# A Simulated Annealing Algorithm for Maximum Likelihood Pedigree Reconstruction

Anthony Almudevar*

July 19, 2000

**Abstract** The calculation of maximum likelihood pedigrees for related organisms using genotypic data is considered. The problem is formulated so that the domain of optimization is a permutation space. This is a feature shared by the travelling salesman problem, for which simulated annealing is known to be effective. Using this technique it is found that pedigrees can be reconstructed with minimal error using genotypic data of a quality currently realizable. This can be done without any *a priori* age or sex information. For smaller numbers of individuals a method of efficiently enumerating all admissible pedigrees of nonzero likelihood is given.

*Key words:* pedigree reconstruction, maximum likelihood, simulated annealing.

## 1 Introduction

The maximum likelihood approach to the reconstruction of pedigrees based on genotypic data was developed in a series of articles by Thompson (1974, 1975, 1976a) and also dis-

*Department of Mathematics and Computing Science, Saint Mary's University, Halifax, Nova Scotia,B3H 3C3

cussed in Cannings and Thompson (1981). This work was largely concerned with human pedigrees. In Meagher and Thompson (1986, 1987) pedigree reconstruction techniques are developed further in the context of natural populations. Pedigree reconstruction has also been explored based on DNA fingerprint data in Geyer *et. al.* (1993). The special case of sibling relationships within a single generation has been undertaken by Blouin (1996), Painter (1994, 1996) and Almudevar and Field (1999).

In the work cited above by Thompson, Cannings and Thompson, and Meagher and Thompson pedigrees are constructed by first enumerating potential parent/offspring triplets. In general, parent/offspring triplets suffice to define an entire pedigree (Cannings and Thompson (1981)), so that pedigree reconstruction can proceed from an inital enumeration of these triplets. In Thompson (1976a) a sequential procedure based on sibship membership is proposed which exploits age and sex information. The method is heuristic, undertaking a sequential consideration of parent/offspring triplets, and will not necessarily achieve the global maximum likelihood value. In a sequential method a single mistaken inference early in the procedure may preclude the determination of the true maximum, and hence the true pedigree. This problem is made more acute by the fact, noted in Thompson (1976a) and (1976b) that a sibling of an individual not genetically excluded as a parent will, on average, have a higher likelihood under the hypothesis of a parent/offspring relationship than an actual true parent.

The techniques proposed in this article can be considered as an extension of triplet based techniques, in that the first step is a triplet enumeration. The problem considered here is that of assembling the triplets into a feasible pedigree, without necessarily using age or sex data. A reformulation of the optimization problem, developed below, proves to be crucial. The effect of this reformulation is to divide the set of all admissible pedigrees into subsets on which the likelihood is easily maximized. These sets are in one-to-one correspondence with a permutation space, which then becomes the domain of optimization. This is a feature

shared by the travelling salesman problem, so that techiniques suitable for that problem can be adapted to this one.

A number of issues regarding the likelihood should be discussed. First, we distinguish between a *complete sample* and an *incomplete sample*. In the former, the parents of each individual are either in the sample, or are unrelated to all other members of the sample. In particular, this implies that founders are mutually unrelated. The likelihood given in Thompson (1976a), and the one used in this article, assumes a complete sample. Clearly, a complete solution to the problem requires treatment of likelihoods for incomplete samples. Such a situuation would arise, for example, when the grandparents but not the parents of an individual are contained in the sample. The inference should then include this set of relationships, which would not happen with a parent/offspring triplet approach.

The second issue is the use of age data. In Cannings and Thompson (1981) it is suggested that age data or some other type of ordering is necessary for the succesful reconstruction of a pedigree. If such data is not present then were the likelihood to be maximized simply by assigning the maximum likelihood parents to each individual independently it is probable that two individuals will be estimated to be parents of each other, resulting in an inadmissible pedigree. The simulated annealing algorithm proposed in this article is designed precisely to arrange individuals into their correct generations, so that maximium likelihood pedigrees can be constructed in the absence of age data. Any available age or sex data could still be incorporated in a straightforward manner.

In Section 2 the likelihood function is introduced, with a discussion of it's relevant properties. In Section 3 a tree based algorithm for the explicit enumeration of all admissible pedigrees of nonzero likelihood is given, suitable for smaller pedigrees. In Section 4 a mathematical representation of admissible pedigrees is developed, leading to an alternative formulation of the maximization problem. In Section 5 an implementation of a simulated annealing algorithm is introduced and applied to the maximization of the likelihood, following which

some numerical examples are given in Section 6. In Section 7 some resulting issues are discussed, and future directions in this area proposed.

