The Application of DNA Fingerprints to the Reconstruction of Genealogies within Populations

Robert Gardner

Certificate of Approval:

Marie W. Wooten Assistant Professor Zoology and Wildlife Science	Michael C. Wooten, Chair Associate Professor Zoology and Wildlife Science
William N. Hudson Professor Algebra, Combinatorics and Analysis	Norman J. Doorenbos, Dean Graduate School

Style manual or journal used Evolution (together with the style known as "aums").

Computer software used The documentation package TEX(secifically LATEX) together with the style-file aums.sty.

The Application of DNA Fingerprints to the Reconstruction of Genealogies within Populations

Robert Gardner

A Thesis

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

Auburn, Alabama

March 20, 1992

THE APPLICATION OF DNA FINGERPRINTS TO THE RECONSTRUCTION OF GENEALOGIES WITHIN POPULATIONS

Robert Gardner

_____
Signature of Author

_____
Date

Copy sent to:

Name Date

VITA

Robert Bentley Gardner, Junior, son of Robert Gardner and Ruby (Gordon) Gardner, was born February 4, 1963, in Montgomery, Alabama. He graduated from Hooper Academy in Hope Hull, Alabama in 1981. In June, 1981, he entered Auburn University at Montgomery and received the degree of Bachelor of Science (Mathematics) in May, 1984. While an undergraduate, he worked at the Alabama Department of Environmental Management, Air Division. He entered the graduate school of Auburn University in September, 1984. He did intermittent work in biology and physics and taught numerous mathematics classes. He received the degree Master of Science (Mathematics-Combinatorics) in August, 1987. He continued graduate work at Auburn University concentrating on a Ph.D. in mathematics (Complex Analysis) which he received in August, 1991. He is currently employed by the Department of Mathematics at Louisiana State University in Shreveport. He is married to E. Standish Paul-Gardner and they have one son, Quincey Paul Gardner.

THESIS ABSTRACT

THE APPLICATION OF DNA FINGERPRINTS TO THE RECONSTRUCTION OF GENEALOGIES WITHIN POPULATIONS

Robert Gardner

Master of Science, March 20, 1992
(Ph.D., Auburn University, 1991)
(M.S., Auburn University, 1987)
(B.S., Auburn University at Montgomery, 1984)

109 Typed Pages

Directed by Michael C. Wooten

This thesis presents a probabilistic approach to the reconstruction of genealogies within populations. The technique uses data such as that provided by DNA fingerprints. Several conditional probabilities are derived which are used to put likelihoods on certain degrees of relationship between pairs of individuals. Simulated data is generated for several different genealogies and the model then takes the data and attempts to reconstruct the genealogy. The results of this maximum likelihood analysis are discussed in light of additional data, such as age structure. Finally, the strengths and weaknesses of this approach are presented.

Acknowledgments

The author would like to thank the Department of Algebra, Combinatorics and Analysis for the generous financial support which it provided while he was a graduate student in the Department of Zoology and Wildlife. He would also like to thank the Department of Physics for support in the summer of 1990 when a significant number of the results in this thesis were determined.

The author would also like to acknowledge Dr. Ann H. Williams for her encouragement in pursuing graduate work in the life sciences.

Finally, the author would like to express his thanks to Dr. Mike (Wooten). In addition to doing all the usual major professor stuff, he was just plain fun to work with!

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction and Background

1.1 Introduction

William Amos of Cambridge University is quoted as saying "DNA fingerprinting is *the* most useful technique to have been introduced into population biology" [Lewin 1989]. This statement was motivated by the idea of applying DNA fingerprinting to problems such as the one addressed here. The primary goal of this project is to use DNA fingerprints of individuals from a population to determine the relative relatedness of one to another. These values will then be used to break the population into related groups. The average relatedness within each group and within the sample population can then be calculated.

This problem is of interest for several reasons. For evolutionary biologists, the ability to backtrack to the ancestoral population would yield clues to who are the "fittest" individuals; that is, those individuals that contribute the most to subsequent gene pools. This information would be especially important in selective breeding programs in small populations or endangered species. Also, inbreeding could be controlled in such a situation if relatedness between individuals were known.

In the area of behavioral ecology, DNA fingerprinting can produce information not traditionally available. Relationships that were previously inferred from behavior, could be determined from genetic data. This determination would give unambiguous measures of reproductive success. In particular, cases of "infidelity" could be discovered and questions of paternity could be answered. In turn, questions of altruism and kin selection

could be addressed. Applications of this type have already appeared. The technique has been applied to a population of house sparrows [Wetton *et al.* 1987 and Burke and Bruford 1987] to detect relationships that otherwise would have gone undetected (a case of parent-offspring mating was detected as well as an extrapair copulation that resulted in an offspring).

Information of this type also has been used in civil and legal cases. The nature of these data allows the determination of certain relationships when intervening relatives are missing (for example, maternity can be exhibited in the absence of the father [Jeffreys *et al.* 1985]). A more direct application of DNA fingerprinting can be found in forensics, where blood, semen, or hair samples are matched to suspects [Gill *et al.* 1987].

The process of producing DNA "fingerprints" was first described by Jeffreys, Wilson, and Thein [1985a]. The chromosomal region adjacent to the centromere in higher eucaryotes is composed of very long blocks of highly repetitive DNA in which simple sequences, called repeat elements, are repeated thousands of times or more. These repeating sequences often have compositions different from most of the organism's other DNA and can be seperated by centrifuging fragmented DNA in a cesium chloride (CsCl) density gradient. DNA seperated by this process is called satellite DNA [see Wilson *et al.* 1987]. The smaller fragments, the *minisatellites*, are highly polymorphic due to variation in repeat unit length. Minisatellite-length-variation can be detected using restriction endonucleases. The digested DNA is then electrophoresed through agarose gel and transferred by blotting to a membrane. The membrane is hybridized with radioactively (^{32}P) labelled single-stranded DNA probes. After exposure to film, the resulting pattern is called a DNA fingerprint.

1.2 History of Applications of DNA Fingerprinting to Population Genetics

Jeffreys, Wilson, and Thein [1985b] were the first to present population genetic type calculations for DNA fingerprint data. DNA fingerprints were made for 20 unrelated British caucasians and pairwise comparisons were made as to the presence or absence of bands. The probability of the presence of a band in an individual is said to be $p^2 + 2p(1-p) = 2p - p^2$ where p is the frequency of the allele producing that band. This indicates that the presence of a band is behaving as a dominant trait would behave in classical Mendelian genetics. As with any dominant trait, band presence does not indicate whether an individual is homozygous dominant or heterozygous at the locus in question. With estimates of band frequencies, Jeffreys *et al.* [1985b] proceeded to estimate the individual specificity of a DNA fingerprint. Using the data from two probes, they calculated probabilities of unrelated individuals having the same DNA fingerprint to be 5×10^{-19} (my recalculation from their data actually puts this probability at 5×10^{-21}), indicating that this method is indeed individual specific.

Jeffreys, Brookfield, and Semeonoff [1985] presented the following example of the use of DNA fingerprints in determining relationships: "The case concerned a Ghanaian boy born in the United Kingdom who emigrated to Ghana to join his father and subsequently returned alone to the United Kingdom to be reunited with his mother, brother and two sisters. However, there was evidence to suggest that a substitution might have occured, either for an unrelated boy, or a son of a sister of the mother... As a result, the returning boy was not granted residence in the United Kingdom." Conventional genetic markers indicated that the woman and boy were related (with 99% probability), but could not determine whether the woman was the boy's mother or aunt. DNA

fingerprints were produced from blood DNA samples taken from the boy, the mother, her three other children, and an unrelated individual. The father was unavailable. Based on the probability of unrelated individuals sharing a band, the allele frequency for a band was calculated to be $p = 0.14$ (all bands were assumed to have the same frequency). The mother and boy were found to share 25 maternal specific bands. Based on this, it was calculated that the probability of these two being unrelated was $(0.26)^{25} = 2 \times 10^{-15}$ (an allele with frequency 0.14 will appear in an individual with probability $2(0.14) - (0.14)^2 = 0.26$). The corresponding probability of the mother actually being the aunt of the boy was said to be 6×10^{-6}. This latter calculation, however, was found to be erroneous and declared irrelevant by Hill [1986] who approached this same problem from a maximum likelihood viewpoint but reached the same final conclusion. The boy was granted residence in the United Kingdom. The method used by Jeffreys *et al.* [1985] is very restrictive and not suitable for general use in testing specified relationships.

The first to present a detailed account of the use of DNA fingerprinting in the estimation of relatedness was Lynch [1988]. He showed that the proportion of shared bands is a poor estimate of relatedness unless the frequencies of the bands are near zero [Lynch 1988, Figure 2]. This is not surprising since a high frequency band would be present in significant numbers of unrelated individuals. So it is necessary to make a compensation in the probability of shared bands for different band frequencies and degrees of relationship (band frequencies in the above references ranged from 0.08 to 0.28). Lynch proposed the following two equations for the relatedness of individual 'B' to individual 'A':

$$\hat{r}_{BA} = \frac{S_{BA} - \theta_{1B}}{1 - \theta_{1B}}$$

where S_{BA} is the observed similarity and θ_{1B} is the expected similarity of B to nonrelatives, and

$$\hat{r}_{BA} \simeq \frac{S_{BA} - \hat{\theta}_1}{1 - \hat{\theta}_1} + \frac{(1 - S_{BA})\mathrm{Var}(\theta_1)}{(1 - \hat{\theta}_1)^3}$$

where $\hat{\theta}_1$ is an estimate of the average similarity between unrelated individuals. Lynch discussed problems with both of these; where his primary concern was with determination of θ_1 and θ_{1B}. I will use a method that is somewhat similar to this which requires the computation of several conditional probabilites.

An additional problem addressed by Lynch was the fact that the variance of the estimate of similarity between unrelated individuals may produce difficulties, in particular in the determination of relatively distant relationships. He concluded that "beyond (and often including) second-degree relationships, DNA fingerprinting does not provide a powerful means of assessing individual relationships." I believe that this problem may be addressed with accurate estimates of band frequencies estimated from large samples, and sufficient numbers of closely related individuals to fill in ambiguities due to poorly determined distant relationships.

1.3 History of the Reconstruction of Genealogies

The study of the construction of phylogenetic trees from genetic data is extensive [see, for example, Hillis 1987 and Pamilo and Nei 1988]. Felsenstein has been particularly active in the quantitative aspects of this problem [see Felsenstein 1981, 1982, and 1983]. However, this problem is only distantly related to the problem at hand, namely the reconstruction of genealogies. One facet of the phylogenetic tree problem that does

carry over is the use of maximum likelihood estimates [Felsenstein 1981 and 1983]. Although the literature on genealogical reconstruction is rather limited [see Cannings and Thompson 1981 and Thompson 1986], the idea of maximum likelihood is always present.

Thompson [1986, Chapter 3] gave an algorithm and example of estimating relationships and reconstructing genealogies. The model was based on genotypic data and maximum likelihood estimates were generated based on the number of alleles shared at a locus (0, 1 or 2) conditional on various relationships. It was commented that remote relationships are difficult to distinguish from one another and that genealogies will most likely have to be rebuilt from the "most readily detected relationships, such as parent-offspring or sib" [p. 55]. This is especially true with inbred populations, where the number of possible sources for a given allele makes other approaches impractical. One problem with this approach is the accurate determination of allele frequencies. Fortunately, it was said [p. 51], the estimation of relationships is not sensitive to small variations in allele frequencies.

Thompson [1986, Chapter 4] also presented methods for putting probabilities (likelihoods) on pedigrees. The key expression was:

P(observed phenotypes | known inheritance model & hypothesized genealogy).

In a pedigree with no inbreeding, each intermediate individual (one with both parents and offspring in the pedigree) will cut the *tree* into two *components*; a component containing the parents and a component containing the offspring (in graph theoretic terms, the vertex of the tree corresponding to this individual would be called a *cut vertex*, see Bondy and Murty [1976] for graph theory definitions). Thompson calls such an individual, X, a *pivot* individual and refers to the component of the pedigree containing X's parents

as the "before" component and the other component as the "after" component. The following probability was also defined:

$$\mathrm{L}_X(i) = \mathrm{P}(\text{data after X} \mid \text{X has genotype } i).$$

This is the likelihood for the genotype of X, given the genetic data on individuals following him. It is possible to work up a genealogy; at each stage the contributions from individuals after X in the genealogy are combined into the terms of L_X for each pivot individual X in turn [p. 86]. If there is inbreeding, then the pedigree will no longer be a tree and will contain loops. So it will no longer contain pivot individuals (cut vertices), but will contain "cutsets" of individuals, whose removal from the pedigree break it into two components. With the collection of cutsets, it is again possible to work through a genealogy sequentially, as above [p. 94].

The computation of ancestoral likelihoods and founder genotypes has also been explored by Thompson [1986, Chapter 4]. The likelihood for given founder genotypes is

$$L_{\text{founders}}(\text{founder genotype set}) = \mathrm{P}(\text{observed data} \mid \text{founder genotypes}).$$

This likelihood depends on a known genealogy. One computation gives a likelihood over all founder combinations and is independent of allele frequencies.

Once a genealogy has been established, average relatedness values are easily computed. The process was described by Wright [1922]. If the individuals concerned are not inbreed, then the length of the path from one to another determines the "coefficient of relatedness" or the "coefficient of kinship". If they are inbreed, then the number and length of the paths from one to another determines this value. In either case, if the genealogy is known, this is a minor problem.

1.4 Objectives

The main objective of this research was to develop mathematical procedures for the use of DNA fingerprints in the reconstruction of genealogies. In addition to this, conclusions were drawn from the results concerning statistical parameters of the population, including probable ancestory. My objectives were:

1. develop methods appropriate for the pairwise determination of relationships using DNA fingerprint data,

2. develop methods for the seperation of sample groups into groups of related individuals,

3. determine the most probable genealogy for each group of related individuals, and

4. determine average relatedness within each group and average relatedness within the entire sample.

Chapter 2

Methods

2.1 Calculation of some Conditional Probabilities

If a population is in Hardy-Weinberg equilibrium for a given trait, then certain conditional probabilites concerning the presence or absence of the trait in individuals that are related can be calculated. For example, if a certain individual demonstrates a trait, the probability that his offspring, sibling, cousin, etc. also shows this trait can be calculated.

Consider a dominate trait with allele frequency p. For an individual X, denote the homozygous dominant state as X_1, the heterozygous state as X_2 and the homozygous recessive state as X_3. The different states of X have the following probabilities: $P(X_1) = p^2$, $P(X_2) = 2p - 2p^2$, $P(X_3) = 1 - 2p + p^2$. However, with a dominant trait, it is unlikely that heterozygous and a homozygous dominant individuals can be distinguished. So denote the presence of the trait as $X-$ and $P(X-) = P(X_1) + P(X_2) = 2p - p^2$.

Similar calculations to determine the probabilities of a dominant trait appearing in an offspring of an individual can be made. Denote the known individual as M and the offspring as D. For this calculation, the other parent of D, say F, must be considered. First, the probability of the genotype of the parent F must be calculated. Next, one can calculate the probabilities for the different genotypes of D given M_i and F_j. Then the probability of each genotype of D given the genotype of M can be derived. With the values in Table 2.1 the probability of each possible genotype of D given any genotype of

M can be calculated as follows:

$$\begin{aligned}
P(D_1|M_1) &= \sum_j P(F_j)P(D_1|M_1 \text{ and } F_j) = p, \\
P(D_1|M_2) &= \sum_j P(F_j)P(D_1|M_2 \text{ and } F_j) = \frac{p}{2}, \\
P(D_1|M_3) &= \sum_j P(F_j)P(D_1|M_3 \text{ and } F_j) = 0, \\
P(D_2|M_1) &= \sum_j P(F_j)P(D_2|M_1 \text{ and } F_j) = 1-p, \\
P(D_2|M_2) &= \sum_j P(F_j)P(D_2|M_2 \text{ and } F_j) = \frac{1}{2}, \\
P(D_2|M_3) &= \sum_j P(F_j)P(D_2|M_3 \text{ and } F_j) = p, \\
P(D_3|M_1) &= \sum_j P(F_j)P(D_3|M_1 \text{ and } F_j) = 0, \\
P(D_3|M_2) &= \sum_j P(F_j)P(D_3|M_2 \text{ and } F_j) = \frac{1-p}{2}, \text{ and} \\
P(D_3|M_3) &= \sum_j P(F_j)P(D_3|M_3 \text{ and } F_j) = 1-p.
\end{aligned}$$

But here, again, there is only concern with the presence or absence of the trait and the following are obtained:

$$\begin{aligned}
P(D-|M-) &= P((D_1 \text{ or } D_2)|(M_1 \text{ or } M_2)) \\
&= \frac{P((D_1 \text{ or } D_2) \text{ and } (M_1 \text{ or } M_2))}{P(M_1 \text{ or } M_2)} \\
&= \frac{P((D_1 \text{ and } M_1) \text{ or } (D_1 \text{ and } M_2) \text{ or } (D_2 \text{ and } M_1) \text{ or } (D_2 \text{ and } M_2))}{P(M_1 \text{ or } M_2)} \\
&= \frac{P(M_1)[P(D_1|M_1)+P(D_2|M_1)]+P(M_2)[P(D_1|X_2)+P(D_2|M_2)]}{P(M_1)+P(M_2)} \\
&= \frac{p^2[p+(1-p)]+2p(1-p)[\frac{p}{2}+\frac{1}{2}]}{p^2+2p(1-p)}
\end{aligned}$$

Table 2.1: Conditional probability of genotype of offspring D given the genotypes of the parents M and F.

M_i	F_j	$P(D_1 \mid M_i \text{ and } F_j)$	$P(D_2 \mid M_i \text{ and } F_j)$	$P(D_3 \mid M_i \text{ and } F_j)$
M_1	F_1	1	0	0
M_1	F_2	$\frac{1}{2}$	$\frac{1}{2}$	0
M_1	F_3	0	1	0
M_2	F_1	$\frac{1}{2}$	$\frac{1}{2}$	0
M_2	F_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
M_2	F_3	0	$\frac{1}{2}$	$\frac{1}{2}$
M_3	F_1	0	1	0
M_3	F_2	0	$\frac{1}{2}$	$\frac{1}{2}$
M_3	F_3	0	0	1

$$
\begin{aligned}
&= \frac{1+p-p^2}{2-p}, \\
P(D-|M_3) &= P(D_1 \text{ or } D_2|M_3)) \\
&= P(D_1|M_3) + P(D_2|M_3) \\
&= 0 + p \\
&= p, \\
P(D_3|M-) &= \frac{P(M_1)P(D_3|M_1) + P(M_2)P(D_3|M_2)}{P(M_1) + P(M_2)} \\
&= \frac{(p^2)(0) + 2p(1-p)[\frac{1-p}{2}]}{2p-p^2} \\
&= \frac{1-2p+p^2}{2-p}, \text{ and} \\
P(D_3|M_3) &= 1-p.
\end{aligned}
$$

Now, consider the probability of the presence of a trait in a parent M given the presence of the trait in the offspring D. In this case, also, the other parent F must be considered. Again the probability of the genotypes of the parents, M_i and F_j, given the genotype of D, and the probabilities of the different states of D, given the different possible genotypes of M and F (see Tables 2.2 and 2.3) are calculated. As before, the following are obtained:

$$
\begin{aligned}
P(M_1|D_1) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = p, \\
P(M_1|D_2) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = \frac{p}{2}, \\
P(M_1|D_3) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = 0,
\end{aligned}
$$

$$
\begin{aligned}
P(M_2|D_1) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = 1-p, \\
P(M_2|D_2) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = \frac{1}{2}, \\
P(M_2|D_3) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = p, \\
P(M_3|D_1) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = 0, \\
P(M_3|D_2) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = \frac{1-p}{2}, \text{ and} \\
P(M_3|D_3) &= \sum_j P(F_j)P(M_1 \text{ and } F_j|D_1) = 1-p.
\end{aligned}
$$

Notice that these are the same conditional probabilities as obtained from Table 2.1. And so, as before:

$$
\begin{aligned}
P(M-|D-) &= \frac{1+p-p^2}{2-p}, \\
P(M-|D_3) &= p, \\
P(M_3|D-) &= \frac{1-2p+p^2}{2-p}, \text{ and} \\
P(M_3|D_3) &= 1-p.
\end{aligned}
$$

These are the same probabilities as above, as expected, since these two cases are symmetric.

