{4\lambda}{3}. \]

Similarly, \( Q_Y(b, b-1) = 4\lambda/b \). Then
\[ \alpha = 1 - \exp\left[-1 \left( \binom{N}{b} \left( \frac{1}{4} \right)^b \left( \frac{3}{4} \right)^{N-b} \right) \right]. \]

Simulations indicate that the cousins approximation is still decently accurate. Table 5 shows these results.

<table>
<thead>
<tr>
<th>one chromosome (20,000 repetitions)</th>
<th>23 chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N = 100 ) \hspace{1cm} ( \lambda = 0.01 ) \hspace{1cm} ( l = 100 )</td>
<td>\hspace{1cm} p-values via</td>
</tr>
<tr>
<td>\hspace{1cm} b</td>
<td>\hspace{1cm} simulation</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------</td>
</tr>
<tr>
<td>41</td>
<td>.01350</td>
</tr>
<tr>
<td>42</td>
<td>.00655</td>
</tr>
<tr>
<td>43</td>
<td>.00305</td>
</tr>
<tr>
<td>44</td>
<td>.00135</td>
</tr>
<tr>
<td>45</td>
<td>.00060</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>\hspace{1cm} p-values via</th>
</tr>
</thead>
<tbody>
<tr>
<td>\hspace{1cm} b</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>45</td>
</tr>
</tbody>
</table>

Table 5.

8. RELATED PROBLEMS

I have applied these methods to more distant relatives, such as second cousins and first cousins once removed. The underlying chains get more complicated, but the methods stay the same. One can also do approximate power computations to get sample sizes and to compare the utility of different types of relative pairs. Another important issue is actually estimating the location of the gene once the test has concluded that one exists. This can be done by likelihood methods. More difficult problems include combining information from different types of relative pairs and using pedigrees that consist of more than two affected
relatives.

Another application of Brown's mapping technique is mapping rare recessive traits by looking at affected children of consanguineous marriages. I.e., instead of comparing the genetic material of two relatives to find regions of identity by descent, compare the two strands of DNA from one person to find regions of homozygosity by descent. The mathematical methods described in this paper can be extended in a very straightforward way to model such data.

REFERENCES