## 2 The pedigree likelihood function

Suppose we have $N_L$ loci of genotype data for $N_I$ individuals labeled $1, \ldots, N_I$. A pedigree can be specified by first partitioning the individuals into three sets $F_0$, $F_1$ and $F_2$. An individual is in set $F_i$ if it has $i$ parents in the sample. According to the definition of a complete sample given in the previous section, any parent not included in the sample is assumed to be unrelated to all other individuals in the sample. Accordingly, the likelihood defined here is for a complete sample, and differs from the likelihood which would be appropriate for an incomplete sample.

To define the pedigree, for individuals $i \in F_1$ we further specify a single parent $a_i$. For individuals $i \in F_2$ we specify a parent pair $(a_i, b_i)$.

A pedigree can be represented as a directed graph. Each individual is represented by a single labeled node. A directed edge points from node $i$ to node $j$ if $i$ is a parent of $j$. A pedigree is feasible if it's graph contains no (directed) loops. This is equivalent to stating that no individual is an ancestor of itself. At this point we do not consider mating constraints due to age or sex. We denote the class of all feasible pedigrees $\mathcal{T}$.

Suppose $\mathcal{A}_l$ is the set of alleles in the population at locus $l$. Let $G_{li}$ be the unordered pair from $\mathcal{A}_l$ representing the genotype of individual $i$ at locus $l$. Let $\alpha_1(g_1|g_2)$ be the probability that an offspring has genotype $g_1$ given that one of it's parents has genotype $g_2$. Let $\alpha_2(g_1|g_2, g_3)$ be the probability that an offspring has genotype $g_1$ given that it's parents have genotypes $g_2, g_3$. Let $P(g)$ be the population frequency of genotype $g$. The likelihood function on $\mathcal{T}$ given $G_1, \ldots, G_N$ for pedigree $T$ is then

$$L(T) = L_0(T)L_1(T)L_2(T) \tag{2.1}$$

where

$$
\begin{aligned}
L_0(T) &= \prod_{i \in F_0} \prod_{l=1}^{N_L} P(G_{li}) \\
L_1(T) &= \prod_{i \in F_1} \prod_{l=1}^{N_L} \alpha_1(G_{li}|G_{la_i}) \\
L_2(T) &= \prod_{i \in F_2} \prod_{l=1}^{N_L} \alpha_2(G_{li}|G_{la_i}, G_{lb_i})
\end{aligned}
$$

The likelihood has the appearance of a function that is multiplicative across the individuals, and so can be maximized by selecting the maximum likelihood parents (or founder status) independently for each individual. However, the resulting pedigree need not be a valid pedigree. For example there would be nothing to prevent this procedure from simultaneously estimating $i$ to be the parent of $j$ and $j$ to be the parent of $i$. The likelihood confined to $\mathcal{T}$ is therefore not muliplicative in this sense. The structure of $\mathcal{T}$ plays a crucial role and this fact represents the central computational difficulty considered in this article.

# 3 Enumeration of feasible pedigrees

The space over which $L(T)$ is maximized is formally $2N_I$ dimensional, each dimension representing one putative parent for each offspring. There would therefore be on the order of ${N_I}^{2N_I}$ possible pedigrees to examine in an enumerative approach, which would be computationally unfeasible except in the smallest problems. Fortunately we may reduce the problem by introducing two constraints. First, that the pedigree be a member of $\mathcal{T}$, and second that the likelihood function be nonzero. The second constraint is satisfied when all parent/offspring combinations are genetically compatible.

The following procedure exploits these two constraints. For each individual a list is constructed consisting of all genetically compatible parental configurations. These include parent pairs, single parents and founder status. Suppose there are $m_i$ such configurations

for individual $i$, and let $D_{i,j} = (a_{i,j}, b_{i,j})$ be the $j$th parental configuration of individual $i$ for $j = 1, \ldots, m_i$. If this configuration identifies $i$ as a founder then set $(a_{i,j}, b_{i,j}) = (0, 0)$. Note that this founder configuration is included in each individual's list. If the configuration assigns only one parent in the sample to $i$ then set $a_{i,j}$ to equal that parent, and set $b_{i,j} = 0$.

The algorithm consists primarily of the construction of a labeled tree. Begin with a null root node at level 0. There follows $N_I$ levels, one corresponding to each individual in the sample in a fixed but arbitrary order. A tree node is labeled by one of the parental configurations corresponding to the individual of that level. The first level is constructed by assigning a node to each parental configuration of individual 1.

A node at level $i > 1$ is constructed according to the following rules. Suppose the $m$th node of level $i - 1$ is labeled with parental configuration $D_{i-1,j}$. Then an offspring node with parental configuration $D_{i,k}$ is created if and only if the pedigree implied by the parental configuration labels located on the path from the root node to the tentative new node implies a feasable pedigree. The tree can be constructed sequentially by level in this manner. To enumerate the feasable pedigrees, we enumerate all paths which extend from the root node to terminal nodes at level $i = N_I$.