Now, for the calculation of the same type probabilities for two individuals that are siblings, say S^1 and S^2, the values from Tables 2.1 and 2.2 are employed. This leads to

Table 2.2: Conditional probabilities of parental genotypes, M_i and F_j, given the genotype of an offspring, D.

M_i	F_j	$P(M_i \text{ and } F_j \mid D_1)$	$P(M_i \text{ and } F_j \mid D_2)$	$P(M_i \text{ and } F_j \mid D_3)$
M_1	F_1	p^2	0	0
M_1	F_2	$p(1-p)$	$p^2/2$	0
M_1	F_3	0	$p(1-p)/2$	0
M_2	F_1	$p(1-p)$	$p^2/2$	0
M_2	F_2	$(1-p)^2$	$p(1-p)$	p^2
M_2	F_3	0	$p(1-p)$	$(1-p)^2/2$
M_3	F_1	0	$p(1-p)/2$	0
M_3	F_2	0	$(1-p)^2/2$	$p(1-p)$
M_3	F_3	0	0	$(1-p)^2$

where :

$$P(M_i \text{ and } F_j \mid D_1) = \frac{P(M_i \text{ and } F_j)P(D_3 \mid M_i \text{ and } F_j)}{P(D_1)},$$

$$P(M_i \text{ and } F_j \mid D_2) = \frac{P(M_i \text{ and } F_j)P(D_2 \mid M_i \text{ and } F_j)}{P(D_2)}, \text{ and}$$

$$P(M_i \text{ and } F_j \mid D_3) = \frac{P(M_i \text{ and } F_j)P(D_3 \mid M_i \text{ and } F_j)}{P(D_3)}.$$

Table 2.3: Conditional probability of the genotype of an individual, S_j^2, given the genotype of a sibling, S^1.

S_i^2	$P(S_j^2\|S_1^1)$	$P(S_j^2\|S_2^1)$	$P(S_j^2\|S_3^1)$
S_1^2	$\frac{1+2p+p^2}{4}$	$\frac{p+p^2}{4}$	$\frac{p^2}{4}$
S_2^2	$\frac{1-p^2}{2}$	$\frac{1+p-p^2}{2}$	$\frac{2p-p^2}{2}$
S_3^2	$\frac{1-2p+p^2}{4}$	$\frac{2-3p+p^2}{4}$	$\frac{4-4p+p^2}{4}$

the following calculations:

$$
\begin{aligned}
P(S_1^1|S_1^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_1^2)P(S_1^1|M_i \text{ and } F_j) \\
&= \frac{4p^2(1-p)^2}{4p^2}\left(\frac{1}{4}\right) + \frac{2p^3(1-p)}{2p^2}\left(\frac{1}{2}\right) + \frac{2p^3(1-p)}{2p^2}\left(\frac{1}{2}\right) + \frac{p^4}{p^2}(1) \\
&= \frac{1+2p+p^2}{4}, \\
P(S_1^1|S_2^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_2^2)P(S_1^1|M_i \text{ and } F_j) \\
&= \frac{4p^2(1-p)^2}{2\times 2p(1-p)}\left(\frac{1}{4}\right) + \frac{2p^3(1-p)}{2\times 2p(1-p)}\left(\frac{1}{2}\right) + \frac{2p^3(1-p)}{2\times 2p(1-p)}\left(\frac{1}{2}\right) \\
&= \frac{p+p^2}{4}, \\
P(S_1^1|S_3^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_3^2)P(S_1^1|M_i \text{ and } F_j) \\
&= \frac{4p^2(1-p)^2}{4(1-p)^2}\left(\frac{1}{4}\right) \\
&= \frac{p^2}{4},
\end{aligned}
$$

$$
\begin{aligned}
P(S_2^1|S_1^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_1^2)P(S_2^1|M_i \text{ and } F_j) \\
&= \frac{4p^2(1-p)^2}{4p^2}\left(\frac{1}{2}\right) + \frac{2p^3(1-p)}{2p^2}\left(\frac{1}{2}\right) + \frac{2p^3(1-p)}{2p^2}\left(\frac{1}{2}\right) \\
&= \frac{1-p^2}{2},
\end{aligned}
$$

$$
\begin{aligned}
P(S_2^1|S_2^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_2^2)P(S_2^1|M_i \text{ and } F_j) \\
&= \frac{2p(1-p)^3}{2\times 2p(1-p)}\left(\frac{1}{2}\right) + \frac{p^2(1-p)^2}{2p(1-p)}(1) + \frac{2p(1-p)^3}{2\times 2p(1-p)}\left(\frac{1}{2}\right) \\
&\quad + \frac{4p^2(1-p)^2}{2\times 2p(1-p)}\left(\frac{1}{2}\right) + \frac{2p^3(1-p)}{2\times 2p(1-p)}\left(\frac{1}{2}\right) + \frac{p^2(1-p)^2}{2p(1-p)}(1) \\
&\quad + \frac{2p^3(1-p)}{2\times 2p(1-p)}\left(\frac{1}{2}\right) \\
&= \frac{1+p-p^2}{2},
\end{aligned}
$$

$$
\begin{aligned}
P(S_2^1|S_3^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_3^2)P(S_2^1|M_i \text{ and } F_j) \\
&= \frac{2p(1-p)^3}{2(1-p)^2}\left(\frac{1}{2}\right) + \frac{2p(1-p)^3}{2p(1-p)^2}\left(\frac{1}{2}\right) + \frac{4p^2(1-p)^2}{4(1-p)^2}\left(\frac{1}{2}\right) \\
&= \frac{2p-p^2}{2},
\end{aligned}
$$

$$
\begin{aligned}
P(S_3^1|S_1^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_1^2)P(S_3^1|M_i \text{ and } F_j) \\
&= \frac{4p^2(1-p)^2}{4p^2}\left(\frac{1}{4}\right) \\
&= \frac{1-2p+p^2}{4},
\end{aligned}
$$

$$
\begin{aligned}
P(S_3^1|S_2^2) &= \sum_{i,j} P(M_i \text{ and } F_j|S_2^2)P(S_3^1|M_i \text{ and } F_j) \\
&= \frac{2p(1-p)^3}{2\times 2p(1-p)}\left(\frac{1}{2}\right) + \frac{2p(1-p)^3}{2\times 2p(1-p)}\left(\frac{1}{2}\right) + \frac{4p^2(1-p)^2}{2\times 2p(1-p)}\left(\frac{1}{4}\right) \\
&= \frac{2-3p+p^2}{4}, \text{ and}
\end{aligned}
$$

$$
P(S_3^1|S_3^2) = \sum_{i,j} P(M_i \text{ and } F_j|S_3^2)P(S_3^1|M_i \text{ and } F_j)
$$

$$= \frac{(1-p)^4}{(1-p)^2}(1) + \frac{2p(1-p)^3}{2(1-p)^2}\left(\frac{1}{2}\right) + \frac{2p(1-p)^3}{2(1-p)^2}\left(\frac{1}{2}\right) + \frac{4p^2(1-p)^2}{4(1-p)^2}\left(\frac{1}{4}\right)$$
$$= \frac{4-4p+p^2}{4}.$$

And so for siblings:

$$\begin{aligned}
P(S^1- \mid S^2-) &= \frac{P(S_1^2)[P(S_1^1|S_1^2)+P(S_2^1|S_1^2)] + P(S_2^2)[P(S_1^1|S_2^2)+P(S_2^1|S_2^2)]}{P(S_1^2)+P(S_2^2)} \\
&= \frac{p^2[\frac{1+2p+p^2}{4}+\frac{1-p^2}{2}] + 2p(1-p)[\frac{p+p^2}{4}+\frac{1+p-p^2}{2}]}{2p-p^2} \\
&= \frac{4+5p-6p^2+p^3}{4(2-p)}, \qquad (2.1) \\
P(S_3^1|S^2-) &= \frac{P(S_1^2)P(S_3^2|S_1^2)+P(S^2)P(S_3^1|S_2^2)}{P(S_1^2)+P(S_2^2)} \qquad (2.2) \\
&= \frac{(p^2)\frac{1-2p+p^2}{4} + 2p(1-p)\frac{2-3p+p^2}{4}}{2p-p^2} \\
&= \frac{4-9p+6p^2-p^3}{4(2-p)}, \\
P(S^1- \mid S_3^2) &= P(S_1^1|S_3^2)+P(S_2^1|S_3^2) \\
&= \frac{p^2}{4}+\frac{2p-p^2}{2} \\
&= \frac{4p-p^2}{4}, \text{ and} \\
P(S_3^1|S_3^2) &= \frac{4-4p+p^2}{4}.
\end{aligned}$$

In fact, equation 2.1 has already appeared [Jeffreys, Brookfield and Semeonoff 1985].

Now I show that the calculations of the probabilities will allow a commuting of the sibling step with the parent/offspring steps by showing that the numbers are the same for an individual's niece/nephew as they are for the individual's aunt/uncle. Again,

denote the known individual as A and the niece/nephew as N. This yields the following (steping through the intermediate relative S, a sibling of A):

$$
\begin{aligned}
P(N_1|A_1) &= P(N_1|S_1)P(S_1|A_1) + P(N_1|S_2)P(S_2|A_1) + P(N_1|S_3)P(S_3|A_1) \\
&= (p)\frac{1+2p+p^2}{4} + \left(\frac{p}{2}\right)\frac{1-p^2}{2} + (0)\frac{1-2p+p^2}{4} \\
&= \frac{p+p^2}{2}, \\
P(N_1|A_2) &= P(N_1|S_1)P(S_1|A_2) + P(N_1|S_2)P(S_2|A_2) + P(N_1|S_3)P(S_3|A_2) \\
&= (p)\frac{p+p^2}{4} + \left(\frac{p}{2}\right)\frac{1+p-p^2}{2} + (0)\frac{2-3p+p^2}{4} \\
&= \frac{p+2p^2}{4}, \\
P(N_1|A_3) &= P(N_1|S_1)P(S_1|A_3) + P(N_1|S_2)P(S_2|A_3) + P(N_1|S_3)P(S_3|A_3) \\
&= (p)\frac{p^2}{4} + \left(\frac{p}{2}\right)\frac{2p-p^2}{2} + (0)\frac{4-4p+p^2}{4} \\
&= \frac{p^2}{2}, \\
P(N_2|A_1) &= P(N_2|S_1)P(S_1|A_1) + P(N_2|S_2)P(S_2|A_1) + P(N_2|S_3)P(S_3|A_1) \\
&= (1-p)\frac{1+2p+p^2}{4} + \left(\frac{1}{2}\right)\frac{1-p^2}{2} + (p)\frac{1-2p+p^2}{4} \\
&= \frac{1+p-2p^2}{2}, \\
P(N_2|A_2) &= P(N_2|S_1)P(S_1|A_2) + P(N_2|S_2)P(S_2|A_2) + P(N_2|S_3)P(S_3|A_2) \\
&= (1-p)\frac{p+p^2}{4} + \left(\frac{1}{2}\right)\frac{1+p-p^2}{2} + (p)\frac{2-3p+p^2}{4} \\
&= \frac{1+4p-4p^2}{4}, \\
P(N_2|A_3) &= P(N_2|S_1)P(S_1|A_3) + P(N_2|S_2)P(S_2|A_3) + P(N_2|S_3)P(S_3|A_3) \\
&= (1-p)\frac{p^2}{4} + \left(\frac{1}{2}\right)\frac{2p-p^2}{2} + (p)\frac{4-4p+p^2}{4}
\end{aligned}
$$

$$= \frac{3p-2p^2}{2},$$

$$\begin{aligned} P(N_3|A_1) &= P(N_3|S_1)P(S_1|A_1) + P(N_3|S_2)P(S_2|A_1) + P(N_3|S_3)P(S_3|A_3) \\ &= (0)\frac{1+2p+p^2}{4} + \left(\frac{1-p}{2}\right)\left(\frac{1-p^2}{2}\right) + (1-p)\frac{1-2p+p^2}{4} \\ &= \frac{1-2p+p^2}{2}, \end{aligned}$$

$$\begin{aligned} P(N_3|A_2) &= P(N_3|S_1)P(S_1|A_2) + P(N_3|S_2)P(S_2|A_2) + P(N_3|S_3)P(S_3|A_2) \\ &= (0)\frac{p+p^2}{4} + \frac{1-p}{2}\frac{1+p-p^2}{2} + (1-p)\frac{2-3p+p^2}{4} \\ &= \frac{3-5p+2p^2}{4}, \text{ and} \end{aligned}$$

$$\begin{aligned} P(N_3|A_3) &= P(N_3|S_1)P(S_1|A_3) + P(N_3|S_2)P(S_2|A_3) + P(N_3|S_3)P(S_3|AA_3) \\ &= (0)\frac{p^2}{4} + \frac{1-p}{2}\frac{2p-p^2}{2} + (1-p)\frac{4-4p+p^2}{4} \\ &= \frac{2-3p+p^2}{2}. \end{aligned}$$

And so:

$$\begin{aligned} P(N-|A-) &= \frac{P(A_1)[P(N_1|A_1) + P(N_2|A_1)] + P(A_2)[P(N_1|A_2) + P(N_2|A_2)]}{P(A_1) + P(A_2)} \\ &= \frac{p^2[\frac{p+p^2}{2} + \frac{1+p-2p^2}{2}] + 2p(1-p)[\frac{p+2p^2}{4} + \frac{1+4p-4p^2}{4}]}{2p-p^2} \\ &= \frac{1+5p-5p^2+p^3}{2(2-p)}, \\ P(N_3|A-) &= \frac{P(A_2)P(N_3|A_1) + P(A_2)P(N_3|A_2)}{P(A_1) + P(A_2)} \\ &= \frac{(p^2)\frac{1-2p+p^2}{2} + 2p(1-p)\frac{3-5p+2p^2}{4}}{2p-p^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{3-7p+5p^2-p^3}{2(2-p)}, \\
P(N-|A_3) &= P(N_1|A_3)+P(N_2|A_3) \\
&= \frac{p^2}{2}+\frac{3p-2p^2}{2} \\
&= \frac{3p-p^2}{2}, \text{ and} \\
P(N_3|A_3) &= \frac{2-3p+p^2}{2}.
\end{aligned}
$$

Now, the same probabilities for an aunt/uncle of N, denoted A, are calculated (stepping through the intermediate relative S, a parent of N):

$$
\begin{aligned}
P(A_1|N_1) &= P(A_1|S_1)P(S_1|N_1)+P(A_1|S_2)P(S_2|N_1) \\
&\quad +P(A_1|S_3)P(S_3|N_1) \\
&= \frac{1+2p+p^2}{4}(p)+\frac{p+p^2}{4}(1-p)+\frac{p^2}{2}(0) \\
&= \frac{p+p^2}{2}, \\
P(A_1|N_2) &= P(A_1|S_1)P(S_1|N_2)+P(A_1|S_2)P(S_2|N_2) \\
&\quad +P(A_1|S_3)P(S_3|N_2) \\
&= \frac{1+2p+p^2}{4}\left(\frac{p}{2}\right)+\frac{p+p^2}{4}\left(\frac{1}{2}\right)+\left(\frac{p^2}{4}\right)\frac{1-p}{2} \\
&= \frac{p+2p^2}{4}, \\
P(A_1|N_3) &= P(A_1|S_1)P(S_1|N_3)+P(A_1|S_2)P(S_2|N_3) \\
&\quad +P(A_1|S_3)P(S_3|N_3)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1+2p+p^2}{4}(0) + \frac{p+p^2}{4}(p) + \frac{p^2}{4}(1-p) \\
&= \frac{p^2}{2}, \\
P(A_2|N_1) &= P(A_2|S_1)P(S_3|N_1) + P(A_1|S_2)P(S_2|N_1) \\
&\quad + P(A_2|S_3)P(S_3|N_1) \\
&= \frac{1-p^2}{2}(p) + \frac{1+p-p^2}{2}(1-p) + \frac{1-p^2}{2}(0) \\
&= \frac{1+p-2p^2}{2}, \\
P(A_2|N_2) &= P(A_2|S_1)P(S_1|N_2) + P(A_2|S_2)P(S_2|N_2) \\
&\quad + P(A_2|S_3)P(S_3|N_2) \\
&= \frac{1-p^2}{2}\left(\frac{p}{2}\right) + \frac{1+p-p^2}{2}\left(\frac{1}{2}\right) + \frac{2p-p^2}{2}\left(\frac{1-p}{2}\right) \\
&= \frac{1+4p-4p^2}{4}, \\
P(A_2|N_3) &= P(A_2|S_1)P(S_1|N_3) + P(A_2|S_2)P(S_2|N_3) \\
&\quad + P(A_2|S_3)P(S_3|N_3) \\
&= \frac{1-p^2}{2}(0) + \frac{1+p-p^2}{2}(p) + \frac{2p-p^2}{2}(1-p) \\
&= \frac{3p-2p^2}{2}, \\
P(A_3|N_1) &= P(A_3|S_1)P(S_1|N_1) + P(A_3|S_2)P(S_2|N_1) \\
&\quad + P(A_3|S_3)P(S_3|N_1) \\
&= \frac{1-2p+p^2}{4}(p) + \frac{2-3p+p^2}{4}(1-p) + \frac{4-4p+p^2}{4}(0) \\
&= \frac{1-2p+p^2}{2}, \\
P(A_3|N_2) &= P(A_3|S_1)P(S_1|N_2) + P(A_3|S_2)P(S_2|N_2) \\
&\quad + P(A_3|S_3)P(S_3|N_2)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1-2p+p^2}{4}\left(\frac{p}{2}\right) + \frac{2-3p+p^2}{4}\left(\frac{1}{2}\right) + \frac{4-4p+p^2}{4}\left(\frac{1-p}{2}\right) \\
&= \frac{3-5p+2p^2}{4}, \text{ and} \\
P(A_3|N_3) &= P(A_3|S_1)P(S_1|N_3) + P(A_3|S_2)P(S_2|N_3) \\
&\quad + P(A_3|S_3)P(S_3|N_3) \\
&= \frac{1-2p+p^2}{4}(0) + \frac{2-3p+p^2}{4}(p) + \frac{4-4p+p^2}{4}(1-p) \\
&= \frac{2-3p+p^2}{2}.
\end{aligned}
$$

Notice that these are the same as with the niece/nephew, and so:

$$
\begin{aligned}
P(A-|X-) &= \frac{1+5p-5p^2+p^3}{2(2-p)}, \\
P(A_3|X-) &= \frac{3-7p+5p^2-p^3}{2(2-p)}, \\
P(A-|N_3) &= \frac{3p-p^2}{2}, \text{ and} \\
P(A_3|N_3) &= \frac{2-3p+p^2}{2}.
\end{aligned}
$$

Now, the same probabilities for a grandparent, G, of X are calculated passing through the parent, M, of X:

$$
P(G_1|X_1) = P(G_1|M_1)P(M_1|X_1) + P(G_1|M_2)P(M_2|X_1)
$$

$$
\begin{aligned}
& & & +P(G_1|M_3)P(M_3|X_1) \\
& & = & (p)(p) + \frac{p}{2}(1-p) + (0)(0) \\
& & = & \frac{p+p^2}{2}, \\
P(G_1|X_2) & & = & P(G_1|M_1)P(M_1|X_2) + P(G_1|M_2)P(M_2|X_2) \\
& & & +P(G_1|M_3)P(M_3|X_2) \\
& & = & (p)\left(\frac{p}{2}\right) + \left(\frac{p}{2}\right)\left(\frac{1}{2}\right) + (0)\frac{1-p}{2} \\
& & = & \frac{p+2p^2}{4}, \\
P(G_1|X_3) & & = & P(G_1|M_1)P(M_1|X_3) + P(G_1|M_2)P(M_2|X_3) \\
& & & +P(G_1|M_3)P(M_3|X_3) \\
& & = & (p)(0) + \frac{p}{2}(p) + (0)(1-p) \\
& & = & \frac{p^2}{2}, \\
P(G_2|X_1) & & = & P(G_2|M_1)P(M_1|X_1) + P(G_2|M_2)P(M_2|X_1) \\
& & & +P(G_2|M_3)P(M_3|X_1) \\
& & = & (1-p)p + \frac{1}{2}(1-p) + (p)(0) \\
& & = & \frac{1+p-2p^2}{2}, \\
P(G_2|X_2) & & = & P(G_2|M_1)P(M_1|X_2) + P(G_2|M_2)P(M_2|X_2) \\
& & & +P(G_2|M_3)P(M_3|X_2) \\
& & = & (1-p)\frac{p}{2} + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + (p)\frac{1-p}{2} \\
& & = & \frac{1+4p-4p^2}{4}, \\
P(G_2|X_3) & & = & P(G_2|M_1)P(M_1|X_3) + P(G_2|M_2)P(M_2|X_3) \\
& & & +P(G_2|M_3)P(M_3|X_3)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-p)(0) + \frac{1}{2}(p) + p(1-p) \\
&= \frac{3p - 2p^2}{2}, \\
P(G_3|X_1) &= P(G_3|M_1)P(M_1|X_1) + P(G_3|M_2)P(M_2|X_1) \\
&\quad + P(G_3|M_3)P(M_3|X_1) \\
&= (0)(p) + \frac{1-p}{2}(1-p) + (1-p)(0) \\
&= \frac{1 - 2p + p^2}{2}, \\
P(G_3|X_2) &= P(G_3|M_1)P(M_1|X_2) + P(G_3|M_2)P(M_2|X_2) \\
&\quad + P(G_3|M_3)P(M_3|X_2) \\
&= (0)\frac{p}{2} + \frac{1-p}{2}\left(\frac{1}{2}\right) + (1-p)\frac{1-p}{2} \\
&= \frac{3 - 5p + 2p^2}{4}, \text{ and} \\
P(G_3|X_3) &= P(G_3|M_1)P(M_1|X_3) + P(G_3|M_2)P(M_2|X_3) \\
&\quad + P(G_3|M_3)P(M_3|X_3) \\
&= (0)(0) + \frac{1-p}{2}(p) + (1-p)(1-p) \\
&= \frac{2 - 3p + p^2}{2}.
\end{aligned}
$$

These are the same as with the niece/nephew and aunt/uncle probabilities, so:

$$
\begin{aligned}
P(G-|X-) &= \frac{1 + 5p - 5p^2 + p^3}{2(2-p)}, \\
P(G_3|X-) &= \frac{3 - 7p + 5p^2 - p^3}{2(2-p)},
\end{aligned}
$$

$$P(G-|X_3) = \frac{3p-p^2}{2}, \text{ and}$$
$$P(G_3|X_3) = \frac{2-3p+p^2}{2}.$$

Due to the symmetry established above, these are also the probabilities for grandchildren. Surprisingly, even though the probabilities for parents and offspring were different from those of siblings, the grandparent/grandchild probabilities are the same as the aunt/uncle and niece/nephew probabilities. So, if the relationship between two individuals is known, the relavent probabilites can be calculated by simply using a series of parent/offspring steps (provided they are not siblings), and a sibling step which, if present, counts the same as a parent/offspring step.