Clearly, the smaller the parental configuration lists, the fewer the feasible pedigrees, which would be a consequence of larger numbers of loci, as well as of greater alllelic diversity within loci. The applicability of this technique will be limited to smaller pedigrees. An numerical example is given in Section 6.

# 4 An algebraic representation of pedigrees

For a given pedigree we will define a *pedigree matrix* as an $N_I \times N_I$ matrix in which element $(r, c)$ is a 1 if $c$ is a parent of $r$ and is 0 otherwise. Let $M_2$ be the set of all $0 - 1$ matrices with at most 2 non zero entries in each row. Clearly, a pedigree matrix must be in $M_2$, but a

member of $M_2$ need not be valid pedigree matrix corresponding to a pedigree in $\mathcal{T}$. Let $M_T$ be the class of $M_2$ matrices representing admissible pedigrees in $\mathcal{T}$. In order to characterize $M_T$ we require the following definitions.

**Definition 1** *An $n \times n$ matrix $A$ is* nilpotent (of order $k$) *if $A^k = 0$ but $A^{k'} \neq 0$ for $k' = 1, \ldots, k-1$.*

**Definition 2** *An $n \times n$ matrix $B$ is a* permutation matrix *if it contains exactly one 1 and $n-1$ 0's in every row and column.*

Thus, if $B$ is an $n \times n$ permutation matrix and $V$ is an $n$ dimensional vector $BV$ is equivalent to $V$ subjected to a permutation of it's elements. Also, if $A$ is an $n \times n$ matrix then $BAB^{-1}$ is equivalent to $A$ subjected to an identical permutation of it's rows and columns.

The following theorem is a well known result from linear algebra theory and is stated without proof.

**Theorem 1** *If $A$ is an $n \times n$ nilpotent matrix then there exists a permutation matrix $B$ and a strictly upper triangular matrix $U$ such that $A = BUB^{-1}$.*

We say that individual $i$ is an *order $k$ ancestor* of $j$ if $j$ inherits genes from $i$ through a path consisting of exactly $k$ inheritance events. Note that $i$ may be, for example, an order 1 ancestor and an order 2 ancestor of $j$, if $i$ mates with it's own offspring $i'$ to yield offspring $j$.

The following theorem is required for the principal result.

**Theorem 2** *If $A$ is an $n \times n$ pedigree matrix then the element $(r, c)$ of $A^k$ is greater than 0 if $c$ is an order $k$ ancestor of $r$ and is zero otherwise.*

*Proof.* We proceed by induction. Assume the theorem holds for some $k \geq 1$. Let $a_{ij}(m)$ be element $(i, j)$ of $A^m$. We have

$$a_{ij}(k+1) = \sum_{l=1}^{n} a_{il}(k) a_{lj}(1). \tag{4.2}$$

By (4.2) $a_{ij}(k+1) > 0$ if and only if there exists some $j'$ such that $a_{ij'}(k)a_{j'j}(1) > 0$. This is equivalent to stating that there exists individual $j'$ who is both an offspring of $j$ and an order $k$ ancestor of $i$, which is in turn equivalent to stating that $j$ is an order $k+1$ ancestor of $i$.

The proof is completed by noting that the theorem holds for $k = 1$ since $A$ is a pedigree matrix. □

*Remark.* Note that $a_{ij}(k)$ need not be zero or one. If two offspring of $j$ mate to produce $i$ then $a_{ij}(2) = 2$.

We can now state the principal result.

**Theorem 3** *An element of $M_2$ is a valid pedigree matrix if and only if it is nilpotent.*

*Proof.* Let $A \in M_2$. First assume that $A$ is nilpotent. By Theorem 1 there exists permutation matrix $B$ and upper triangular matrix $U$ such that $A = BUB^{-1}$. Suppose $B$ permutes the vector $(1, \ldots, n)$ to $(i_1, \ldots, i_n)$. This implies that for any $m = 1, \ldots, n$ any ancestor of $i_m$ is in the set $\{i_{m+1}, \ldots, i_n\}$ (empty for $m = N_I$), so that no ancestor of $i_m$ of any order can have $i_m$ as an ancestor. This constraint suffices to ensure that $A$ is a valid pedigree matrix.

Conversely, assume that $A$ is a valid predigree matrix. Suppose that the maximum number of generations of any line of descent in the pedigree is $k$. Then there is no order $k+1$ ancestor in the pedigree, hence $A^{k+1} = 0$, which implies that $A$ is nilpotent. □

Let $\mathcal{C}$ be the space of all permutations of $\{1, \ldots, N_I\}$. A permutation implies a ranking. A pedigree *conforms to ranking* $C \in \mathcal{C}$ if a parent always has higher rank than an offspring. Thus, Theorem 3 states that a pedigree is admissible if and only if it conforms to some ranking.