If two (noninbred) individuals, denoted X and Y, are related, then they share many common ancestors. In one particular generation, they must share exactly two ancestors. Suppose these two ancestors occur n_1 generations before individual X and n_2 generations before individual Y. Then X and Y are *n-degree relatives*, where $n = n_1 + n_2 - 1$.

Now, consider an individual that is an n-degree relative of X (e.g. if $n = 2$ this is a grandparent, a grandchild, an aunt, a niece, etc.), say G^n. Then the following probabilities are derived, which have been established for $n = 1$. Again, siblings are special and these formulae do not hold in that case:

$$P(G_1^n|X_1) = \frac{p+(2^{n-1}-1)p^2}{2^{n-1}},$$
$$P(G_1^n|X_2) = \frac{p+(2^n-2)p^2}{2^n},$$

$$
\begin{aligned}
P(G_1^N|X_3) &= \frac{2^{n-1}-1}{2^{n-1}}p^2, \\
P(G_2^n|X_1) &= \frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}}, \\
P(G_2^n|X_2) &= \frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n}, \\
P(G_2^n|X_3) &= \frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}}, \\
P(G_3^n|X_1) &= \frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}}, \\
P(G_3^n|X_2) &= \frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n}, \text{ and} \\
P(G_3^n|X_3) &= \frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}.
\end{aligned}
$$

Each of these can be established by mathematical induction, assuming all to be true for n and showing that this yields the expected answer for $n+1$.

$$
\begin{aligned}
P(G_1^{n+1}|X_1) &= P(G_1^{n+1}|G_1^n)P(G_1^n|X_1) + P(G_1^{n+1}|G_2^n)P(G_2^n|X_1) \\
&\quad + P(G_1^{n+1}|G_3^n)P(G_3^n|X_1) \\
&= (p)\frac{p+(2^{n-1}-1)p^2}{2^{n-1}} + \left(\frac{p}{2}\right)\frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}} \\
&\quad + (0)\frac{(2^{n-1})-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}} \\
&= \frac{p+(2^n-1)p^2}{2^n}, \\
P(G_1^{n+1}|X_2) &= P(G_1^{n+1}|G_1^n)P(G_1^n|X_2) + P(G_1^{n+1}|G_2^n)P(G_2^n|X_2) \\
&\quad + P(G_1^{n+1}|G_3^n)P(G_3^n|X_2) \\
&= (p)\frac{p+(2^n-2)p^2}{2^n} + \left(\frac{p}{2}\right)\frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n}
\end{aligned}
$$

$$
\begin{aligned}
& +(0)\frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n} \\
= \; & \frac{p+(2^{n+1}-2)p^2}{2^{n+1}},
\end{aligned}
$$

$$
\begin{aligned}
P(G_1^{n+1}|X_3) = \; & P(G_1^{n+1}|G_1^n)P(G_1^n|X_3)+P(G_1^{n+1}|G_2^n)P(G_2^n|X_3) \\
& +P(G_1^{n+1}|G_3^n)P(G_3^n|X_3) \\
= \; & (p)\frac{2^{n-1}-1}{2^{n-1}}p^2+\left(\frac{p}{2}\right)\frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}} \\
& +(0)\frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}} \\
= \; & \frac{2^n-1}{2^n}p^2,
\end{aligned}
$$

$$
\begin{aligned}
P(G_2^{n+1}|X_1) = \; & P(G_2^{n+1}|G_1^n)P(G_1^n|X_1)+P(G_2^{n+1}|G_2^n)P(G_2^n|X_1) \\
& +P(G_2^{n+1}|G_3^n)P(G_3^n|X_1) \\
= \; & (1-p)\frac{p+(2^{n-1}-1)p^2}{2^{n-1}}+\left(\frac{1}{2}\right)\frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}} \\
& +(p)\frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}} \\
= \; & \frac{1+(2^{n+1}-3)p-(2^{n+1}-2)p^2}{2^n},
\end{aligned}
$$

$$
\begin{aligned}
P(G_2^{n+1}|X_2) = \; & P(G_2^{n+1}|G_1^n)P(G_1^n|X_2)+P(G_2^{n+1}|G_2^n)P(G_2^n|X_2) \\
& +P(G_2^{n+1}|G_3^n)P(G_3^n|X_2) \\
= \; & (1-p)\frac{p+(2^n-2)p^2}{2^n}+\left(\frac{1}{2}\right)\frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n} \\
& +(p)\frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n} \\
= \; & \frac{1+(2^{n+2}-4)p-(2^{n+2}-4)p^2}{2^{n+1}},
\end{aligned}
$$

$$
\begin{aligned}
P(G_2^{n+1}|X_3) = \; & P(G_2^{n+1}|G_1^n)P(G_1^n|X_3)+P(G_2^{n+1}|G_2^n)P(G_2^n|X_3) \\
& +P(G_2^{n+1}|G_3^n)P(G_3^n|X_3) \\
= \; & (1-p)\frac{2^{n-1}-1}{2^{n-1}}p^2+\left(\frac{1}{2}\right)\frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}}
\end{aligned}
$$

$$
\begin{aligned}
& +(p)\frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}} \\
= & \frac{(2^{n+1}-1)p-(2^{n+1}-2)p^2}{2^n},
\end{aligned}
$$

$$
\begin{aligned}
P(G_3^{n+1}|X_1) = & P(G_3^{n+1}|G_1^n)P(G_1^n|X_1)+P(G_3^{n+1}|G_2^n)P(G_2^n|X_1) \\
& +P(G_3^{n+1}|G_3^n)P(G_3^n|X_1) \\
= & (0)\frac{p+(2^{n-1}-1)p^2}{2^{n-1}}\left(\frac{1-p}{2}\right)\frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}} \\
& +(1-p)\frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}} \\
= & \frac{(2^n-1)-(2^{n+1}-2)p+(2^n-1)p^2}{2^n},
\end{aligned}
$$

$$
\begin{aligned}
P(G_3^{n+1}|X_2) = & P(G_3^{n+1}|G_1^n)P(G_1^n|X_2)+P(G_3^{n+1}|G_2^n)P(G_2^n|X_2) \\
& +P(G_3^{n+1}|G_3^n)P(G_3^n|X_2) \\
= & (0)\frac{p+(2^n-2)p^2}{2^n}+\left(\frac{1-p}{2}\right)\frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n} \\
& +(1-p)\frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n} \\
= & \frac{(2^{n+1}-1)-(2^{n+2}-3)p+(2^{n+1}-2)p^2}{2^{n+1}}, \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
P(G_3^{n+1}|X_3) = & P(G_3^{n+1}|G_1^n)P(G_1^n|X_3)+P(G_3^{n+1}|G_2^n)P(G_2^n|X_3) \\
& +P(G_3^{n+1}|G_3^n)P(G_3^n|X_3) \\
= & (0)\frac{2^{n-1}-1}{2^{n-1}}p^2+\left(\frac{1-p}{2}\right)\frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}} \\
& +(1-p)\frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}} \\
= & \frac{2^n-(2^{n+1}-1)p+(2^n-1)p^2}{2^n}.
\end{aligned}
$$

From which:

$$
\begin{aligned}
P(G^n - |X-) &= \{P(X_1)[P(G_1^n|X_1) + P(G_2^n|X_1)] + P(X_2)[P(G_1^n|X_2) \\
&\quad +P(G_2^n|X_2)]\}/[P(X_1) + P(X_2)] \\
&= \left\{p^2\left[\frac{p+(2^{n-1}-1)p^2}{2^{n-1}} + \frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}}\right]\right. \\
&\quad \left. +2p(1-p)\left[\frac{p+(2^n-2)p^2}{2^n}\,\frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n}\right]\right\} \Big/ (2p-p^2) \\
&= \frac{1+(2^{n+1}-3)p-(2^{n+1}-3)p^2+(2^{n-1}-1)p^3}{2^{n-1}(2-p)}, \qquad (2.3) \\
P(G^n - |X_3) &= P(G_1^n|X_3) + P(G_2^n|X_3) \\
&= \frac{2^{n-1}-1}{2^{n-1}}(p^2) + \frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}} \\
&= \frac{(2^n-1)p-(2^{n-1}-1)p^2}{2^{n-1}}, \qquad (2.4) \\
P(G_3^n|X-) &= \frac{P(X_1)P(G_3^n|X_1) + P(X_2)p(G_3^n|X_2)}{P(X_1)+P(X_2)} \\
&= \frac{(p^2)\frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}} + 2p(1-p)\frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n}}{2p-p^2} \\
&= \frac{(2^n-1)-(5\times 2^{n-1}-3)p+(2^{n+1}-3)p^2-(2^{n-1}-1)p^3}{2^{n-1}(2-p)}, \text{ and} \qquad (2.5) \\
P(G_3^n|X_3) &= \frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}. \qquad (2.6)
\end{aligned}
$$

Now some observations about these formulae can be made. Table 2.4 lists these conditional probabilities for various values of n and p. Notice that the probabilities for siblings are slightly higher than the corresponding probabilities for degree one relatives, although the numbers are very close in size. Notice that if a limit as n approaches infinity is taken, in each of the above formulae, the probability of the presence or absence of the

Table 2.4: Conditional probabilities of an individual displaying a dominant trait of allele frequency p, given an n degree relative displays the trait.

	Allele Frequency (p)										
n	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
SIBS	0.50	0.58	0.66	0.73	0.80	0.85	0.90	0.94	0.97	0.99	1.00
1	0.50	0.57	0.64	0.71	0.78	0.83	0.89	0.93	0.97	0.99	1.00
2	0.25	0.38	0.50	0.61	0.71	0.79	0.86	0.92	0.96	0.99	1.00
3	0.13	0.29	0.43	0.56	0.67	0.77	0.85	0.92	0.96	0.99	1.00
4	0.06	0.24	0.40	0.54	0.66	0.76	0.85	0.91	0.96	0.99	1.00
∞	0.00	0.19	0.36	0.51	0.64	0.75	0.84	0.91	0.96	0.99	1.00

trait is (as expected, since random mating with unrelated individuals is assumed):

$$\begin{aligned} P(G^{\infty} - |X-) &= 2p - p^2, \\ P(G^{\infty} - |X_3) &= 2p - p^2, \\ P(G_3^{\infty}|X-) &= (1-p)^2, \text{ and} \\ P(G_3^{\infty}|X_3) &= (1-p)^2. \end{aligned}$$

Notice that if $p = 0$, then $P(G^n - |X-) = \left(\frac{1}{2}\right)^n$ which is twice the coefficient of relationship [Lynch 1988, Smith 1989 and Wright 1922].

Although the following probabilities are not employed in my subsequent modelling, I do include them here because of their similarity to the previous results. I wish to explore certain special cases where inbreeding is present. If two siblings S^1 and S^2 have an offspring D then Tables 2.1 and 2.3 yield:

$$
\begin{aligned}
P(D_1|S_1^1) &= \sum_j P(S_j^2|S_1^1)P(D_1|S_1^1 \text{ and } S_j^2) = \frac{1+p}{2}, \\
P(D_1|S_2^1) &= \sum_j P(S_j^2|S_2^1)P(D_1|S_2^1 \text{ and } S_j^2) = \frac{1+2p}{8}, \\
P(D_1|S_3^1) &= \sum_j P(S_j^2|S_3^1)P(D_1|S_3^1 \text{ and } S_j^2) = 0, \\
P(D_2|S_1^1) &= \sum_j P(S_j^2|S_1^1)P(D_2|S_1^1 \text{ and } S_j^2) = \frac{1-p}{2}, \\
P(D_2|S_2^1) &= \sum_j P(S_j^2|S_2^1)P(D_2|S_2^1 \text{ and } S_j^2) = \frac{1}{2}, \\
P(D_2|S_3^1) &= \sum_j P(S_j^2|S_3^1)P(D_2|S_3^1 \text{ and } S_j^2) = \frac{p}{2}, \\
P(D_3|S_1^1) &= \sum_j P(S_j^2|S_1^1)P(D_3|S_1^1 \text{ and } S_j^2) = 0, \\
P(D_3|S_2^1) &= \sum_j P(S_j^2|S_2^1)P(D_3|S_2^1 \text{ and } S_j^2) = \frac{3-2p}{8}, \text{ and} \\
P(D_3|S_3^1) &= \sum_j P(S_j^2|S_3^1)P(D_3|S_3^1 \text{ and } S_j^2) = \frac{2-p}{2}.
\end{aligned}
$$

From which:

$$
\begin{aligned}
P(D-|S^1-) &= \frac{P(S_1^1)[P(D_1|S_1^1)+P(D_2|S_1^1)]+P(S_2^1)[P(D_1|S_2^1)+P(D_2|S_2^1)]}{P(S_1^1)+P(S_2^1)} \\
&= \frac{(p^2)[\frac{1+p}{2}+\frac{1-p}{2}]+2p(1-p)[\frac{1+2p}{8}+\frac{1}{2}]}{2p-p^2} \\
&= \frac{5+p-2p^2}{4(2-p)}, \\
P(D-|S_3^1) &= P(D_1|S_3^1)+P(D_2|S_3^1)
\end{aligned}
$$

$$
\begin{aligned}
&= 0 + \frac{p}{2} \\
&= \frac{p}{2}, \\
P(D_3|S^1-) &= \frac{P(S_1^1)P(D_3|S_1^1) + P(S_2^1)P(D_3|S_2^1)}{P(S_1^1) + P(S_2^1)} \\
&= \frac{(p^2)(0) + 2p(1-p)\frac{3-2p}{8}}{2p - p^2} \\
&= \frac{3 - 5p + 2p^2}{4(2-p)}, \text{ and} \\
P(D_3|S_3^1) &= \frac{2-p}{2}.
\end{aligned}
$$

Now consider a similar case, where an individual X mates with an n-degree relative G^n producing an offspring D. See Tables 2.1 and 2.5. These follow:

$$
\begin{aligned}
P(D_1|X_1) &= \sum_j P(G_j^n|X_1)P(D_1|X_1 \text{ and } G_j^n) = \frac{1 + (2^n - 1)p}{2^n}, \\
P(D_1|X_2) &= \sum_j P(G_j^n|X_2)P(D_1|X_2 \text{ and } G_j^n) = \frac{1 + (2^{n+1} - 2)p}{2^{n+2}}, \\
P(D_1|X_3) &= \sum_j P(G_j^n|X_3)P(D_1|X_3 \text{ and } G_j^n) = 0, \\
P(D_2|X_1) &= \sum_j P(G_j^n|X_1)P(D_2|X_1 \text{ and } G_j^n) = \frac{(2^n - 1) - (2^n - 1)p}{2^n}, \\
P(D_2|X_2) &= \sum_j P(G_j^n|X_2)P(D_2|X_2 \text{ and } G_j^n) = \frac{1}{2}, \\
P(D_2|X_3) &= \sum_j P(G_j^n|X_3)P(D_2|X_3 \text{ and } G_j^n) = \frac{(2^n - 1)p}{2^n}, \\
P(D_3|X_1) &= \sum_j P(G_j^n|X_1)P(D_3|X_1 \text{ and } G_j^n) = 0, \\
P(D_3|X_2) &= \sum_j P(G_j^n|X_2)P(D_3|X_2 \text{ and } G_j^n) = \frac{(2^{n+1} - 1) - (2^{n+1} - 2)p}{2^{n+2}}, \text{ and} \\
P(D_3|X_3) &= \sum_j P(G_j^n|X_3)P(D_3|X_3 \text{ and } G_j^n) = \frac{2^n - (2^n - 1)p}{2^n}.
\end{aligned}
$$

Table 2.5: Conditional probability of genotype of an n-degree relative, G_j^n, of an individual X.

G_j^n	$P(G_j^n \mid X_1)$	$P(G_j^n \mid X_2)$	$P(G_j^n \mid X_3)$
G_1^n	$\frac{p+(2^{n-1}-1)p^2}{2^{n-1}}$	$\frac{p+(2^n-2)p^2}{2^n}$	$\frac{(2^{n-1}-1)p^2}{2^{n-1}}$
G_2^n	$\frac{1+(2^n-3)p-(2^n-2)p^2}{2^{n-1}}$	$\frac{1+(2^{n+1}-4)p-(2^{n+1}-4)p^2}{2^n}$	$\frac{(2^n-1)p-(2^n-2)p^2}{2^{n-1}}$
G_3^n	$\frac{(2^{n-1}-1)-(2^n-2)p+(2^{n-1}-1)p^2}{2^{n-1}}$	$\frac{(2^n-1)-(2^{n+1}-3)p+(2^n-2)p^2}{2^n}$	$\frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}$

And hence:

$$
\begin{aligned}
P(D-|X-) &= \frac{P(X_1)[P(D_1|X_1)+P(D_2|X_1)]+P(X_2)[P(D_1|X_2)+P(D_2|X_2)]}{P(X_1)+P(X_2)} \\
&= \frac{(p^2)[\frac{1+(2^n-1)p}{2^n}+\frac{(2^n-1)-(2^n-1)p}{2^n}]+2p(1-p)[\frac{1+(2^{n+1}-2)p}{2^{n+2}}+\frac{1}{2}]}{2p-p^2} \\
&= \frac{(2^{n+1}+1)+(2^{n+1}-3)p-(2^{n+1}-2)p^2}{2^{n+1}(2-p)}, \\
P(D-|X_3) &= P(D_1|X_3)+P(D_2|X_3) \\
&= 0+\frac{(2^n-1)p}{2^n} \\
&= \frac{(2^n-1)p}{2^n}, \\
P(D_3|X-) &= \frac{P(X_1)P(D_3|X_1)+P(X_2)P(D_3|X_2)}{P(X_1)+P(X_2)} \\
&= \frac{(p^2)(0)+2p(1-p)\frac{(2^{n+1}-1)-(2^{n+1}-2)p}{2^{n+2}}}{2p-p^2} \\
&= \frac{(2^{n+1}-(2^{n+2}-3)p+(2^{n+1}-2)p^2}{2^{n+1}(2-p)}, \text{ and} \\
P(D_3|X_3) &= \frac{2^n-(2^n-1)p}{2^n}.
\end{aligned}
$$

Notice that for $n = 1$, this reduces to the case of mated siblings. Also, for any particular type of inbreeding, say continued brother/sister mating, it is possible to calculate these type probabilities. The calculations (although tedious) would employ the methods used here, that is, the construction of tables of the nine possible genotypes of the parents with the corresponding conditional probabilities.