Clearly, with age playing the role of a ranking, this results seem reasonable, but for a mathematically rigorous treatment it is crucial to establish that there are no special cases or additional conditions required to establish the admissibility of a pedigree.

We can now reformulate the problem of maximizing $L$ on $\mathcal{T}$ in a way which will reduce the complexity of the problem. Let $L^*(C)$ for $C \in \mathcal{C}$ be the maximum of $L$ over all pedigrees $T \in \mathcal{T}$ which conform to $C$. Subject to $C$, $L$ can be maximized by separately selecting as parents for each individual the maximium likelihood parents (or founder status) among those of higher rank, in this way exploiting the multiplicative form of $L$. By Theorem 3 any pedigree constructed in this way will be admissible, and every admissible pedigree conforms to some $C$. Therefore, the problem of maximizing $L$ on $\mathcal{T}$ is logically equivalent to maximizing $L^*$ on $\mathcal{C}$.

# 5 A simulated annealing algorithm

When the number of feasible pedigrees is too large for exhaustive enumeration, an alternative is the use of simulated annealing, an optimization tool which has proven effective in a large variety of combinatorial optimization problems (Kirkpatrick *et. al.* (1983)).

Suppose we wish to maximize a function $f$ on a state space $X$. We construct a Monte Carlo Markov chain (MCMC) on $X$ as follows. For each $x \in X$ there is a *neighbourhood* of states, assumed to be a constant size $N$. A transition from state $x_i$ to $x_{i+1}$ is defined by first selecting at random a *proposal state* $x'_{i+1}$ from the neighbourhood of $x_i$. Then $x'_{i+1}$ is accepted as the subsequent state with probability

$$\alpha_c(x'_{i+1}; x_i) = \begin{cases} 1 & \text{if } f(x'_{i+1}) \geq f(x_i) \\ \exp\left(\frac{f(x'_{i+1}) - f(x_i)}{c}\right) & \text{if } f(x'_{i+1}) < f(x_i) \end{cases} \tag{5.3}$$

where the constant $c$ is refered to as the *temperature* and, in a simulated annealing algorithm, is allowed to decrease to zero. If the proposal state is accepted we set $x_{i+1} = x'_{i+1}$, otherwise

$x_{i+1} = x_i$. This acceptance rule is known as the *Metropolis criterion* (Metropolis *et. al.* (1953)).

Our objective is to maximize $f = L^*$ on the permutation space $X = \mathcal{C}$. Optimization on a permutation space is a feature shared with the well known *travelling salesman problem*, hence it would seem appropriate to adopt methodology known to function well for that problem. Accordingly, the simulated annealing algorithm adopted here is one proposed in Aarts and van Laarhoven (1985a, 1985b), and summarized in Aarts and Korst (1989), and is one with established effectiveness for that problem. We give below a sketch of the argument. Readers may consult the references for further details.

The initial temperature $c_0$ is determine by specifying the initial acceptance probability $\chi_0$. Ideally, this would be 1, but a considerable savings in computation time with little or no decrease in accuracy can be attained by selecting $\chi_0$ to be somewhat smaller than 1, say 0.95. The relationship between temperature $c$ and acceptance probability $\chi$ is

$$c = \frac{d^{(-)}}{\ln\left(\frac{m_2}{m_2\chi - m_1(1-\chi)}\right)} \tag{5.4}$$

where given a long sequence of $m_0 = m_1 + m_2$ proposed transitions $m_1$ and $m_2$ are number for which $f(x'_{i+1}) \geq f(x_i)$ and $f(x'_{i+1}) < f(x_i)$ respectively, with $d^{(-)}$ defined as the average magnitude of change in $f$ among the proposed transitions resulting in a decrease in $f$. Fix $\chi = \chi_0$ in (5.4), and set $c_0 = 0$. Generate $m_0$ trials, updating $m_1, m_2, d^{(-)}$ according their definition while updating $c_0$ using (5.4). The final value of $c_0$ is taken to be the initial temperature. Below we use $m_0 = 100$.

The *cooling schedule* determines how temperature $c$ is decreased. The schedule consists of a sequence of *temperature stages*. In stage $i$ the process remains at constant temperature $c_i$ for a fixed number of transitions $\beta N$, then proceeds to temperature $c_{i+1}$, using a fixed *neighbourhood size factor* $\beta$. Note that each temperature results in a distinct steady state distribution of the process $\{f(x_i) : i \geq 1\}$. The change in temperature $\Delta c$ is calibrated to result in a change in variational distance between steady state distributions proportional to

a fixed constant $\delta$. If the observed variance of the process is $\sigma_{c_i}$ at temperature $c_i$, then temperature $c_{i+1}$ can be given by formula

$$c_{i+1} = \frac{c_i}{1 + \frac{c_i \ln(1+\delta)}{3\sigma_{c_i}}}. \tag{5.5}$$

For each temperature $c_i$ at stage $i$, the stage process mean $\mu_{c_i}$ can be estimated. Convergence of the process can be inferred when the change in process mean is negligible. In this article convergence is inferred when the condition

$$0 \leq \frac{\mu_{c_{i+1}} - \mu_{c_i}}{\sigma_{c_0}} \leq \epsilon \tag{5.6}$$

is satisfied for $N_\epsilon$ consecutive temperature changes. At this point, the current state is assumed to be a global maximum.

In Aarts and Korst (1989) the neighbourhood is defined by considering each pairwise exhange of elements of $C$ (a *swap*). However, it was found that an alternative proposal performed better, in which a single element of $C$ is chosen at random, then placed in a new position chosen at random (a *step*). This is discussed further in the following section.

To summarize, the process depends on 5 control parameters:

$\chi_0$ = initial acceptance rate.

$\delta$ = increment in steady state distribution variational distance.

$\epsilon$ = convergence tolerance.

$\beta$ = neighbourhood size factor.

$N_\epsilon$ = convergence event number.

The advantage of this approach is that a single set of control parameters can be expected to result in a similar performance for an entire class of input problems.

Finally, it should be noted that this implementation assumes that a postulate holds in which the values of the objective function $f$ are are divided into two groups $R_1$ and $R_2$. Group

$R_1$ consists of those values within a few standard deviations of the mean, and $R_2$ consists of those values outside of $R_1$ close to the maximum value. Values in $R_1$ are normally distributed, and values in $R_2$ are exponentially distributed. Furthermore, the number of values in $R_1$ is much larger than the number of values in $R_2$.

# 6 Examples

We consider here 3 examples. Genotypes are simulated for test pedigrees 1 and 2 given in Figures 1 and 2. Ten loci were assumed, each with 8 equally frequent alleles, for a heterozygosity of 0.875 per locus. This is comparable in allelic diversity to a set of DNA collected from a sample of 17 sperm whales analyzed in Example 3. No sex distinction is made in test pedigrees 1 and 2.

## 6.1 Example 1

In this example 1000 simulation trials were conducted for test pedigree 1. The exact maximum likelihood pedigree was determined. Then, using the same set of simulated DNA the simulated annealing algorithm was repeated 32 times under various control parameter settings, obtained by varying $\chi_0 = 0.90, 0.95$; $\delta = 0.1, 0.01$; $\epsilon = 10^{-2}, 10^{-5}$; $\beta = 0.5, 1.5$ and $N_\epsilon = 1, 3$. The calculation time to convergence in milliseconds was also captured, and a check was performed to see of the pedigree to which the algorithm converged was the true maximum likelihood pedigree (not necessarily the true pedigree). The experiment was conducted with a standard personal computer. As discussed above, we have two alternative proposal rules. Therefore this experiment was done using the swap proposal, then for the step proposal. It was found that the step proposal worked uniformly better, so the following discussion will deal with the step proposal only.

In Figure 3, configurations are represented by a binary string in which the digits 0 and 1

represent, in order and respectively, $\chi_0 = 0.90, 0.95$; $\delta = 0.1, 0.01$; $\epsilon = 10^{-2}, 10^{-5}$; $\beta = 0.5, 1.5$ and $N_\epsilon = 1, 3$. In general, a setting marked as 1 should result in greater accuracy but longer convergence time. This rule appears to hold generally, but not absolutely. As expected, configuration 11111 gives nearly the highest accuracy and highest convergence time. There is a cluster of configurations in the low convergence time/high convergence accuracy quadrant. A closeup of this region is given in Figure 4, in which configuration 00011 appears to be the configuration giving optimal or near optimal accuracy with low convergence time. To summarize, this corresponds to control parameters $\chi_0 = 0.90$, $\delta = 0.1$, $\epsilon = 10^{-2}$, $\beta = 1.5$ and $N_\epsilon = 3$. This configuration will be adopted for all trials which follow.

Using the exact algorithm, the correct pedigree was uniquely estimated in 943 of 1000 trials. The adopted configuration converged to the true maximum likelihood pedigree in 993 of 1000 trials.

To more closely scrutinize the algorithm and likelihood function a single trial was examined in detail. This trial resulted in the a correct likelihood estimate and in correct convergence of the simulated annealing algorithm. There were 2184 transitions and 26 temperature stages. In Figure 5 a plot of the log likelihood against the number of transitions is given, which suggests a reasonably smooth transition of the steady state distribution, and a plausible convergence event. In Figure 6 summaries of mean log likelihood, temperature, variance and proposal acceptance rate are given for each temperature stage. Because the maximum likelihood pedigree will probably conform to more that one ranking, the global maximum will occur at more than one state, hence we do not expect the acceptance rate to converge to zero. Again, the summaries suggest acceptable convergence behaviour. A histogram of all log likelihood values is given in Figure 7, showing no obvious deviation from the distributional postulate given in the previous section. There were 30 unique likelihood values observed.