2.2 The Model

I will use the conditional probabilities of the previous section to put likelihoods on different degrees of relationship of pairs of individuals. In order to test this approach, I have written programs that generate DNA fingerprint type data. See the Appendix for the program listings. The programs are written in Fortran 77 and were ran on a VAX minicomputer and an IBM compatible personal computer. The programs require the input of allele frequencies and a genealogy for the model population. Each individual of the population must have either no parents present in the model population or exactly two parents present. The "parentless" individuals are randomly assigned genotypes according to the allele frequencies and then genotypes are generated for the remaining individuals assuming Mendelian inheritance and a given rate of mutation. The *phenotypes* of the individuals are then used to reconstruct the genealogy using maximum likelihood estimators.

The program FREQDAT.FOR allows for the input of the number of loci to be considered, the number of individuals in the hypothetical population, and the number of alleles at each locus. The alleles at a given locus are assumed to be equally frequent.

For example, if the number of alleles at a given locus is 20, then each allele is assumed to have a frequency of 5%. Although this may seem a bit restrictive, if several alleles of different frequencies are desired, this can be accomplished by using several loci with different numbers of alleles at each. The information that the user gives this program is stored in a file.

The program GENDAT.FOR takes as input the size of the model population and the genealogy for that population. The genealogy is stored in a file as a square matrix, A, which is the *adjacency matrix* for the parent $\rightarrow$ offspring *directed graph*, G (see Bondy and Murty [1976] for definitions of graph theoretic terms). That is, the entry $a_{i,j}$ of A is 1 if individual i is the parent of individual j and 0 otherwise. In order for the following programs to work, each individual must have either 0 or 2 parents present in the model population. In graph theoretic terms, the *in-degree* of each vertex of the graph G is either 0 or 2. Notice that the *out-degree* of vertex i is the number of offspring of individual i that are present in the model population.

The program GENOME.FOR is the program that actually generates the genotypes for the individuals of the model population. The program reads population size and allele frequencies from the file created with the FREQDAT.FOR program and reads the genealogy matrix from the file which was created by GENDAT.FOR. As described above, the members of the model population fall into two categories: those with no parents present in the model population and those with two parents in the model population. Individuals of this first type are randomly assigned a genotype based on allele frequencies. For example, if a particular locus contains 20 alleles, then a first allele is chosen with there being a 5% probability of any one allele getting picked, and then a second allele

is (independently) chosen under the same conditions. So at this locus, the probability of homozygosity of any given allele is 0.25% and the probability of heterozygosity is 95%. Once the genotypes of the "parentless" individuals are assigned, the remaining genotypes are generated according to the genealogy which was input. Simple Mendelian inheritance is followed. At a given locus, one allele is inherited from each parent. The allele inherited from each parent is chosen with an equal probability from the two present at that locus. The program also allows for a certain level of mutation. The mutation rate is measured per allele. Once a genotype has been given to an individual with parents in the model population, each of the alleles is "tested" for mutation. Depending on the mutation rate, substitutions are randomly made for any allele at the locus. This process carries the advantage of allowing not only for mutation, but also for laboratory error. Since this model is intended to deal with DNA fingerprint data, it is quiet likely that particular bands could be misread and confused with other bands. Also, if all the alleles are flawlessly passed on, it is rather easy to check pairwise for parents (each allele in the offspring must come from one or the other of the parents in the absence of mutation). Finally, the genotypic data for all the individuals is stored in a file. It is this data that will be used to reconstruct the genealogy.

The program RECONSTR.FOR is the heart of this project. It is this program that would evaluate data gathered in the lab. The input which is read from the file created by GENOME.FOR is simply the number of individuals and the genotypes of each. The program views the data "with a blind eye" in the sense that it does not make a distinction between an allele being present in the homozygous or heterozygous state. That is, the program only recognizes the presence or absence of an allele (i.e. the phenotype of the

individual). First, the frequency of each trait is estimated. Assuming Hardy-Weinberg equilibrium, if a dominant trait is determined by an allele with frequency p, then the trait has frequency $f = 2p - p^2$. So the allele frequency can be estimated as $p = 1 - \sqrt{1-f}$. Next, the likelihood of each relationship is calculated for all possible pairs of individuals. For any trait (or allele) present in the model population, there are four possibilities when one compares two individuals. There is an equation from the previous section associated with each possibility. The four cases are:

1. the trait is present in both individuals, in which case equation (2.3) is the associated equation,

2. the trait is present in the first individual, but not in the second, in which case equation (2.4) is the associated equation,

3. the trait is absent in the first individual, but present in the second, in which case equation (2.5) is the associated equation, and

4. the trait is absent in both individuals in which case equation (2.6) is the associated equation.

The probability of each case must be computed for various degrees of relationship. Given that i and j are n degree relatives, the probabilities associated with the four cases for a trait with allele frequency p are, respectively:

$$\begin{aligned} P(i- \text{ and } j-) &= P(i-)P(j-|i-) \\ &= (2p-p^2)\frac{1+(2^{n+1}-3)p-(2^{n+1}-3)p^2+(2^{n-1}-1)p^3}{2^{n-1}(2-p)}, \qquad (2.7) \\ P(i- \text{ and } j_3) &= P(i-)P(j_3|i-) \end{aligned}$$

$$= (2p-p^2)\frac{(2^n-1)-(5\times 2^{n-1}-3)p+(2^{n+1}-3)p^2-(2^{n-1}-1)p^3}{2^{n-1}(2-p)},$$

$$= \frac{(2^n-1)p-(5\times 2^{n-1}-3)p^2+(2^{n+1}-3)p^3-(2^{n-1}-1)p^4}{2^{n-1}}, \quad (2.8)$$

$$P(i_3 \text{ and } j-) = P(i_3)P(j-|i_3)$$

$$= (1-p)^2\frac{(2^n-1)p-(2^{n-1}-1)p^2}{2^{n-1}},$$

$$= \frac{(2^n-1)-(5\times 2^{n-1}-3)p+(2^{n+1}-3)p^2-(2^{n-1}-1)p^3}{2^{n-1}(2-p)},$$

$$= \frac{(2^n-1)p-(5\times 2^{n-1}-3)p^2+(2^{n+1}-3)p^3-(2^{n-1}-1)p^4}{2^{n-1}}, \quad (2.9)$$

$$P(i_3 \text{ and } j_3) = P(i_3)P(j_3|i_3)$$

$$= (1-p)^2\frac{2^{n-1}-(2^n-1)p+(2^{n-1}-1)p^2}{2^{n-1}}. \quad (2.10)$$

Let the events described by cases 1, 2, 3 and 4 be denoted by E_1, E_2, E_3, and E_4, respectively. From equations (2.8) and (2.9), it follows that the probabilities for events E_2 and E_3 are the same for each allele. This means that there is, as expected, a symmetry and so the likelihood of "i is an n degree relative of j" and the likelihood of "j is an n degree relative of i" are the same. Hence one need only deal with the event "i and j are n degree relatives".

Now, introduce the variable x_a. Let x_a take on the values -1, 0, or 1 as follows:

$$x_a = \begin{cases} -1 & \text{if } E_4 \text{ describes the situation for allele } a \\ 0 & \text{if } E_2 \text{ or } E_3 \text{ describes the situation for allele } a \\ 1 & \text{if } E_1 \text{ describes the situation for allele } a. \end{cases}$$

If p is the frequency of allele a then:

$$\begin{aligned} x_a &= -1 \text{ with the probability given in 2.10} \\ x_a &= 0 \text{ with the probability given in 2.8} \\ x_a &= 1 \text{ with the probabiltiy given in 2.7} \end{aligned}$$

If these three probabilities are denoted as q_a, r_a, and t_a, respectively, then the likelihood that i and j are n degree relatives is

$$\prod_{a \text{ is an allele}} q_a^{x_a(x_a-1)/2} \, r_a^{1-x_a^2} \, t_a^{x_a(x_a+1)/2} \tag{2.11}$$

RECONSTR.FOR uses expression (2.11) to put likelihoods on different degrees of relationship between i and j. For computational reasons, the program calculates the logarithm of the likelihood, as opposed to the likelihood itself. This will help avoid roundoff error that might arise when dealing directly with the likelihoods since they will, in general, be very small numbers. Since the logarithm function is an increasing function, the maximum logarithm of the likelihoods will correspond to the maximum likelihood. Once the maximum likelihood is chosen, the corresponding degree of relationship is associated with the pair i and j. These values are stored in a distance matrix which RECONSTR.FOR writes to a file.

The final program in the model, MATRIX.FOR, simply puts the output in a readable format. This program prints in a file the names of the genealogy file, the frequency file, and the genome file, as well as the size of the model population and the data in the frequency file. Finally, of course, the distance matrix is printed in the file. In order to

view the results of this model run, it is the file created by MATRIX.FOR that must be used.

Chapter 3

Results

The results of running the model described in Chapter 2 on several genealogies are discussed in this chapter. The genealogies will be refered to as "Genealogy 1" through "Genealogy 6". Each genealogy was ran in the model several times. Each run involves a different value for one of the following parameters: number of loci, allele frequencies, and mutation rate per allele. Genealogies 1, 2 and 4 are given in Figures 3.1, 3.2 and 3.3, respectively. Genealogy 3 is a sample of 10 unrelated individuals, Genealogy 5 includes the population from Genealogy 4 along with 10 unrelateds, and Genealogy 6 includes the population from Genealogy 4 along with 20 unrelateds.

When estimating relationships between pairs of individuals, there are two ways to view the results. The percentage of actual degree n relationships recognized as degree n relationships is one statistic that can be associated with the results. This number describes the ability of the algorithm to recognize certain degrees of relationships. A second statistic associated with the output is the percentage of relationships said to be degree n, that actually are. This describes the level of confidence one can put in the results of the model. For example, the model might only accurately recognize 70% of the first degree relationships, but each time the model indicates a first degree relationship, it could be correct. That is, one can put 100% confidence in the model when it indicates a degree one relationship. However, there must be less confidence in the degree two or higher relationships since it has said that 30% of the first degrees must fall in these categories.

3.1 Genealogy 1

Genealogy 1 consists of 33 individuals with 4 groups of related individuals, 5 individuals unrelated to any others in the sample, two pairs of siblings, and several 3^{rd} and 4^{th} degree relationships (see Figure 3.1). Namely, there are 28 first degree relationships, 14 second degree relationships, and 486 third or higher degree relationships (including unrelated).

Genealogy 1 was chosen for analysis because it contains a variety of different relationships. It includes one large clan of 15 related individuals spanning four generations. It includes a smaller group (of size 5) which spans three generations. It also has two pairs of parents which each produce a pair of siblings.

Genealogy 1 might represent samples taken from different areas, or a single sample from a large population with small groups of related individuals. These groups would be unrelated or distantly related to one another.

Tables 3.1 and 3.2 summarize the results of running Genealogy 1 with no mutations, 10, 20 and 30 loci, and 2%, 5%, 10% and 20% allele frequencies. As can be seen, the model is quite successful with first degree relationships and unrelated pairs. It is significantly less successful with second degree relationships, however.

As expected, the higher the mutation rate, the less successful the model. However, with the mutation rate as high as 30%, the model is still surprisingly accurate with first degree relationships and with unrelateds.

To test the effects of mutations, the model was ran several times with a constant number of loci (20) and various allele frequencies. Tables 3.3 and 3.4 report the results

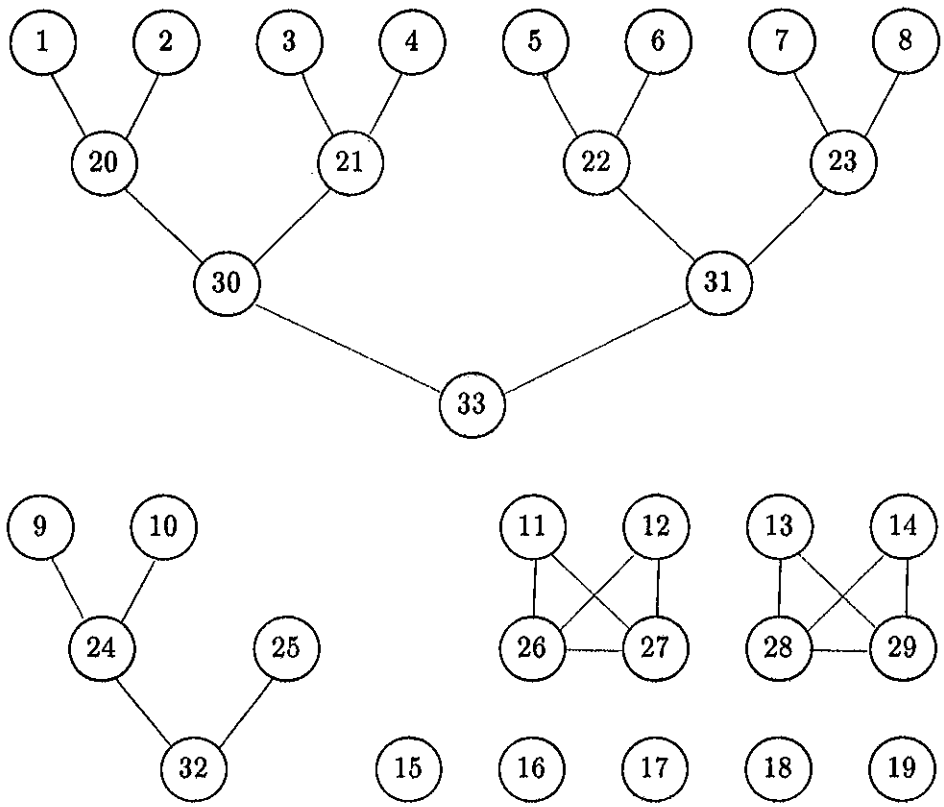

Figure 3.1: Genealogy 1. The edges represent degree one relationships, either parent-offspring or siblings.

Table 3.1: These entries are the results for Genealogy 1 with no mutations and various values of the number of loci and allele frequencies. In this table, "% ... correct" is the percentage of the model output that is correct. These values indicate the confidence one can put in the model output.

number of loci	allele frequencies	% relationships correct	% 1^{st} degree correct	% 2^{nd} degree correct	% unrelated correct
10	2%	98.67	100.00	88.89	98.79
10	5%	97.54	100.00	52.94	98.97
10	10%	93.56	91.30	22.86	98.93
10	20%	84.47	80.77	5.41	98.36
20	2%	99.05	100.00	90.91	99.18
20	5%	98.86	100.00	83.33	99.18
20	10%	98.67	96.43	73.33	99.59
20	20%	93.75	82.35	21.05	98.73
30	2%	99.24	100.00	91.67	99.39
30	5%	90.05	100.00	90.91	99.18
30	10%	98.86	100.00	78.57	99.39
30	20%	95.08	96.55	25.93	98.74

Table 3.2: These entries are the results for Genealogy 1 with no mutations and various values of the number of loci and allele frequencies. In this table, "% ... recognized" indicates the percentage of a particular type of relationship that was correctly identified.

number of loci	allele frequencies	% relationships recognized	% 1^{st} degree recognized	% 2^{nd} degree recognized	% unrelated recognized
10	2%	98.67	100.00	57.14	99.79
10	5%	97.54	100.00	64.29	98.35
10	10%	93.56	82.14	57.14	95.27
10	20%	84.47	75.00	28.57	86.63
20	2%	99.05	96.43	71.43	100.00
20	5%	98.86	96.43	71.43	99.79
20	10%	98.67	96.43	78.57	99.38
20	20%	93.75	82.14	35.71	96.09
30	2%	99.24	96.43	78.57	100.00
30	5%	90.05	96.43	71.43	100.00
30	10%	98.86	92.86	78.57	99.79
30	20%	95.08	85.71	50.00	96.91

Table 3.3: Results for Genealogy 1 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% relationships correct	% 1^{st} degree correct	% 2^{nd} degree correct	% unrelated correct
10	2%	97.16	100.00	46.15	98.38
10	5%	96.78	100.00	40.00	98.37
10	10%	96.59	96.55	40.00	98.77
10	20%	94.13	100.00	25.71	98.94
20	2%	94.51	100.00	14.29	97.79
20	5%	95.27	100.00	21.05	97.98
20	10%	93.94	96.55	12.50	97.79
20	20%	90.15	100.00	10.42	98.08
30	2%	93.56	100.00	8.33	97.79
30	5%	93.75	100.00	15.38	97.98
30	10%	92.61	100.00	12.90	97.57
30	20%	88.45	100.00	9.26	97.45

Table 3.4: Results for Genealogy 1 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% relationships recognized	% 1^{st} degree recognized	% 2^{nd} degree recognized	% unrelated recognized
10	2%	97.16	75.00	42.86	100.00
10	5%	96.78	75.00	42.86	99.59
10	10%	96.59	71.43	57.14	99.18
10	20%	94.13	71.43	64.86	96.30
20	2%	94.51	35.71	21.43	100.00
20	5%	95.27	50.00	28.57	99.79
20	10%	93.94	42.86	21.43	98.97
20	20%	90.15	42.86	35.71	94.44
30	2%	93.56	21.43	14.29	100.00
30	5%	93.75	21.43	28.57	99.79
30	10%	92.61	14.29	28.57	98.97
30	20%	88.45	14.29	35.71	94.24

for a mutation rate of 10%, 20%, and 30% per allele and for allele frequencies of 2%, 5%, 10%, and 20%.

3.2 Genealogy 2

Genealogy 2 consists of 26 individuals which make up four groups of related individuals. The sample has six pairs of siblings and several second degree relationships (see Figure 3.2). Namely, there are 34 first degree relationships, 12 second degree relationships and 279 unrelated pairs. Genealogy 2 does not contain quite the diversity of relationships of Genealogy 1. This example might represent a sample from a population of organisms with a long generation time. With the absence of inbreeding, this genealogy could represent a population with many isolated groups. One important difference between this genealogy and Genealogy 1 is the absence of unrelateds in this population sample.

Genealogy 2 was chosen to test if the model can actually break the population into the proper related groups and accurately distinguish between individuals from different demes.

Tables 3.5 and 3.6 summarize the results of running Genealogy 2 with no mutations, 10, 20 and 30 loci, and 2%, 5%, 10% and 20% allele frequencies. Tables 3.7 and 3.8 report the results for mutation rates of 10%, 20% and 30% and for allele frequencies of 2%, 5%, 10%, and 20%. As with Genealogy 1, the model is very successful in recognizing first degree relationships and unrelated pairs. However, it recognizes second degree relationships with a varying degree of success, and often rather poorly.

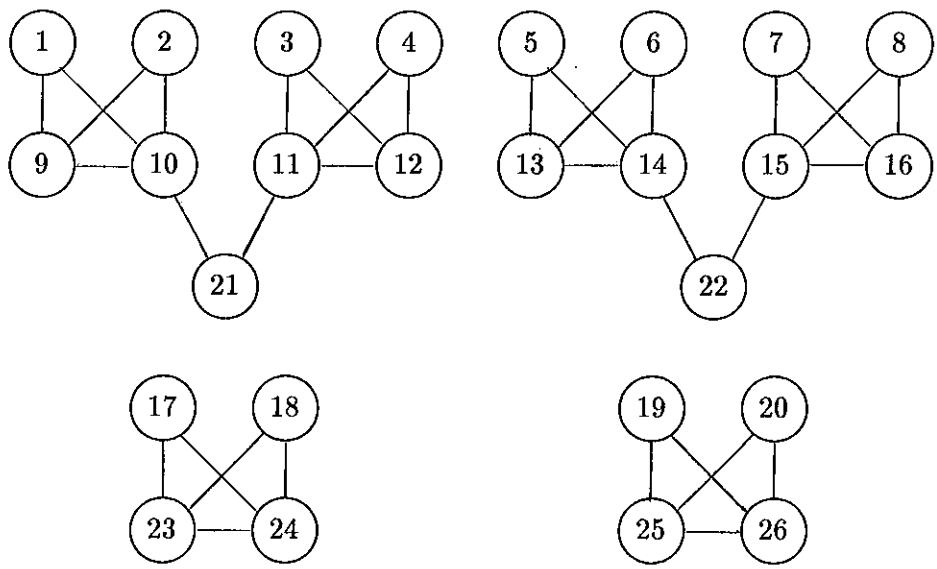

Figure 3.2: Genealogy 2. The edges represent degree one relationships.