In addition, each of the 8!=40,320 permutation states was examined separately. A dis-

tance between two permutations was taken to be the minimum number of elements which need to be removed in order to leave the remaining partial permutations equal. This gives the minimum number of proposals needed to reach one permutation state from another. (This can be efficiently calculated using an appropriate recursive algorithm, adapted from one given in Almudevar and Field (1999)). For each permutation state the constrained maximum likelihood pedigree was calculated and the corresponding log likelihood recorded. In addition, the minimum permutation distance to a state which achieves the global maximum likelihood was calculated. The term *gradient* is here used to indicate a statistical tendency for states nearer a maximum to have higher likelihoods. In Figure 8 a boxplot of the log likelihoods grouped by minimum permutation distance to the global maximum is given. It indicates the existence of a clear gradient leading to the maximum likelihood ranking, which is crucial to the succesful convergence of the algorithm. We also note that there appears to be a smaller local maximum located among the permutation states which are a distance of 4 from a global maximum. The maximum likelihood pedigree among this group was the one in which individuals 1 and 2 have parent 5 only, individuals 3 and 4 have parent 6 only, individuals 7 and 8 have parents 5 and 6, and individuals 5 and 6 are founders. The second boxplot gives the log likelhoods for this group, in turn grouped by the number of times individual 5 or 6 is inferred as a parent. We clearly see a second gradient, leading to a pedigree in which individuals 5 and 6 dominate as parents. Thus, in addition to fluctuation of likelihood between transitions, we can expect larger scale multimodality of the likelihood surface. This feature was observed consistently by the author over a number of trials.

## 6.2 Example 2

A second simulation experiment was conducted using the pedigree in Figure 2. This pedigree consists of 59 individuals, and contains examples of inbreeding, as well as of half sibling relationships. A total of 200 trials were conducted. The mean convergence time was observed to

be 18.4 minutes, with minimum, quartiles and maximum of the convergence times observed to be 7.9, 11.5, 12.8, 17.3, 93.2 minutes, suggesting a clear right skewness. Error was determined by noting that a pedigree of 59 individuals is defined by the 118 parent specifications (with the population defined as a parent for founders). The number of these parents incorrectly inferred is taken as the error. The mean observed error was 1.9 parent specifications out of 118. The error frequencies are given in detail in Table 1. A more detailed breakdown of the errors is given in Table 2. Here, errors are reported as the average number of parent mispecifications per indivudual. The average individual rate can be converted to the average rate for the pedigree by multiplying by 59. Note that there appears to be differing error rates for different classes of individuals. In particular, the error rate for founders appears to be lower than that for other types of pedigree members from all other generations. The error rate for the inbred members seems higher than that of all other member, within and outside of their generation.

An individual trial was examined separately. The number of transitions until convergence was 970,137. A plot of the log likelihood against number of transitions is given in Figure 9, which indicates a smooth transition of the steady state distribution and a plausible convergence event. The temperature stage summaries are given in Figure 10, again indicating the desired behaviour. A histogram of 5,000 randomly chosen log likelhood values given in Figure 11 suggests that the distributional postulate holds.

### 6.3 Example 3

As a further example DNA markers from 17 sperm whale from a Galapagos Islands breeding ground were subjected to the simulated annealing algorithm. There were 10 loci with heterozygosities of 0.75, 0.68, 0.87, 0.89, 0.75, 0.89, 0.91, 0.83, 0.78 and 0.62, making the information content comparable to that used in the previous examples. Population genotype frequencies were estimated from an independent sample from the same population.

Only two parent/offspring pairs exist which are not genetically excluded. Age data exists which establishes generation. The algorithm infered these two relationships in the correct generational order. The convergence time was several seconds. It is notable that the same control parameters were used in this example as were used in Examples 1 and 2, suggesting that a single set of well chosen values can be expected to result in a well behaved algorithm applicable to a wide variety of input problems.

# 7 Discussion

A simulated annealing algorithm adapted to maximize the likelihood function of a joint pedigree was presented. The algorithm appears to converge properly, and has demonstrated success in infering pedigrees without using age or sex data. In addition, a single set of well chosen control parameters can be used for problems of differing complexity.

A number of issues remain. It was assumed that founders are unrelated. The statistical properties of the genotype distribution on the pedigree has been shown to be noticeably dependent on the relatedness of the founders in a study of a Guam rail founder pedigree in Haig *et. al.* (1994). This issue assumes importance when the requirement of a complete sample is relaxed. The author is planning future work in this direction.

The algorithm assumes knowledge of population genotype frequencies. Using the sample itself to estimate these frequencies is problematic, since all information about the frequencies is contained in the founders, which may be small in number. However, in Thompson (1976a), it is argued that kinship estimation is insensitive to errors in allele frquencies, except for especially rare alleles. A consequence is that adequate allele frequency estimates can be obtained by assuming unrelatedness among the sampled individuals. The allele frequency estimates can be updated once the pedigree has been estimated. Also, population genotype frequencies have been replaced with Bayesian prior distributions in Painter (1994, 1996).

Age and sex data may be incorprated into the algorithm in a straighforward way, by retaining only those parent/offspring triplets which conform to whatever age or sex information is available. The algorithm is therefore especially useful in a situation in which age and sex data exists for some but not all individuals.

The algorithm first proceeds by enumerating all potential parent/offspring triplets, then assembling the triplets into a pedigree. It was shown in Meagher and Thompson (1986) that an acceptable compromise is to first enumerate for each individual all high likelihood single parents, then to construct the triplets from these, which results in a significant reduction of computing time.

Linkage presents another problem. The likelihood used here assumes that loci are unlinked. If linkage was to be incorporated into the likelihood, recombination fractions, known or estimated, would need to be incorporated into the calculation of joint genotype probabilities. Otherwise, the structure of the problem remains unchanged.

Microsatellite DNA markers are especially appropriate for kinship and mating behaviour inference They tend to be highly variable, with heterozygosities often above 0.8 ($k$ equally frequent alleles result in a heterozygosity of $(k-1)/k$ under random mating). Especially important is the fact that they tend to be generally neutral with respect to selection, which is a crucial assumption for pedigree inference. Some exceptions have been observed. Also, mutation rates for highly variable markers higher than $10^{-2}$ have been reported, so a small number of mutations may be expected in a reasonably large pedigree. This suggests that modifications which render the algorithm robust to mutations and other anomolies must be part of a genuinely complete solution to the problem of pedigree reconstruction. For a discussion of the role of microsatellites in kinship inference, see Queller *et. al.* (1989) and Bruford and Wayne (1993).

It is suggested in Thompson (1976a) that the definition of the likelihood could be expanded to account for family size, mating behaviour, or other geographic and demographic

variates. It is anticipated that the methodology proposed here would be adaptable to the more complicated likelihood models which would result.

The DNA evidence is assumed to be genotypic, and the likelihood is therefore calculated using the probability laws of Mendelian inheritance. On the other hand, DNA fingerprints (observable patterns of bands created by fragments of DNA obtained by cutting the DNA at identifiable sites, cf. Bailey (1995)) cannot be so modeled, since they are not assignable to specific locus sites. It would be, in principle, possible to modify the techniques proposed in this article to use finger print data, but significant drawbacks have been reported with their use in kinship estimation (Lynch (1988)).

# 8 Acknowledgments

This research is supported by a grant from the Natural Sciences and Engineering Research Council of Canada. In addition, the author wishes to acknowledge the generous support of Chris Field. The author is also grateful to Heydar Radjavi for some invaluable discussions. The sperm whale data is courtesy of Jenny Christal.

# 9 Bibliography

Aarts, E. and van Laarhoven, P.J.M. (1985a). Statistical Cooling: A General Approach to Combinatorial Optimization Problems. *Philips Journal of Research* **40**, 193-226.

Aarts, E. and van Laarhoven, P.J.M. (1985b). A New Polynomial-time Cooling Schedule. *Proc. IEEE Int. Conference on Computer-Aided Design*, Santa Clara, Nov. 1985, 206-208.

Aarts, E. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines.* John Wiley & Sons

Almudevar, A. and Field, C. (1997). Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 2, 212-229.

Blouin, M.S., Parsons, M., Lacaille, V. and Lotz, S. (1996). Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology* **5**, 393-401.

Brookfield, J.F.Y. and Parkin, D.T. (1993). Use of single-locus DNA probes in the establishment of relatedness in wild populations. *Heredity* **70**, 660-663.

Bruford, M.W. and Wayne, R.K. (1993). Microsatellites and their application to population genetic studies. *Current Opinion in Genetics and Development* **3**, 939-943.

Cannings, C. and Thompson, E. (1981). *Genealogical and genetic structure.* Cambridge University Press.

Geyer, C. J., Ryder, O. A., Chemnick, L. G. and Thompson, E. A. (1993). Analysis of relatedness in the California condors: from DNA fingerprints. Molecular Biology and Evolution, 10 571-589.

Haig, S.M., Ballou, J.D. and Casna, N.J. (1994). Identification of kni structure among Guam rail founders: a comparison of pedigrees and DNA profiles. *Molecular Ecology* **3**, 109-119.