Table 3.5: These entries are the results for Genealogy 2 with no mutations and various values of the number of loci and allele frequencies. As in Table 3.1, "% ... correct" is the percentage of the model output that is correct. These values indicate the confidence one can put in the model output.

number of loci	allele frequencies	% relationships correct	% 1^{st} degree correct	% 2^{nd} degree correct	% unrelated correct
10	2%	96.61	100.00	55.56	97.55
10	5%	95.38	100.00	33.33	96.86
10	10%	91.08	100.00	20.69	97.79
10	20%	84.92	97.14	7.14	96.55
20	2%	99.38	100.00	100.00	99.28
20	5%	98.77	100.00	100.00	98.59
20	10%	98.46	100.00	100.00	98.23
20	20%	94.77	97.14	39.13	99.28
30	2%	98.46	100.00	88.89	98.59
30	5%	98.77	100.00	90.00	98.94
30	10%	97.85	97.14	77.78	98.59
30	20%	95.69	100.00	43.75	98.22

Table 3.6: These entries are the results for Genealogy 2 with no mutations and various values of the number of loci and allele frequencies. As in Table 3.2, "% ... recognized" indicates the percentage of a particular type of relationship that was correctly identified.

number of loci	allele frequencies	% relationships recognized	% 1^{st} degree recognized	% 2^{nd} degree recognized	% unrelated recognized
10	2%	96.61	88.24	41.67	100.00
10	5%	95.38	85.29	25.00	99.64
10	10%	91.08	70.59	50.00	95.34
10	20%	84.92	58.82	25.00	90.32
20	2%	99.38	100.00	83.33	100.00
20	5%	98.77	100.00	66.67	100.00
20	10%	98.46	100.00	58.33	100.00
20	20%	94.77	73.52	75.00	98.21
30	2%	98.46	97.06	66.67	100.00
30	5%	98.77	97.06	75.00	100.00
30	10%	97.85	94.12	58.33	100.00
30	20%	95.69	82.35	58.33	98.92

Table 3.7: Results for Genealogy 2 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% relationships correct	% 1^{st} degree correct	% 2^{nd} degree correct	% unrelated correct
10	2%	94.77	100.00	33.33	97.55
10	5%	93.85	100.00	27.78	97.55
10	10%	93.85	100.00	26.67	96.88
10	20%	89.23	100.00	18.18	97.45
20	2%	90.46	100.00	12.00	96.88
20	5%	88.92	100.00	7.69	95.88
20	10%	87.38	100.00	3.33	95.85
20	20%	87.38	100.00	11.11	96.81
30	2%	86.46	100.00	0.00	94.90
30	5%	87.38	100.00	3.45	95.55
30	10%	88.31	100.00	7.41	95.55
30	20%	84.31	100.00	2.70	94.67

Table 3.8: Results for Genealogy 2 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% relationships recongnized	% 1^{st} degree recognized	% 2^{nd} degree recognized	% unrelated recognized
10	2%	94.77	70.59	41.67	100.00
10	5%	93.85	61.76	41.67	100.00
10	10%	93.85	64.71	33.33	100.00
10	20%	89.23	50.00	50.00	95.70
20	2%	90.46	35.29	25.00	100.00
20	5%	88.92	23.53	16.67	100.00
20	10%	87.38	17.65	8.33	99.28
20	20%	87.38	20.59	33.33	97.85
30	2%	86.46	5.88	0.00	100.00
30	5%	87.38	11.76	8.33	100.00
30	10%	88.31	17.65	16.67	100.00
30	20%	84.31	20.59	8.33	95.34

3.3 Genealogy 3

Genealogy 3 consists of 10 unrelated individuals. The purpose of running the model on this genealogy is to see if it confuses unrelateds with relateds. One advantage of having such a sample is that allele frequencies can be accurately estimated, without the bias that would arise from having related individuals in the sample. The model was run on Genealogy 3 for 20 loci, allele frequencies of 2%, 5%, 10% and 20%, and no mutations. All relationships were correctly identified. The model was also run for mutation rates of 10%, 20% and 30% per allele. In each case, all relationships were again correctly identified. This in itself is not surprising: The alleles are randomly assigned and the mutations occur at random. It is expected that mutations would have no effect on the results of the model for Genealogy 3. In fact, the model was ran with 20 loci, 20% allele frequency and a mutation rate of 50%. Again, the model was correct in 100% of the cases.

3.4 Genealogies 4, 5 and 6

Genealogy 4 differs significantly from Genealogies 1, 2 and 3. It includes 10 individuals, 8 of which are descended from two founders (see Figure 3.3). The population is highly inbred. It is difficult to talk about "degrees of relationship" in the presence of inbreeding.

This genealogy was chosen to see how the model would react to a small inbred population. Under such circumstances, other factors might be available to help establish relationships. Such factors include behavior and location or range. Genealogy 4 would likely represent a sample taken from a single site. In Chapter 4 it will be argued that the

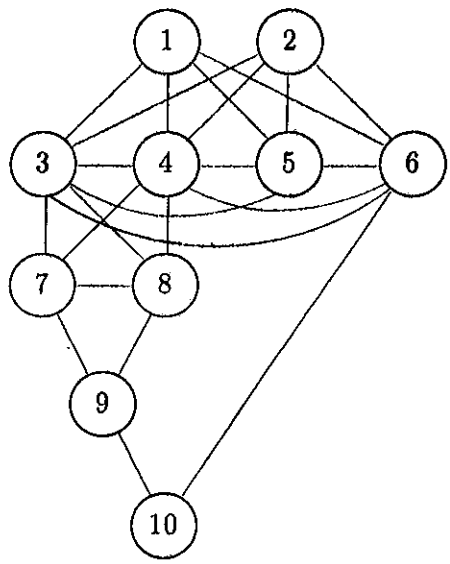

Figure 3.3: Genealogy 4. The edges represent degree one relationships.

model should report all pairs of individuals as degree one relatives, with the exception of 1 and 2. Table 3.9 presents the percentage of these relationships correctly recognized for various mutation rates and allele frequencies.

Genealogy 5 includes the individuals of Genealogy 4 along with 10 additional individuals each unrelated to any others in the sample population. This genealogy was chosed to see if the added individuals have any effect on the model output. Genealogy 6 includes the individuals of Genealogy 4 along with 20 unrelateds. This was chosen to see if the results the model gives with the above two genealogies are refined. Tables 3.10 and 3.11 present the summaries of the model output for Genealogies 5 and 6, respectively.

Table 3.9: The percentage of degree one relationships recognized by the model for Genealogy 4 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% 1^{st} degrees recoginized
0	2%	4.55
0	5%	4.55
0	10%	4.55
0	20%	4.55
10	2%	2.27
10	5%	2.27
10	10%	4.55
10	20%	4.55
20	2%	2.27
20	5%	0.00
20	10%	2.27
20	20%	4.55
30	2%	0.00
30	5%	0.00
30	10%	0.00
30	20%	2.27

Table 3.10: The percentage of degree one relationships recognized by the model and the percentage of unrelateds said to be related for Genealogy 5 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% 1^{st} degrees % recognized	% unrelated said to be degree 1
0	2%	40.91	0.00
0	5%	45.45	0.00
0	10%	43.18	0.33
0	20%	43.45	3.67
10	2%	18.18	0.00
10	5%	13.64	0.00
10	10%	13.64	0.66
10	20%	20.00	2.00

Table 3.11: The percentage of degree one relationships recognized by the model and the percentage of unrelateds said to be related for Genealogy 6 with 20 loci and various values of allele frequencies and mutation rates.

mutation rate	allele frequencies	% 1^{st} degrees % recognized	% unrelated said to be degree 1
0	2%	70.45	0.00
0	5%	63.64	0.25
0	10%	68.18	1.25
0	20%	75.00	3.75
10	2%	29.55	0.00
10	5%	25.00	0.00
10	10%	27.27	0.50
10	20%	34.09	3.13

Chapter 4

Discussion

In this chapter, the strengths and weaknesses of the model are discussed. In addition, certain information (namely, age structure) is added to the knowledge about the sample population and some of the results of Chapter 3 are presented graphically.

4.1 Genealogy 1

As revealed in Tables 3.1 and 3.3, a great deal of confidence can be put in the model when it says two individuals are first degree relatives or when it says they are unrelated. One can put 95% confidence in a relationship said to be first degree, provided allele frequencies are less than 20% when 20 or 30 loci are used, or less than 10% when 10 loci are used. With mutations present, the degree of confidence is even greater. As shown in Table 3.3, there is more than a 96% confidence in "first degree relationship" output in every case tested. This is because the mutations have shifted pairs that may have been bordering on first and second degree to second degree. Unfortunately, comparing Table 3.2 and 3.4, one sees that the mutations also lessen the number of degree one relationships that are recognized. With or without mutations, the model is very successful with unrelated pairs. From Tables 3.1 and 3.3, one can put over 97% confidence in the output when it indicates that a pair is unrelated, from Tables 3.2 and 3.4, the model is over 94% accurate in recognizing unrelated pairs in all cases tested, except one (10 loci, 20% allele frequencies and no mutations).

The model is much less successful with second degree relationships. From Table 3.1, one can see that as the number of loci increase or as the allele frequencies decrease (and therefore the number of alleles increases), the confidence in degree two output increases. However, there is never 95% confidence in this output, although a few cases reveal 90% confidence. Things are even worse in the presence of mutations (see Table 3.3). In fact, in the cases tested, one can never even put 50% confidence in the degree two output. This is primarily because the model confuses degree one relationships for degree two relationships. This is also why a smaller percentage of degree one relationships are recognized when mutations are present (compare Tables 3.2 and 3.4). The recognition rate of degree two relationships is also rather poor. In the absence of mutation, the average rate of recognition in the cases tested is around 60%. This rate is much less when mutations are present, averaging around 40%.

One source of error, is the determination of allele frequencies. The frequencies are estimated assuming the population sample consists of unrelateds. These frequencies are then used to estimate relatedness. The result is that allele frequencies are overestimated and in turn, degrees of relationship are underestimated. Attempts were made to modify the model by adjusting allele frequencies based on average relatedness, re-estimating relatedness, and iterating this process. However, they proved unsuccessful.

Therefore, in the reconstruction of genealogies, only the degree one relationships will be directly incorporated. In each case, this will leave certain "gaps" in the genealogy, due to the general underestimation of degrees of relationship. At other times, false relationships will appear. An informal attempt will be made to detect the false missing

relationships using the degree two relationships, information on "degree infinity" relationships (unrelateds) and, if available, age structure.

Of all the model runs using Genealogy 1, the best results are obtained with the largest number of loci (30) and the smallest allele frequencies (2% for each allele). Figure 4.1 illustrates the results for this run with no mutations. In fact, this is precisely Genealogy 1 with the exception of the fact that individuals 28 and 29 are siblings. The model reports these as degree two relatives. It also says that 13 and 14 are unrelated. With the degree one relationships given for 13, 14, 28 and 29, either 13 and 14 must be siblings, or 28 and 29 must be siblings (assuming no inbreeding). Even in the absence of knowledge of the age structure, it would be suspected that 28 and 29 are siblings and therefore are degree one relatives. If there were a knowledge of the age structure, that is if it were known that individuals 13 and 14 both occurred in the generation preceding 28 and 29, it would have to be the case, accepting the degree one relationships, that 28 and 29 are siblings. In this case, Genealogy 1 is perfectly reproduced. Each additional error of the model, each of which involves a confusion of degree two and unrelateds, can also be detected from Figure 4.1. A similar analysis will hold for the runs with: 30 loci, allele frequencies of 5% each; 20 loci, allele frequencies of 2% each; and 20 loci, allele frequencies of 5% each. In each case, a single degree one relationship is omitted and can be inferred.

Figure 4.2 shows the results for Genealogy 1 with 20 loci and allele frequencies of 10% for each allele. Again, only the first degree relationships are shown. As in the above run, a pair of siblings, individuals 26 and 27, was missed and the relationship was said to be degree two. As above, the relationship can be inferred. A problem a bit more difficult to resolve, is the alleged degree one relationship between individuals 20 and 33.

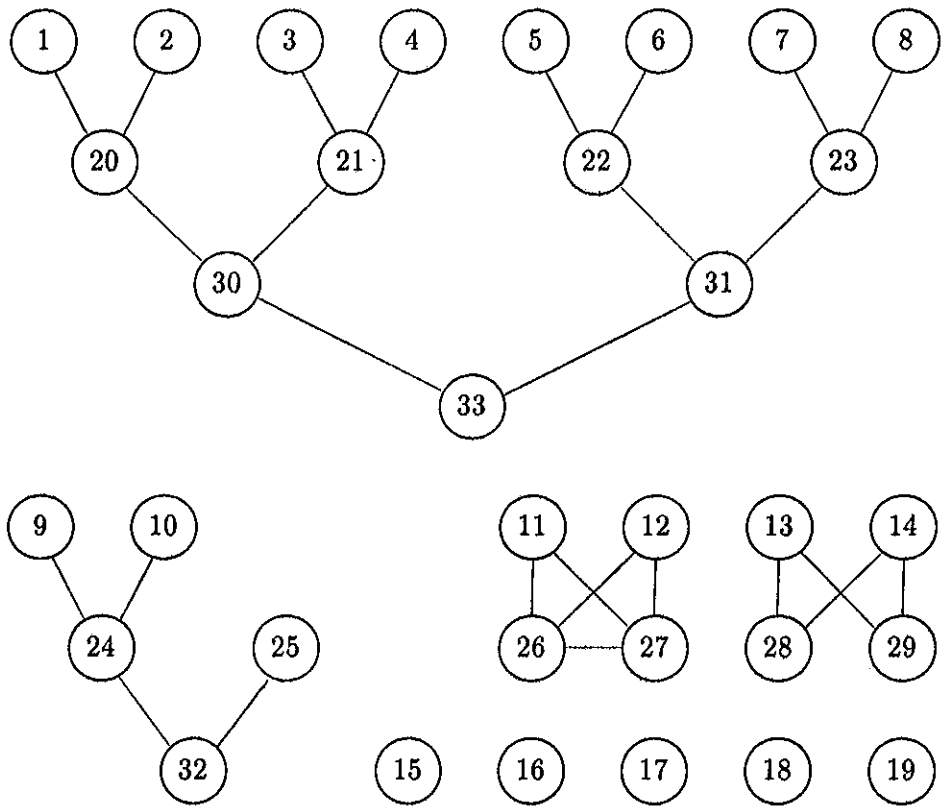

Figure 4.1: Degree one relationships as determined by the model using Genealogy 1, 30 loci and 2% allele frequencies. This is precisely Genealogy 1 with the exception of the fact that 28 and 29 are siblings. In this run, they were reported as degree two relatives.

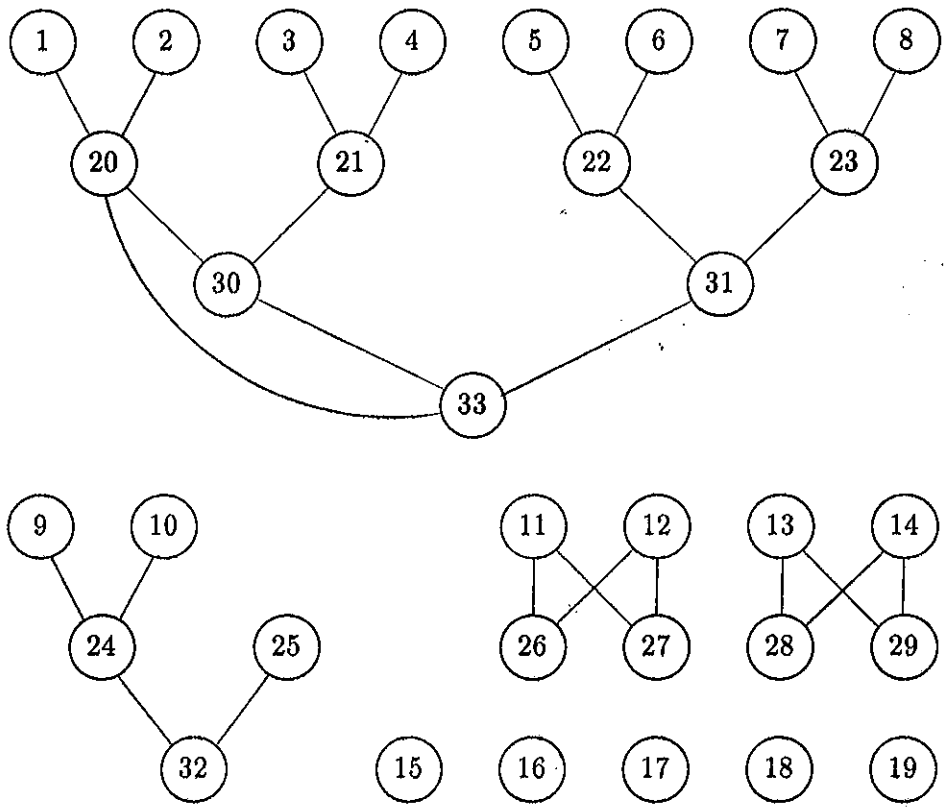

Figure 4.2: Degree one relationships as determined by the model using Genealogy 1, 20 loci and 10% allele frequencies.

If there is knowledge of age structure, the problem can be resolved. Since 20 appears two generations earlier than 30, it is unlikely that these two could be degree one relatives. In the absence of this information, it might be possible that 20 and 30 are siblings with 33 as one parent and the other parent not occurring in the sample. However, this would contradict several alleged unrelated pairs. For example, this would require 31 to be a degree two relative (grandparent) of both 20 and 30. The model says that 31 is not related to either of these. It would also require a degree two relationship between 20 and 21, which the model does not indicate. Nonetheless, if the degree two relationships do not carry a great deal of confidence, certainly the best way to resolve the dilemma is from the age structure. Again, the model is very successful in reconstructing the genealogy, although somewhat less successful than in the above discussed runs.

Pushing the model a bit, Figure 4.3 illustrates the degree one results in the weakest run of Genealogy 1 not involving mutations. The model used 10 loci with allele frequencies of 20% for each allele. This corresponds to a total of at most 50 traits detected in the population. As can be seen, several degree one relationships are missing, three degree two relationships were mistaken for degree one, a third degree relationship was mistaken for a first degree and one pair that was actually unrelated, was said to be in a first degree relationship. In fact, as an artifact of the small number of alleles present in the population, several relationships were overestimated (see Tables 3.1 and 3.2). The model did a fair job of breaking the sample into related groups. Only one individual (17) was pulled into a group to which it was not related, although two of the related groups were fragmented into smaller related groups. The presence of the knowledge of age structure in this case would be of little help, although some of the erroneous degree

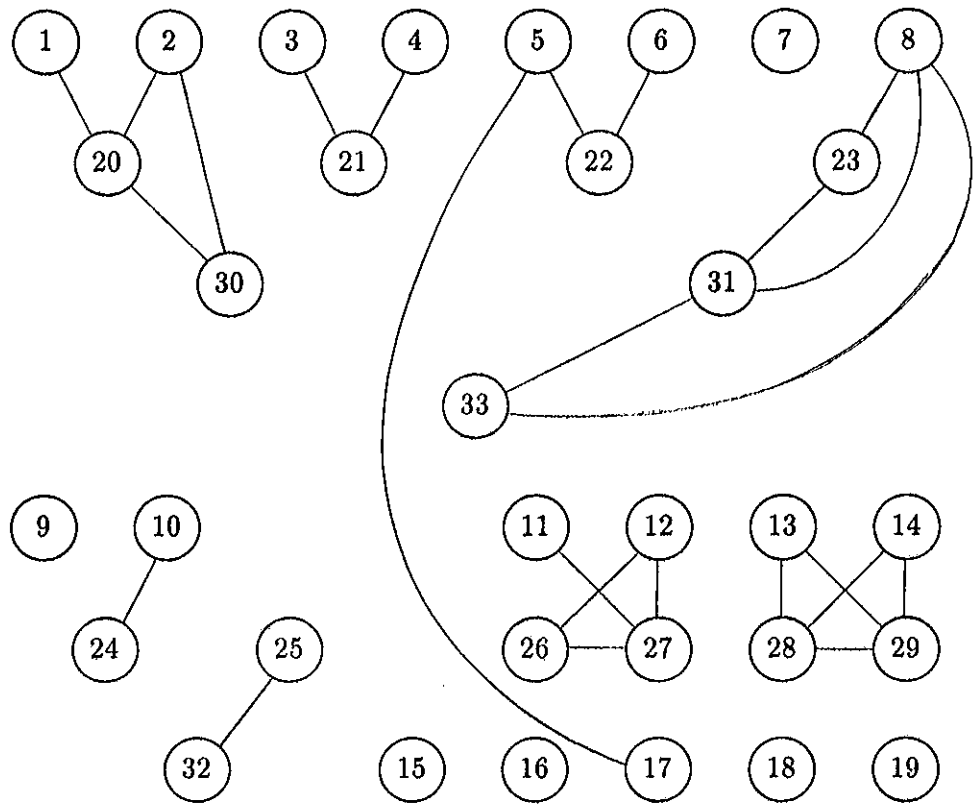

Figure 4.3: Degree one relationships as determined by the model using Genealogy 1, 10 loci and 20% allele frequencies.

one relationships might be detected. It appears that the model is not very successful with genealogy reconstruction with the limited data used as input in this run.

The effect of mutations is now considered. The programs described in Chapter 2 are set up in such a way that when a mutation (or laboratory error) occurs, it exchanges one allele at the given locus for another (not necessarily different) allele at that locus. This means that if there are a large number of alleles at a particular locus, mutations are not likely to make two individuals appear more closely related. However, if there are only a few alleles at a locus, it is quite possible that mutations could make individuals appear more closely related than they actually are. This trend can be observed with Genealogy 1 in Tables 3.1-3.4. This anomaly will have quite an effect on the output.

Figure 4.4 gives the results for Genealogy 1 with 20 loci, 2% frequencies for each allele and a mutation rate of 10% per allele. Since with small allele frequencies, mutations make individuals appear more distantly related, it is not surprising that each time the model indicates a degree one relationship, it is correct. Several of the degree two relationships indicated in the model output are verified from the degree one relationships. However, some are not. These are indicated with dotted lines in Figure 4.4. Including these corrections, almost gives the correct topology of the genealogy, although some of the distances are not accurate. Again, some knowledge of the age structure would be very helpful here. If some confidence is put in the second degree relationships, the model is quite successful in this case. Figure 4.5 gives the results for 2% allele frequencies, 20 loci and a 20% mutation rate. Including the degree two relationships introduces two topological errors (one involving 1, 20 and 30 and the other involving 10, 24 and 32). Again, the model is reasonably accurate, especially if age structure data is available.

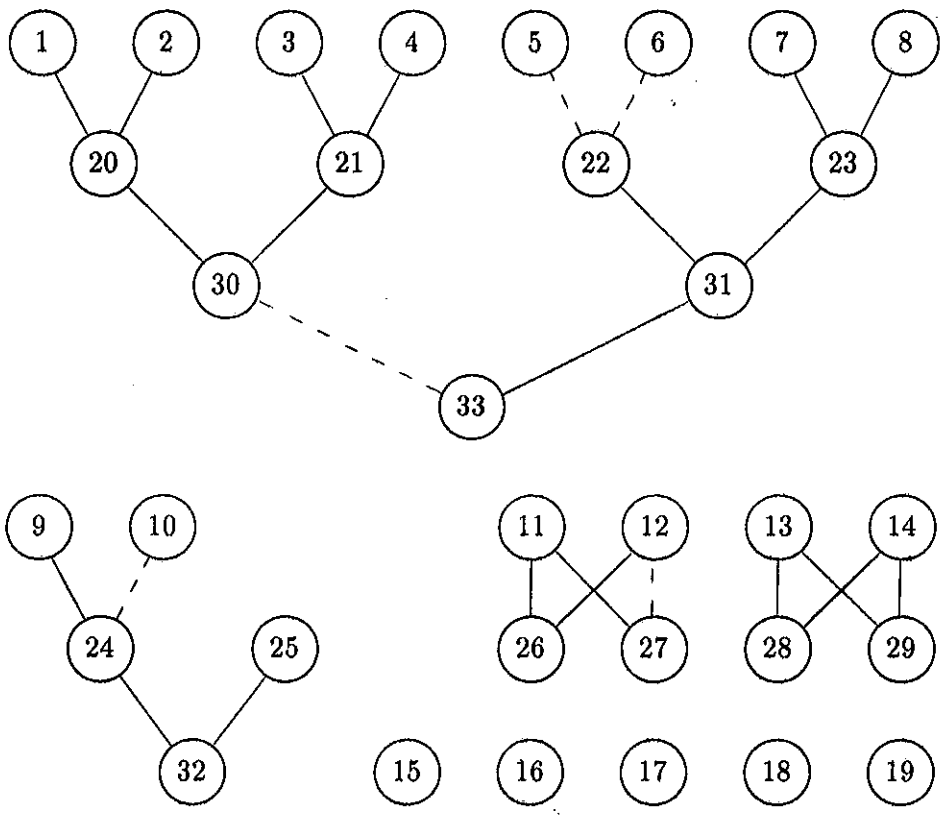

Figure 4.4: Degree one and degree two relationships as determined by the model using Genealogy 1, 20 loci, 2% allele frequencies and a 10% mutation rate. The solid lines represent degee one relationshps and the dotted lines represent the degree two output that is not explained by degree one relationships.

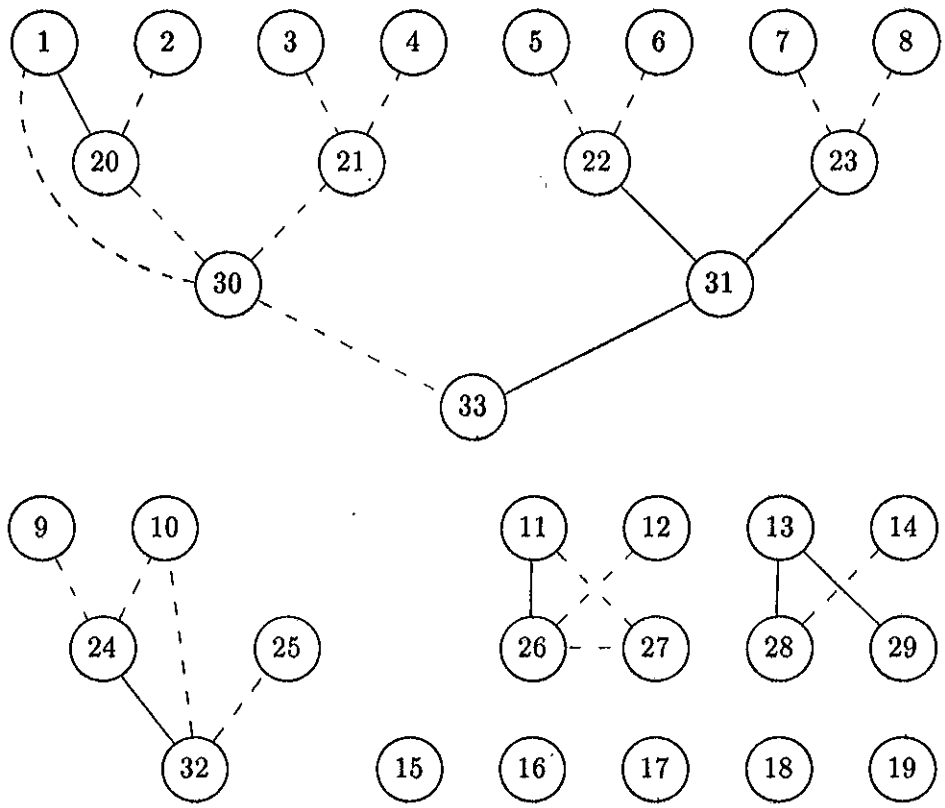

Figure 4.5: Degree one and degree two relationships as determined by the model using Genealogy 1, 20 loci, 2% allele frequencies and a 20% mutation rate. The solid lines represent degee one relationshps and the dotted lines represent the degree two output that is not explained by degree one relationships.

These results indicate that with small allele frequencies, it is still quite possible to reconstruct genealogies, even with a very high mutation rate (or laboratory error). Similar results, although somewhat less precise, are also obtained from the runs where a 5% allele frequency is used.

The model becomes much less reliable with a 20% mutation rate, as the allele frequencies become larger. Figures 4.6 and 4.7 show the results for Genealogy 1 with 20 loci, 10% allele frequencies and a 20% mutation rate. Figure 4.6 includes the first degree relations only. The model is rather poor at recognizing these relationships, although a great deal of confidence can be put in a "first degree relationship" output. In this case, the second degree relationships are of little help (see Figure 4.7). All of the actual first degree relationships were said to be either first degree or second degree. However, five pairs of unrelateds were said to be degree two relatives. Again, this is a combination of the high mutation rate and the few number of alleles allowable at a particular locus. Adding the second degree relationships to the first degree, give some idea of the structure of the genealogy. However, the extra relationships effectively mask the true intricacies of the genealogy. It seems that the model is not very successful at genealogy reconstruction under these conditions of high mutation rates and high allele frequencies.

4.2 Genealogy 2

As with Genealogy 1, a great deal of confidence can be put in the model when it says two individuals are first degree relatives or are unrelated when using Genealogy 2 (see Tables 3.5 and 3.7). In the absence of mutations, there is at least a 97% confidence in degree one output. With mutations this confidence rises to 100%. Without mutations,

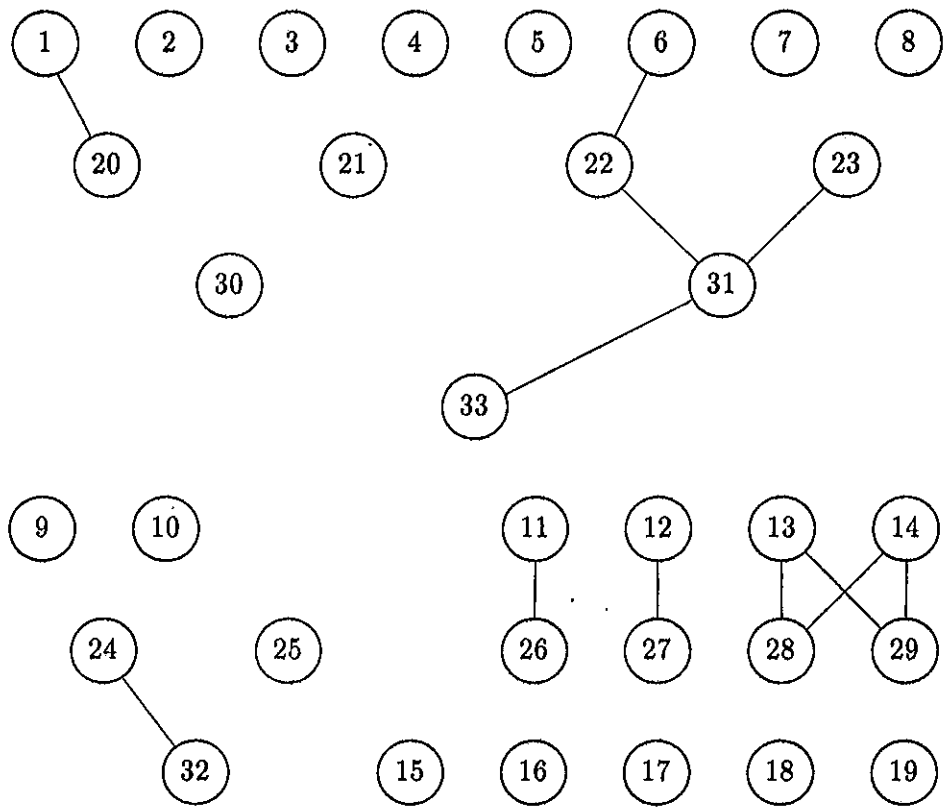

Figure 4.6: Degree one relationships as determined by the model using Genealogy 1, 20 loci, 10% allele frequencies and a 20% mutation rate.

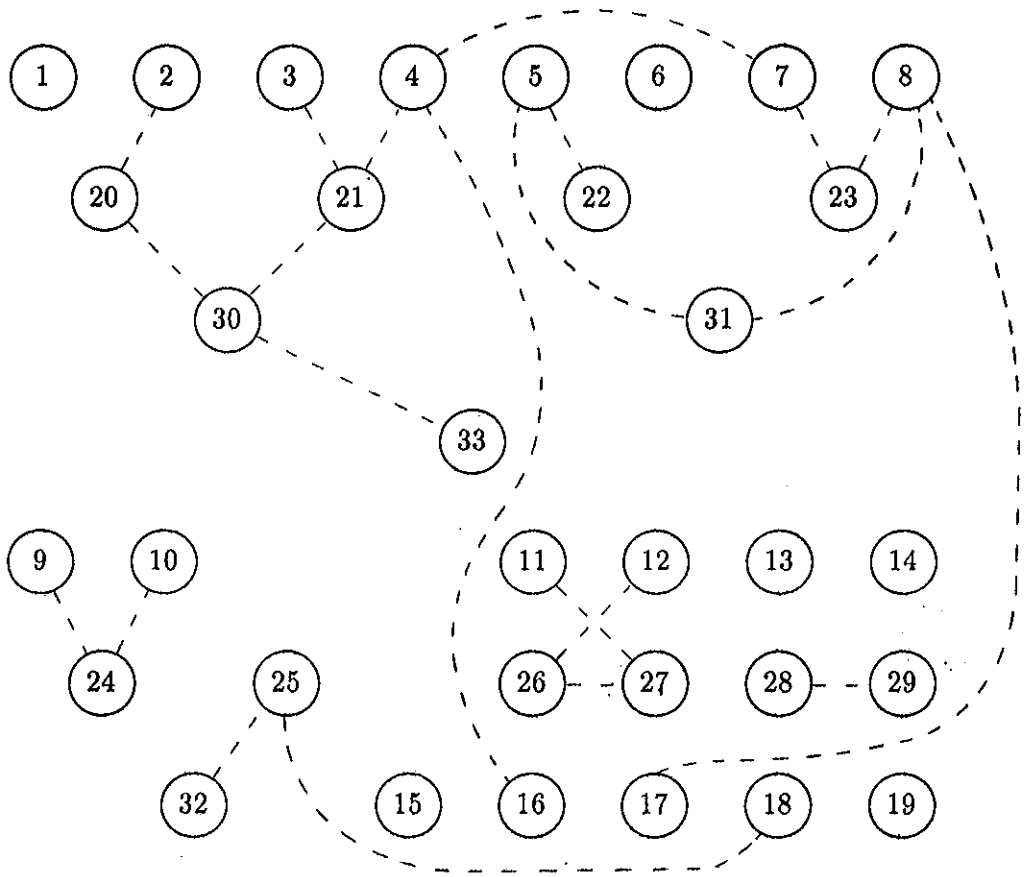

Figure 4.7: Degree two relationships as determined by the model using Genealogy 1, 20 loci, 10% allele frequencies and a 20% mutation rate.

the percentage of actual degree one relationships that are recognized is comparable to the results for Genealogy 1 (in fact, they are a bit better). With mutations, the percentage of degree one relatives recognized is a bit less than it was for Genealogy 1 (compare Tables 3.4 and 3.8). The model is very successful with unrelated pairs. One can put 95% confidence in an "unrelated" output in all tested cases except for two (both of which involve mutation rates of 30%, but there is still a 94% confidence in these worst cases, see Table 3.7). From Tables 3.6 and 3.8, one can see that the model recognizes over 95% of the unrelated pairs, except in one case (in which there are the fewest number of loci [10] and the fewest number of alleles per locus [5]).

The model is less successful with the second degree relationships. Table 3.5 reveals that both the level of confidence and the percentage of recognition decrease as allele frequencies increase. There is occasionally a high degree of confidence in the degree two output, but quite often the degree of confidence is less than 95%. However, in each case the confidence in degree two output for Genealogy 2 was either comparable to or better than the analogous output for Genealogy 1 (compare Tables 3.1 and 3.5) in the absence of mutations. The percentage of degree two relationships recognized for Genealogy 2 are also comparable to the results for Genealogy 1 in the absence of mutations, although the deviation from a general trend is a bit larger here (compare Tables 3.2 and 3.6). In general, with mutations present, the level of confidence in degree two output drops significantly (except in the cases of 10% mutation rates and 10% and 20% allele frequencies, compare Tables 3.5 and 3.7). In fact, the highest level of confidence in the cases tested is 33%. One sees a similar trend when comparing the

percentage of degree two relationships recognized with and without mutations (compare Tables 3.6 and 3.8).

Again, a source of error is the determination of allele frequencies, due to the sampling of related individuals. However, this problem is somewhat alleviated by the fact that Genealogy 2 consists of several unrelated groups. The bias introduced with the relatedness within groups is somewhat balanced by the nonrelatedness between groups.

Figure 4.8 presents the model output for degree one relationships for Genealogy 2 when ran with 30 loci and allele frequencies of 2%. All degree one relationships are correctly identified except for one (individuals 25 and 26 are siblings). As in the first discussed run of Genealogy 1, the sibling relationship between 25 and 26 can be inferred from either the degree two relationships or from the age structure. A similar analysis holds for the case of 30 loci and 5% allele frequencies. Interestingly, the model is a bit more successful when there are only 20 loci. With 20 loci and allele frequencies of 2%, 5% and 10 %, all first degree relationships are identified and there is 100% confidence in the "first degree" output. This means that the genealogy is perfectly reconstructed in these cases.

The run for Genealogy 2 with 10 loci and 2% allele frequencies is presented in Figure 4.9. Again, the genealogy is reconstructed from the degree one relationships, with the exception of four pairs of siblings and again, this relationship can be inferred. These results seem to indicate that, even with a small number of loci (10), if the traits being scored are rare enough (2% allele frequency), then the genealogy can be correctly reconstructed.

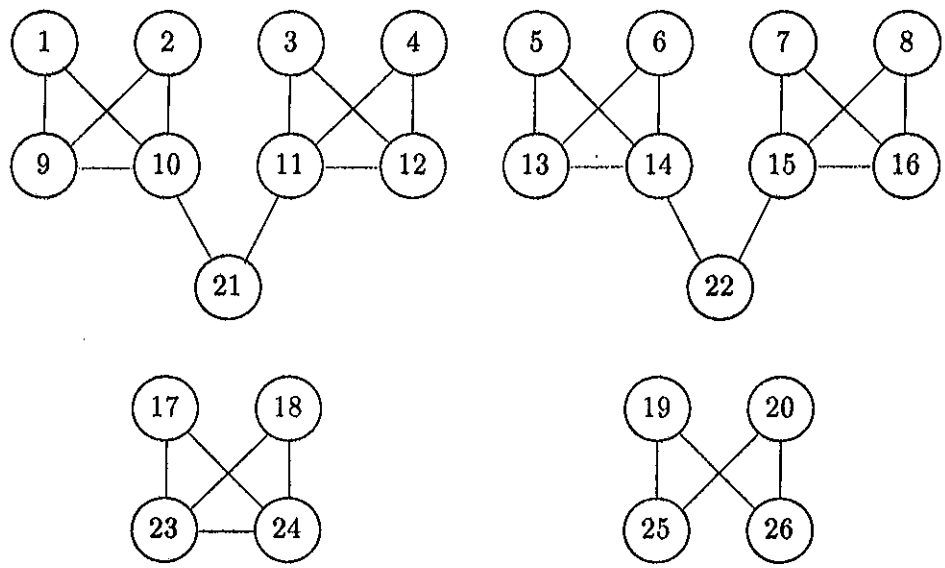

Figure 4.8: Degree one relationships as determined by the model using Genealogy 2, 30 loci and 2% allele frequencies.

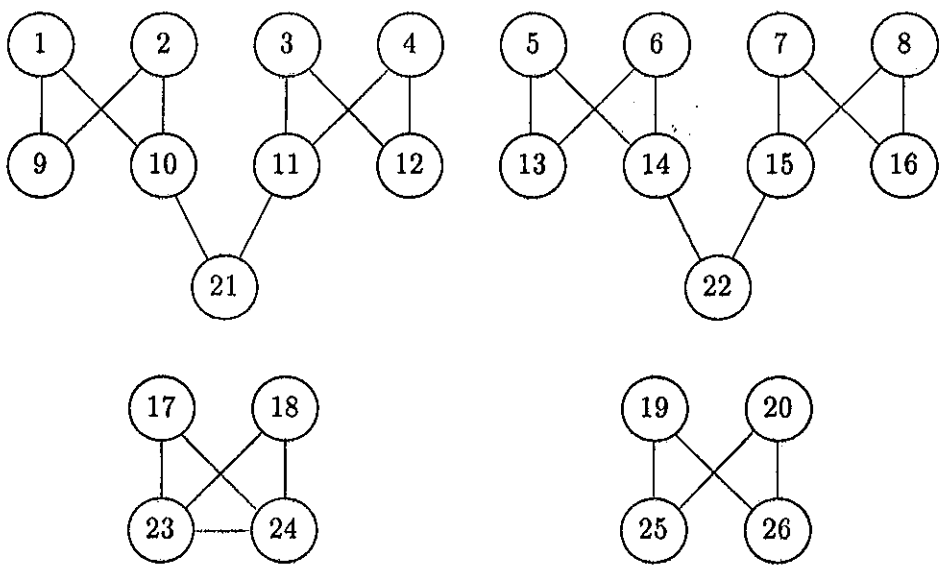

Figure 4.9: Degree one relationships as determined by the model using Genealogy 2, 10 loci and 2% allele frequencies.

Figure 4.10 illustrates the degree one results for Genealogy 2 with 20 loci and allele frequencies of 20%. One false degree one relationship is present (individuals 9 and 10 are actually degree two relatives). However, no false connections have been made between unrelated groups. The population has been broken into related groups, although the groups are not as large as they should be. Including the degree two relationships starts to cloud the structure. It seems that the model is only marginally successful in this case.

Tables 3.7 and 3.8 give the results for Genealogy 2 when mutations are present. Similar to Genealogy 1, a great deal of confidence can be put in degree one output. In fact, everytime the model said a relationship was degree one, with mutations present, it was correct. The percent of degree one relationships recognized was much less, as was the case in Genealogy 1. In fact, the percentages are for the most part, comparable on this point between the two genealogies, although Genealogy 1 does do better in general with these.

Figure 4.11 gives the degree one output of the model for Genealogy 2 with 20 loci, 2% allele frequencies, and a 10% mutation rate. Several of the degree two relationships are also given as dotted lines. The topology is correct except for the 11-12-21 triangle and the 13-14-22 triangle. Since 11 and 12 are each degree one relatives of both 3 and 4, it must be (assuming no inbreeding) that either 3 and 4 are siblings or 11 and 12 are siblings. Since 11 and 12 are reported to be degree two relatives and 3 and 4 are reported to be unrelated, it is more likely that 11 and 12 are the pair of siblings (this same analysis was carried out above). Also, if 11 and 12 are siblings, this resolves the degree two relationship between 12 and 21. This dilemma can be similarly resolved with

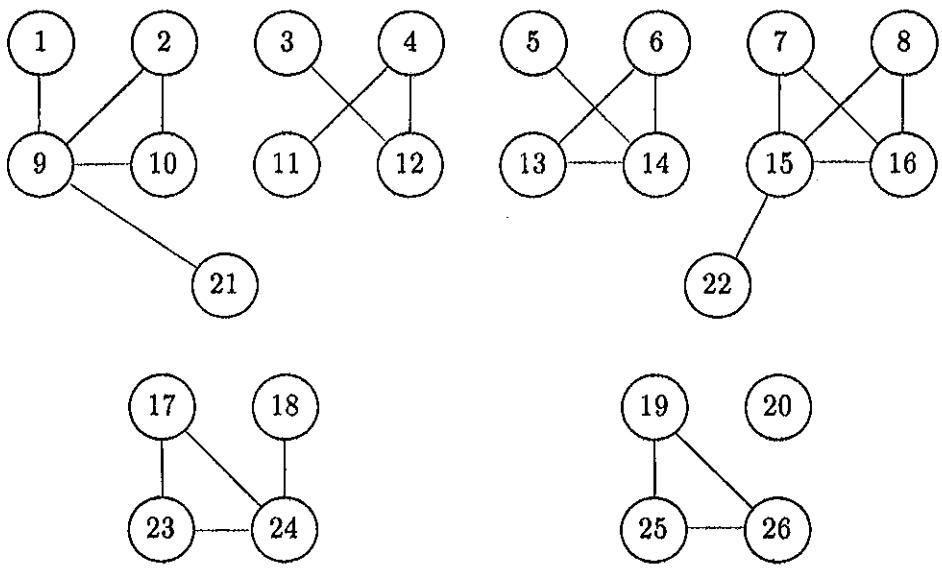

Figure 4.10: Degree one relationships as determined by the model using Genealogy 2, 20 loci and 20% allele frequencies.

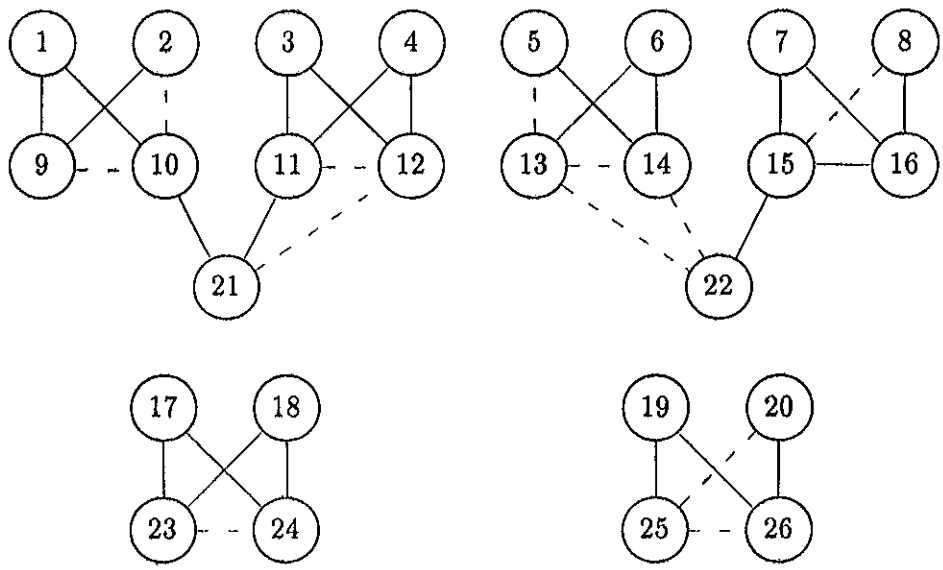

Figure 4.11: Degree one and degree two relationships as determined by the model using Genealogy 2, 20 loci, 2% allele frequencies and a 10% mutation rate. The solid lines represent degree one and the dotted lines represent some of the degree two output.

age structure information. The 13-14-22 triangle is a bit more difficult to resolve. One option, is that the three are siblings. This means that the relationship between 6 and 22 (reported as degree two) was underestimated, and the relationship between 5 and 22 (reported as unrelated) was grossly underestimated. This is possible, although unlikely. Also, the possibility could quickly be ruled out in the presence of age structure . Another option is that 22 is the niece or nephew of 13 and 14 and that the parent of 22, a sibling of 13 and 14, is not present in the sample. This possibility leaves all reported degrees of relationship in tact and also, age structure would not be of any help in ruling this out. If there is not a great deal of confidence in the reported degree two relationships, another possibility is that 13 and 14 are siblings and one of them is a parent of 22 and the other is an aunt or uncle. Simply from the degrees of relationship, it is impossible to tell which is the parent and which the aunt or uncle, however. Again, age structure is not a great deal of help with this particular point. However, if it is known that individual 15 occurs in the generation before 22 and their relationship is the first degree, then 15 must be a parent of 22. One other piece of information that is easily obtained but has not yet been discussed could be useful at this point. Since the sex of a pair of parents must be different, if 15 is one parent of 22, the other parent must certainly be of the opposite sex. Again, however, this may not completely solve the problem. If it is suspected that 15 is one parent and either 13 or 14 is the other, closer scrutiny of the raw data might be reveilling. In the absence of mutation, every dominant trait that an individual expresses must be inherited from one of the parents. Searching the phenotypes of the four individuals concerned with this in mind, might provide additional insight. However, mutations have not been ruled out in this case. Another approach might involve the

use of rare or "unique" alleles. See the "Discussion" section of this chapter for more on the use of unique markers. This example illustrates the difficulty involved in the type of analysis being attempted. Quite often, the simple data of pairwise degrees of relationship will be insufficient to accurately reconstruct the true genealogy. The results of this run are similar to the results for 20 loci, 2% allele frequencies and a 20% mutation rate and for 20 loci, 5% allele frequencies and a 10% mutation rate, although less degree one relationships were recognized in these runs.

As above, high mutation rates make the relationships seem more distant. Figure 4.12 gives the results of the model when run with Genealogy 2, 20 loci, 10% allele frequencies and a 30% mutation rate. Using the first and second degree relationships gives the approximate shape of the genealogy. However, as above, there is a problem with a 9-10-21 triangle. Since 10 and 21 are said to be (with a good deal of confidence) degree one relatives, the dilemma is more easily resolved this time. The mutations have had another severe effect. Three pairs of first degree relatives were said to be unrelated (6-14, 15-16, and 23-24). It seems that the model is weak in this case. Its performance is even less accurate with larger allele frequencies.

4.3 Genealogy 3

Recall that Genealogy 3 is a population of 10 unrelated individuals. The model correctly identified each relationship in all tested cases of loci number, allele frequencies and mutation rates. This is an important point. Even though the population sample was small, the structure was correctly identified. The source of error identified above, namely poor estimates of allele frequencies due to relatedness in the sample, is not a

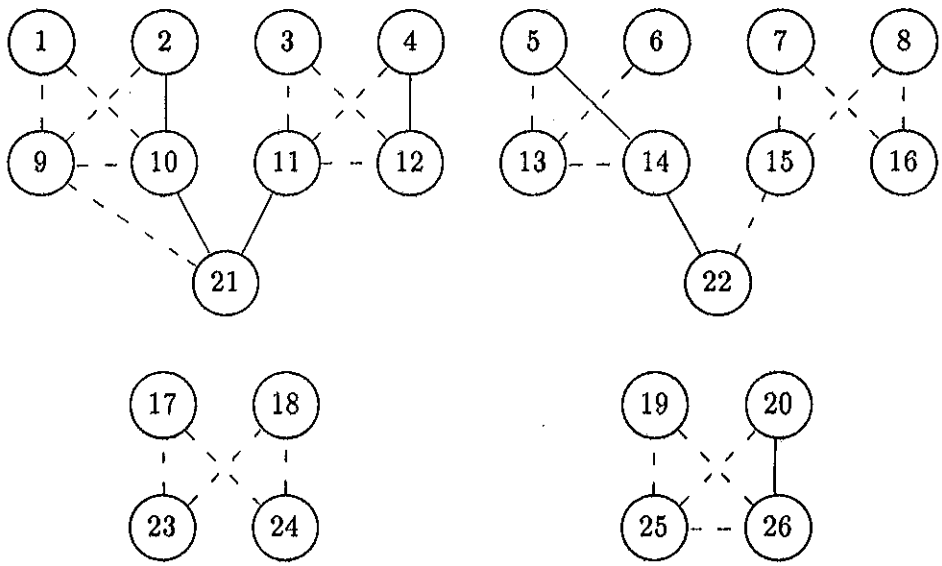

Figure 4.12: Degree one and degree two relationships as determined by the model using Genealogy 2, 20 loci, 10% allele frequencies and a 30% mutation rate. The solid lines represent degree one and the dotted lines represent some of the degree two output that is not explained by degree one relationships.

problem in this case. This indicates that the presence of unrelateds in a sample is desirable. They enhance the estimates of allele frequencies and help refine the degree of relationship estimates between related individuals. See the "Discussion" section below for more details.

4.4 Genealogies 4, 5 and 6

Genealogy 4 involves a high degree of inbreeding. The results for this model are presented in a slightly different form. Since the model only tests for "degree one", "degree two", and "degree 3 or higher" relationships, it will be difficult to reconcile the output with the actual relationships. To shed some light on this, consider the so called *coefficient of kinship*, F_{JK}, between two individuals J and K. This is defined as

$$F_{JK} = \sum \left(\frac{1}{2}\right)^{N+1} \tag{4.1}$$

where N is the number of steps in a path from J to a common ancestor and back to K, and the summation is over all such paths [Smith, 1989]. This formula assumes that the common ancestors are not inbred and are unrelated. In the absence of inbreeding, the J and K are n-degree relatives where

$$n = -\log_2(2F_{JK}) = \frac{-\log(2F_{JK})}{\log 2} = -1 - \frac{\log F_{JK}}{\log 2}. \tag{4.2}$$

Since Genealogy 4 involves inbreeding and the model only outputs degrees of relationship, define the degree of relationship, n, between two inbred individuals by equation (4.2). Table 4.1 gives the coefficients of relatedness for the pairs of individuals of Genealogy 4

Table 4.1: The coefficients of kinship for the pairs of individuals in Genealogy 4 calculated from equation (4.1).

	1	2	3	4	5	6	7	8	9	10
1	-	0	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4
2	0	-	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4
3	1/4	1/4	-	1/4	1/4	1/4	3/8	3/8	1/4	1/4
4	1/4	1/4	1/4	-	1/4	1/4	3/8	3/8	1/4	1/4
5	1/4	1/4	1/4	1/4	-	1/4	1/4	1/4	1/4	1/4
6	1/4	1/4	1/4	1/4	1/4	-	1/4	1/4	1/4	3/8
7	1/4	1/4	3/8	3/8	1/4	1/4	-	3/8	7/16	9/32
8	1/4	1/4	3/8	3/8	1/4	1/4	3/8	-	7/16	9/32
9	1/4	1/4	1/4	1/4	1/4	1/4	7/16	7/16	-	3/8
10	1/4	1/4	1/4	1/4	1/4	3/8	9/32	9/32	3/8	-

and Table 4.2 gives the degrees of relationship as defined above. In every case except for the founders, 1 and 2, the degree of relationship between pairs is one or greater. So, the best possible output from the model would be a "degree one" output for each of these relationships. Certain relationships are particularly close, for example between 7 and 9, and between 8 and 9, and even if the model has trouble accurately recognizing the other relationships, it should recognize these.

Table 3.9 presents a brief summary of the results of running Genealogy 4 in the model with 20 loci and various allele frequencies and mutation rates as input. In each run, either none, one, or two degree-one relationships were recognized. In the absence of mutations, two degree-one relationships were recognized and in each case these were the pair 7 and 9 and the pair 8 and 9. It might initially seem surprising that the model performs so poorly. The reason for this lies in the fact that allele frequencies are very badly estimated. The estimates are calculated assuming the individuals are unrelated. The only pair of unrelated individuals in the sample is the pair 1 and 2. This bias is reflected in the fact

Table 4.2: The degrees of relationship for the pairs of individuals in Genealogy 4 which are calculated using equation (4.2).

	1	2	3	4	5	6	7	8	9	10
1	-	∞	1	1	1	1	1	1	1	1
2	1	-	1	1	1	1	1	1	1	1
3	1	1	-	1	1	1	0.415	0.415	1	1
4	1	1	1	-	1	1	0.415	0.415	1	1
5	1	1	1	1	-	1	1	1	1	1
6	1	1	1	1	1	-	1	1	1	0.415
7	1	1	0.415	0.415	1	1	-	0.415	0.193	0.830
8	1	1	0.415	0.415	1	1	0.415	-	0.193	0.830
9	1	1	1	1	1	1	0.193	0.193	-	0.415
10	1	1	1	1	1	0.415	0.830	0.830	0.415	-

that only a small percentage of the pairs are correctly given their degree of relationship. As discussed above, relatedness has the effect of making individuals appear less closely related. It follows that the output would be greatly improved if the allele frequencies could be better estimated.

Genealogy 5 is an attempt to do just that. Genealogy 5, recall, includes all the individuals of Genealogy 4 plus 10 new individuals unrelated to any others in the sample. It is expected that the presence of this new outgroup will dramatically improve allele frequency estimates. As Table 3.10 reveals, the percentage of first degree relationships recognized, increases roughly 10-fold over the results from Genealogy 4. The effects of mutations seem rather severe, with a 10% mutation rate cutting the percentage of degree one relationships recognized in half. One further statistic that can be associated with the output is the percentage of pairs of unrelateds mistaken for related pairs. This error occurs a small percentage of the time (see Table 3.10), mostly when the allele frequencies are relatively large. Again, however, the true degree of relationship between individuals

1 through 10 is not well recognized by the model, with the average recognition in the absence of mutations at around 43%.

Genealogy 6 consists of all members of Genealogy 5 along with an additional group of 10 individuals unrelated to all others in the sample population. The percentage of recognition of degree one relatives is much better than with Genealogy 5, at an average of 69% in the absence of mutations (see Table 3.11). Again, a mutation rate of 10% roughly cuts these values in half. It is apparent, however, that a large number of unrelateds would have to be present in the sample for the percentage of degree one relationships recognized to reach a substantial level (say 95%). This is due to the fact that any allele present in individuals 3 through 10 is derived either from individual 1 or 2, or is from a mutation. Frequency estimates will be large as long as these 10 individuals comprise a significant percentage of the sample population.

One interesting fact about Genealogy 4 is the following. Even if the model were to give the optimal output (all relationships first degree except for the pair 1 and 2), then it is still impossible to reconstruct the true genealogy. The output would indicate a closely related population, but the details of the structure are completely hidden. The presence of age structure would be a bit helpful, but the number of possible genealogies would still be staggering.

It seems that the model is at its worst when dealing with Genealogies 4, 5 and 6. That is, the model is least effective when dealing with a small population of closely related individuals. This is precisely the time, however, that other methods might be much more useful. Testing individuals pairwise to see if they are possible parents of other individuals in the sample or using unique markers to trace relationships both might shed light on the

true genealogical structure. Unique markers might be particularly useful in an inbred population. The use of unique markers in the absence of inbreeding is discussed in the next section.

4.5 Discussion

In this section, the data presented above are summarized and conclusions drawn. Strengths and limitations of the model are discussed.

The general trend throughout the model is that, in the absence of mutations, the number of alleles present in the population is directly proportional to the accuracy of the output of the model. In fact, if alleles were "unique", that is if the allele is present in only one lineage and the probability of it occuring in other lineages is zero, then it is expected that the model would give perfect output. In this case, unrelated individuals would share none of these markers. The use of unique markers in pedigree reconstruction has been explored in Wooten and Gardner [submitted manuscript]. Unique markers are unambiguous indicators of relatedness. The continued discovery of highly polymorphic sequences and the use of gene amplification techniques for both fragment detection and direct sequencing has vastly increased the ability to identify large numbers of individual-specific markers [Horn *et al.*, 1989 and Jeffreys *et al.* 1990]. If a large enough sample of unique markers could be identified, it might be possible to trace relationships over many generations. A question of interest is "How many unique markers must an individual possess for there to be a $1-\alpha$ probability that a degree n relative also possesses one of the markers?" For any given unique marker, the probability of sharing is $\left(\frac{1}{2}\right)^n$ and the probabilty of not sharing is $1-\left(\frac{1}{2}\right)^n$. If an individual possesses x unique markers,

the probability that none are shared with an n degree relative is $\left\{1-\left(\frac{1}{2}\right)\right\}^x$, and the probability that at least one is shared is $1-\left\{1-\left(\frac{1}{2}\right)\right\}^x$. So for there to be a $1-\alpha$ probability of shared markers, it must be that

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

or

$$x = \frac{\log \alpha}{\log(1-(\frac{1}{2})^n)}.$$

For $n = 1$ and $\alpha = 0.05$, one gets that $x = 4.3$. That is, an individual must possess 5 unique markers for one to have 95% confidence that this individual shares a unique marker with a degree one relative. For $n = 2$ and $\alpha = 0.05$, one gets that $x = 10.4$. With $n = 3$ and $\alpha = 0.05$, one has $x = 22.4$ and for $n = 4$ and $\alpha = 0.05$, one needs $x = 46.4$ unique markers. So it seems even with unique markers, it is difficult to carry pairwise determination of relationships very far (more than 2 or 3 steps) unless there is a rather large number of such markers.

The model almost always performed better when allele frequencies were low. This point was suggested in Chapter 1 and has been explored by Lynch [1988]. This suggests that relationships are rather accurately identified when allele frequencies are less than 5% or so. The degree one relationships are particularly well recognized. When dealing with real world data, it is easy to check the trait or band frequencies and to derive the associated allele frequencies from these. The genealogies above seem to indicate that if the observed allele frequencies are low, then one can put a fairly high degree of confidence in the output of the model.

With relatedness within a sample, estimations of allele frequencies will be high. This can be partially remedied by adding an "outgroup". The model was most successful with Genealogy 3 because there was no bias due to relatedness in the estimation of allele frequencies. The model was least successful with Genealogy 4 because of the very high bias in these estimates. The level of success was also reasonably high with Genealogies 1 and 2. The presence of related individuals, of course, produced a bias in frequency estimates; however in each population there was an absence of relationships between several groups. In particular, Genealogy 1 contains five individuals unrelated to all others in the sample. The presence of this division into related groups (and also the presence of unrelateds in Genealogy 1) helps compensate somewhat for the bias in frequency estimations. This would be especially true when the groups are of similar sizes. This also explains the lack of success with Genealogies 4, 5 and 6. As more unrelateds are introduced into the sample, allele frequency estimates become more accurate and the degrees of relatedness becomes more apparent. The high degree of relatedness between the 10 individuals of Genealogy 4 implies that many unrelateds will be needed to compensate for this bias. One interesting result from Genealogies 4, 5 and 6 is that the true allele frequencies used to generate the data seemed to have little effect on the effectiveness of the model when dealing with individuals 1 through 10 (see Tables 3.9, 3.10 and 3.11). This is due to the fact that the frequencies of alleles used to determine relatedness in this group were estimated based heavily on their presence in this group. That is, the estimates of allele frequencies were so dominated by the inbred group that alleles which would have been rare in a noninbred population, appeared rather common in this population. Again, this is a point where unique markers could be helpful.

Another trend that the model seems to follow is that it is more accurate with a larger number of loci. This is not true in general, instead holds most dramatically when allele frequencies are large. The model is set up in such a way (see Chapter 2) that if m loci are used, each individual will display between m and $2m$ of the possible traits. This indicates that when higher frequency markers are used in genealogical reconstruction, more markers (data) must be employed.

Under the best circumstances tested, the model is quite accurate with second degree relationships. However, the greatest confidence is in the degree one output. In an ideal situation, all necessary degree one relatives would be present and these could be used to check the degree two output. Advantage was taken of the fact that all intermediate individuals were present between two relateds in the above discussed genealogies. In the absence of the relevant degree one relatives, if two individuals are said to be second degree relatives, it can only be assumed that the output is correct and the relatives linking these two together are not present in the sample. This could be a problem when little confidence is put in the degree two output.

One very strong point of the model is its behavior in the presence of "mutations". The situation written into the model is to have a certain percentage of the alleles randomly changed to other alleles. This could correspond not only to a mutation, but to an error in scoring the raw data in the laboratory. The model is adversely affected by such action, although in the better cases that were tested, still fairly accurate when faced with the astronomical mutation rates of 10% or 20% per allele. Certainly no natural population could have such a mutation rate, and it is hoped that no laboratory could be so poor at recording data!

4.6 Conclusion

The model proposed in this thesis, namely genealogical reconstruction within a population using pairwise distances determined with a maximum likelihood algorithm, seems quite useful under a certain set of circumstances. The model is strong when

1. it is applied to a large population consisting of several unrlelated groups or having a large number of individuals unrelated to all other members of the sample,
2. the number of markers used as raw data are low in frequency, with associated allele frequencies of around 5% or less,
3. there are a sufficient number of pairs of degree one relatives in the sample to piece together more distant relationships, and when
4. the rate of mutation or laboratory error is small ("small" might be taken to be less than 10%).

The model is less successful when

1. it is applied to small populations which have a high average degree of relatedness or a high level of inbreeding,
2. the markers used as raw data are few in number and high in frequency,
3. the population only contains distantly related pairs and the genealogy must be reconstructed without the intermediate relatives, and when
4. there is a very high rate of mutation or laboratory error.

When applied to the appropriate situation, this model should perform quite satisfactorally and provide the population geneticist, conservationist, ecologist, or field biologist with valuable information that would otherwise be unavailable.

Bibliography

Bondy, J., and Murty, U. 1976. Graph Theory with Applications. Elsevier Science Publishing Company, New York, NY.

Burke, T., and M. Bruford. 1987. DNA Fingerprinting in Birds. Nature 327:149-152.

Cannings, C., and E. Thompson. 1981. Genealogical and Genetic Structure. Cambridge University Press, Cambridge, MA.

Felsenstein, J. 1981. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. Journal of Molecular Evolution 17:368-376.

Felsenstein, J. 1982. Numerical Methods for Inferring Evolutionary Trees. The Quarterly Review of Biology 57:379-404.

Felsenstein, J. 1983. Statistical Inference of Phylogenies. Journal of the Royal Statistical Society-A 146:246-272.

Gill, P., Lygo, J., Fowler, S., and D. Werrett. 1987. An Evaluation of DNA Fingerprinting for Forensic Purposes. Electrophoresis 8:38-44.

Hill, W. 1986. DNA Fingerprint Analysis in Immigration Test-Case. Nature 322:290-291.

Hillis, D. 1987. Molecular Versus Morphological Approaches to Systematics. Annual Review of Ecology and Systematics 18:23-42.

Horn, G., Richards, B., and K. Klinger. 1989. Amplification of a Highly Polymorphic VNTR Segment by the Polymerase Chain Reaction. Nucl. Acid Res. 17:2140.

Jeffreys, A., Brookfield, F., and R. Semeonoff. 1985. Positive Identification of an Immigration Test-Case using Human DNA Fingerprints. Nature 317:818-819.

Jeffreys, A., Wilson, V., and S. Thein. 1985a. Hypervariable 'Minisatellite' Regions in Human DNA. Nature 314:67-73

. 1985b. Individual-Specific 'Fingerprints' of Human DNA. Nature 316:76-79.

Jeffreys, A., Neumann, R., and V. Wilson. 1990. Repeat Unit Sequence Variation in Minisatellites: A Novel Source of DNA Polymorphism for Studying Variation and Mutation by Single Molecular Analysis. Cell 60:473-485.

Lewin, R. 1989. Limits to DNA Fingerprinting. Science 243:1549-1551.

LYNCH, M. 1988. Estimation of Relatedness by DNA Fingerprinting. Molecular Biology and Evolution 5(5):584-599.

PAMILO, P., AND M. NEI. 1988. Relationships between Gene Trees and Species Trees. Molecular Biology and Evolution 5(5):568-583.

SMITH, J. MAYNARD. 1989. Evolutionary Genetics. Oxford University Press, New York, NY.

THOMPSON, E. 1986. Pedigree Analysis in Human Genetics. Johns Hopkins University Press, Baltimore, MD.

WETTON, J., CARTER, R., PARKIN, D., AND D. WALTERS. 1987. Demographic Study of a Wild House Sparrow Population by DNA Fingerprinting. Nature 327:147-149.

WOOTEN, M. AND R. GARDNER. submitted manuscipt. A Comment on the Number of Unique Markers Required for Pedigree Reconstruction.

WRIGHT, S. 1922. Coefficients of Inbreeding and Relationship. American Naturalist 56:330-338.

Appendix

This appendix includes listings of the fortran programs FREQDAT.FOR, GENDAT.FOR, GENOME.FOR, RECONSTR.FOR and MATRIX.FOR. The listings are the versions of the programs which were run in Microsoft Fortran on an IBM compatible personal computer.

GENDAT.FOR

```
C
C   GENDAT.FOR
C
C     THIS PROGRAM ALLOWS FOR THE INPUT OF DATA THAT
C     DESCRIBES A GENEALOGY FOR A CERTAIN POPULATION
C     OF SIZE N.  IT IS THE ADJACENCY MATRIX FOR THE
C     PARENT TO OFFSPRING DIRECTED GRAPH.
C
      DIMENSION IG(50,50)
      CHARACTER*20 GENEALFILE
      WRITE(*,*)'THIS PROGRAM IS FOR THE INPUT OF THE'
      WRITE(*,*)'ADJACENCY MATRIX FOR THE'
      WRITE(*,*)'PARENT-OFFSPRING DIRECTED GRAPH.'
      WRITE(*,*)' '
      WRITE(*,*)'IN WHAT FILE WILL THE GENEALOGY BE STORED?'
      READ(*,10)GENEALFILE
   10 FORMAT(A9)
      OPEN (UNIT=1, FILE = GENEALFILE, STATUS='OLD')
      WRITE(*,*)' '
      WRITE(*,*)'WHAT IS THE SIZE OF THE POPULATION?'
      READ(*,*)N
      WRITE(1,*)N
  200 WRITE(*,*)'WHICH EDGES DO YOU WANT TO INCLUDE?'
      WRITE(*,*)'INPUT "0,0" TO END'
      WRITE(*,*)'INPUT PARENT#, OFFSPRING#'
      READ(*,*)J,I
      IF(I.EQ.0) GO TO 300
      IG(I,J)=1
      GOTO 200
  300 DO 400 I=1,N
      DO 400 J=1,N
      WRITE(1,*)IG(I,J)
  400 CONTINUE
      END
```

FREQDAT.FOR

```
C
C   GENDAT.FOR
C
C     THIS PROGRAM ALLOWS FOR THE INPUT OF DATA THAT
C     DESCRIBES A GENEALOGY FOR A CERTAIN POPULATION
C     OF SIZE N.  IT IS THE ADJACENCY MATRIX FOR THE
C     PARENT TO OFFSPRING DIRECTED GRAPH.
C
      DIMENSION IG(50,50)
      CHARACTER*20 GENEALFILE
      WRITE(*,*)'THIS PROGRAM IS FOR THE INPUT OF THE'
      WRITE(*,*)'ADJACENCY MATRIX FOR THE'
      WRITE(*,*)'PARENT-OFFSPRING DIRECTED GRAPH.'
      WRITE(*,*)' '
      WRITE(*,*)'IN WHAT FILE WILL THE GENEALOGY BE STORED?'
      READ(*,10)GENEALFILE
   10 FORMAT(A9)
      OPEN (UNIT=1, FILE = GENEALFILE, STATUS='OLD')
      WRITE(*,*)' '
      WRITE(*,*)'WHAT IS THE SIZE OF THE POPULATION?'
      READ(*,*)N
      WRITE(1,*)N
  200 WRITE(*,*)'WHICH EDGES DO YOU WANT TO INCLUDE?'
      WRITE(*,*)'INPUT "0,0" TO END'
      WRITE(*,*)'INPUT PARENT#, OFFSPRING#'
      READ(*,*)J,I
      IF(I.EQ.0) GO TO 300
      IG(I,J)=1
      GOTO 200
  300 DO 400 I=1,N
      DO 400 J=1,N
      WRITE(1,*)IG(I,J)
  400 CONTINUE
      END
```

GENOME.FOR

```
C
C   GENOME.FOR
C
C     THIS PROGRAM USES THE GENEALOGY IN FILE "GENEALOGY1.DAT" AND
C     THE ALLELE FREQUENCIES IN "FREQUENCY1.DAT" TO RANDOMLY GENERATE
C     GENOMES FOR THE ANCESTORS IN THE GENEALOGY AND TO BREED THEM
C     ACCORDING TO THE GENEALOGY AND PRODUCE GENOMES FOR THEIR
C     DECENDANTS. THE GENOMIC DATA IS THEN STORED IN A FILE.
C
C     THE PROGRAM ALSO ALLOWS A CERTAIN AMOUNT OF MUTATIONS.  SOME
C     PERCENTAGE OF OF THE TIME (MUT%) AN ALLELE AT A LOCUS IS
C     RANDOMLY REPLACED WITH ANOTHER ALLELE AT THAT LOCUS.
C
      DIMENSION IGEN(50,100),IF(50),IG(50,50),A(50,50)
      CHARACTER*20 GENEALFILE, FREQFILE, GENOMEFILE
C
C   IGEN( , ) IS THE GENOTYPE MATRIX WITH ENTRIES (INDIVIDUAL #, LOCUS1)
C   AND (INDIVIDUAL #, LOCUS2).
C
C   IF( ) HOLDS THE NUMBER OF ALLELES PER LOCUS, ALLELES ARE ASSUMED TO BE
C   EQUALLY FREQUENT, P=1/IF().
C
C   IG( ) IS THE MATRIX IN WHICH STORES THE GENEALOGY
C
C   A( , ) IS THE OBSERVED FREQUENCY OF THE ALLELES.
C
      WRITE(*,*)'WHAT FILE WILL BE USED FOR THE GENEALOGY?'
      READ(*,10)GENEALFILE
   10 FORMAT(A9)
      WRITE(*,*)'WHAT FILE FOR THE FREQUENCY DATA?'
      READ(*,10)FREQFILE
      WRITE(*,*)'WHAT FILE FOR THE GENOME OUTPUT?'
      READ(*,10)GENOMEFILE
      WRITE(*,*)'WHAT MUTATION RATE DO YOU WANT?'
      WRITE(*,*)'EXPRESS AS A PERCENTAGE PER ALLELE.'
      READ(*,*)MUT
      RMUT=FLOAT(MUT)/100.
      OPEN (UNIT=1, FILE=GENEALFILE, STATUS='OLD')
      OPEN (UNIT=2, FILE=FREQFILE, STATUS='OLD')
      OPEN (UNIT=3, FILE=GENOMEFILE, STATUS='OLD')
C
C     THE RANDOM NUMBERS ARE READ FROM THE FILE 'RAN.DAT'
C
      OPEN (UNIT=4, FILE='RAN.DAT', STATUS='OLD')
      READ (1,*)N
      DO 100 I=1,N
      DO 100 J=1,N
      READ (1,*)IG(I,J)
  100 CONTINUE
      READ (2,*)L
      DO 200 I=1,L
      READ (2,*)IF(I)
  200 CONTINUE
C
C    NOW, TO DETERMINE THE MAXIMUM IF()
C
      M1=IF(1)
      DO 300 I=2,L
      IF(M1 .LT. IF(I)) M1=IF(I)
  300 CONTINUE
```

```
C
C
C    THE COMPUTATION TO CHECK FOR PARENTS.   THOSE WITHOUT PARENTS ARE
C    RANDOMLY ASSIGNED A GENOME ACCORDING TO ALLELE FREQUENCIES.  THOSE
C    WITH PARENTS ARE GIVEN A GENOME DEPENDENT ON THE GENOME OF THEIR
C    PARENTS (ASSUMING MENDELIAN INHERITANCE).
C
      DO 500 IX=1,N
      IP1=0
      IP2=0
        DO 410 IY=1,N
        IF(IG(IX,IY) .EQ. 0) GO TO 410
        IF(IP1 .EQ. 0) IP1=IY
        IF(IP1 .EQ. 0) GO TO 410
        IP2=IY
  410 CONTINUE
      IF(IP1 .NE. 0) GO TO 430
C
C    RANDOM GENERARATION OF GENOTYPE FOR ANCESTORS (BASED ON ALLELE FREQS)
C
        DO 420 IZ=1,L
        CALL RANDOM(SEED,R)
        IGEN(IX,2*IZ-1)=IFIX(R*FLOAT(IF(IZ))+1.)
        CALL RANDOM(SEED,R)
        IGEN(IX,2*IZ)=IFIX(R*FLOAT(IF(IZ))+1.)
  420   CONTINUE
      GO TO 500
C
C    GENERATION OF GENOTYPE OF INDIVIDUAL IX WITH PARENTS IP1 AND IP2.
C
  430 DO 440 IZ=1,L
        CALL RANDOM(SEED,R)
        IGEN(IX,2*IZ-1)=IGEN(IP1,2*IZ-IFIX(2.*R))
        CALL RANDOM(SEED,R)
        IGEN(IX,2*IZ)=IGEN(IP2,2*IZ-IFIX(2.*R))
C
C    MUTATION OF ALLELE AT LOCUS 2*IZ-1
C
        CALL RANDOM(SEED,R)
        IF(R .GT. RMUT) GOTO 435
C
C       ASSIGN A (POSSIBLY NEW) ALLELE
C
        CALL RANDOM(SEED,R)
        IGEN(IX,2*IZ-1)=IFIX(R*FLOAT(IF(IZ))+1.)
C
C    MUTATION OF ALLELE AT LOCUS 2*IZ
C
  435 CALL RANDOM(SEED,R)
      IF(R .GT. RMUT) GOTO 440
C
C       ASSIGN A (POSSIBLY NEW) ALLELE
C
        CALL RANDOM(SEED,R)
        IGEN(IX,2*IZ)=IFIX(R*FLOAT(IF(IZ))+1.)
  440   CONTINUE
  500 CONTINUE
```

```
C
C    NOW TO CHOOSE THE MAXIMUM RL(I) AND MAKE THE DISTANCE MATRIX
C    M(IX,IY) BASED ON THIS
C
          D2=RL(1)
          M3=1
              DO 600 IZ=2,3
              IF(RL(IZ) .GT. D2) THEN
                 D2=RL(IZ)
                 M3=IZ
              ENDIF
  600         CONTINUE
       M(IX,IW)=M3
      IF(M(IX,IW) .EQ. 3) THEN
        M(IX,IW)=0
      ENDIF
      RL(1)=0.
      RL(2)=0.
      RL(3)=0.
  700 CONTINUE
C
C    NOW TO MAKE M(,) SYMMETRIC
C
       DO 800 IX=1,N
           DO 800 IY=1,IX-1
           M(IX,IY)=M(IY,IX)
  800 CONTINUE
C
C    NOW TO WRITE TO A FILE
C
       DO 900 IX=1,N
          DO 900 IY=1,N
          WRITE(2,*)M(IX,IY)
  900 CONTINUE
      END
```

MATRIX.FOR

```
C
C    MATRIX.FOR
C
C      THIS PROGRAM TAKES A DISTANCE MATRIX FROM RECONSTRUCT.FOR AND
C      BUILDS A DISTANCE MATRIX THAT IS COMPLETE.  THAT IS, IT ADDS
C      ANY NECESSARY MISSING VERTICES.
C
       DIMENSION M(50,50), IF(50), IHEAD(50)
       CHARACTER*20 DISTFILE, MATFILE, GENEALFILE, FREQFILE, GENOMEFILE
       WRITE(*,*)'WHAT IS THE DISTANCE FILE TO BE CONVERTED?'
       READ(*,10)DISTFILE
    10 FORMAT(A9)
       WRITE(*,*)'WHAT IS THE MATRIX FILE TO RECIEVE OUTPUT?'
       READ(*,10)MATFILE
       OPEN (UNIT=1, FILE=DISTFILE, STATUS='OLD')
       OPEN (UNIT=2, FILE=MATFILE, STATUS='OLD')
C
C    READ IN OF DISTANCE MATRIX FROM RECONSTRUCT
C
       READ(1,10)GENEALFILE
       READ(1,10)FREQFILE
       READ(1,10)GENOMEFILE
       READ(1,*)MUT
       OPEN (UNIT=3, FILE=GENEALFILE, STATUS='OLD')
       READ(3,*)N
       OPEN (UNIT=4, FILE=FREQFILE, STATUS='OLD')
       READ(4,*)L
       DO 30 I=1,L
       READ(4,*)IF(I)
    30 CONTINUE
       DO 100 IX=1,N
         DO 100 IY=1,N
         READ(1,*)M(IX,IY)
   100 CONTINUE
       WRITE(2,*)'THE GENEALOGY FILE WAS: ',GENEALFILE
       WRITE(2,*)'THE FREQUENCY FILE WAS: ',FREQFILE
       WRITE(2,*)'THE GENOME FILE WAS: ',GENOMEFILE
       WRITE(2,*)'  '
       WRITE(2,*)'THERE WERE ',N,' INDIVIDUALS IN THE POPULATION.'
   510 FORMAT(I2)
       WRITE(2,*)' '
       WRITE(2,*)'THIS RUN USED ',L,' LOCI WITH THE NUMBER OF ALLELES'
       WRITE(2,*)'AT EACH LOCUS AS DESRIBED BY THIS VECTOR:'
       WRITE(2,520) (IF(J), J=1,L)
   520 FORMAT(' ',50I3)
       WRITE(2,*)' '
       WRITE(2,*)'THE MUTATION RATE USED WAS ',MUT,'% PER ALLELE.'
       WRITE(2,*)'  '
       WRITE(2,*)'HERE ARE THE RESULTS:'
       DO 530 I=1,N+NEW
       IHEAD(I)=I-10*IFIX(FLOAT(I)/10.)
   530 CONTINUE
       WRITE(2,540)IO, (IHEAD(J), J=1,N+NEW)
   540 FORMAT(' ',I2,I3,49I2)
       DO 600 IX=1,N+NEW
       WRITE(2,550)IX, (M(IX,IY), IY=1,N+NEW)
   550 FORMAT(' ',I2,I3,49I2)
   600 CONTINUE
       END
```