Kirkpatrick, S., Gelatt Jr., C.D., and Vecchi, M.P. (1983). Optimization by Simulated Annealing, *Science* **220**, 671-680.

Lynch, M. (1988). Estimation of Relatedness by DNA fingerprinting. *Mol. Biol. Evol.* **5**, 584-599.

Meagher, R.M. and Thompson, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theor. Pop. Biol.* **29**, 87-106.

Meagher, R.M. and Thompson, E. (1987). Analysis of parentage for naturally established seedlings of *chamaelirium luteum* (liliacae). *Ecology* **68**, 803-812.

Metroplis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-92.

Painter, I. (1994) Analysis of full-sibship configurations in a maternal half-sibship. Technical Report No. 272, Dept. of Statistics, University of Washington, Seattle, Washington.

Painter, I. (1996). Sibling reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 2, 212-229.

Queller, D.C. and Goodnight, K.F. (1989). Estimating relatedness using genetic markers. *Evolution* **43**, 2, 258-279.

Thompson, E. (1974). Gene identities and multiple relationships. *Biometrics* **30**, 667-

680.

Thompson, E. (1975). Estimation of pairwise relationships. *Annals of Human Genetics* **39**, 173-188.

Thompson, E. (1976a). Inference of genealogical structure. *Soc. Sci. Inform.* **15**, 2/3, 477-526.

Thompson, E. (1976b). A paradox of genealogical inference. *Adv. Appl. Prob.* **8**, 2/3, 648-650.

Table 1: Summary of total errors for Experiment 2

| Error out of 118 | Frequency out of 200 |
|---|---|
| 0 | 41 |
| 1 | 59 |
| 2 | 40 |
| 3 | 30 |
| 4 | 17 |
| 5 | 7 |
| 6 | 4 |
| ⋮ | ⋮ |
| 8 | 1 |
| ⋮ | ⋮ |
| 16 | 1 |

Table 2: Summary of individual error rates for Experiment 2 by generation and type

| Generation | Founder | Inbred | Halfsib | No Offspring | Total |
|---|---|---|---|---|---|
| 1 | 0.014 | | | | 0.014 |
| 2 | | | 0.028 | 0.029 | 0.027 |
| 3 | 0.006 | | | 0.050 | 0.034 |
| 4 | | | | 0.029 | 0.031 |
| 5 | | 0.070 | 0.034 | 0.052 | 0.052 |
| Total | 0.012 | 0.070 | 0.031 | 0.039 | 0.032 |

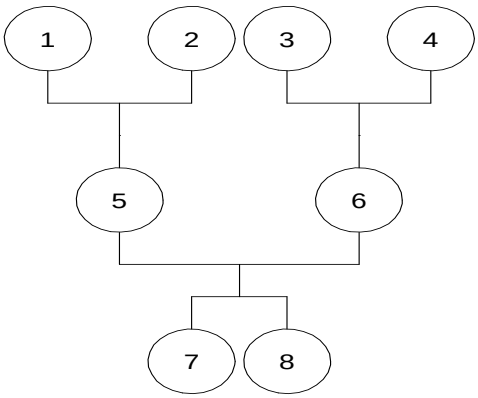


Figure 1: Test pedigree 1.

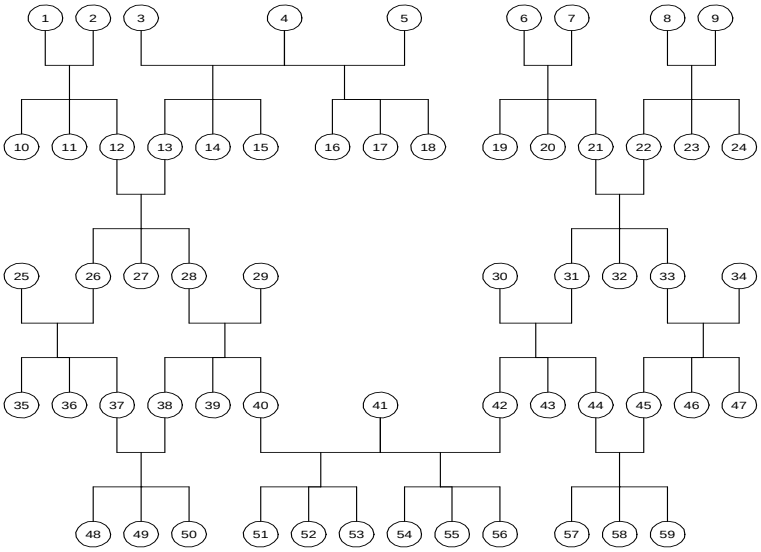


Figure 2: Test pedigree 2.

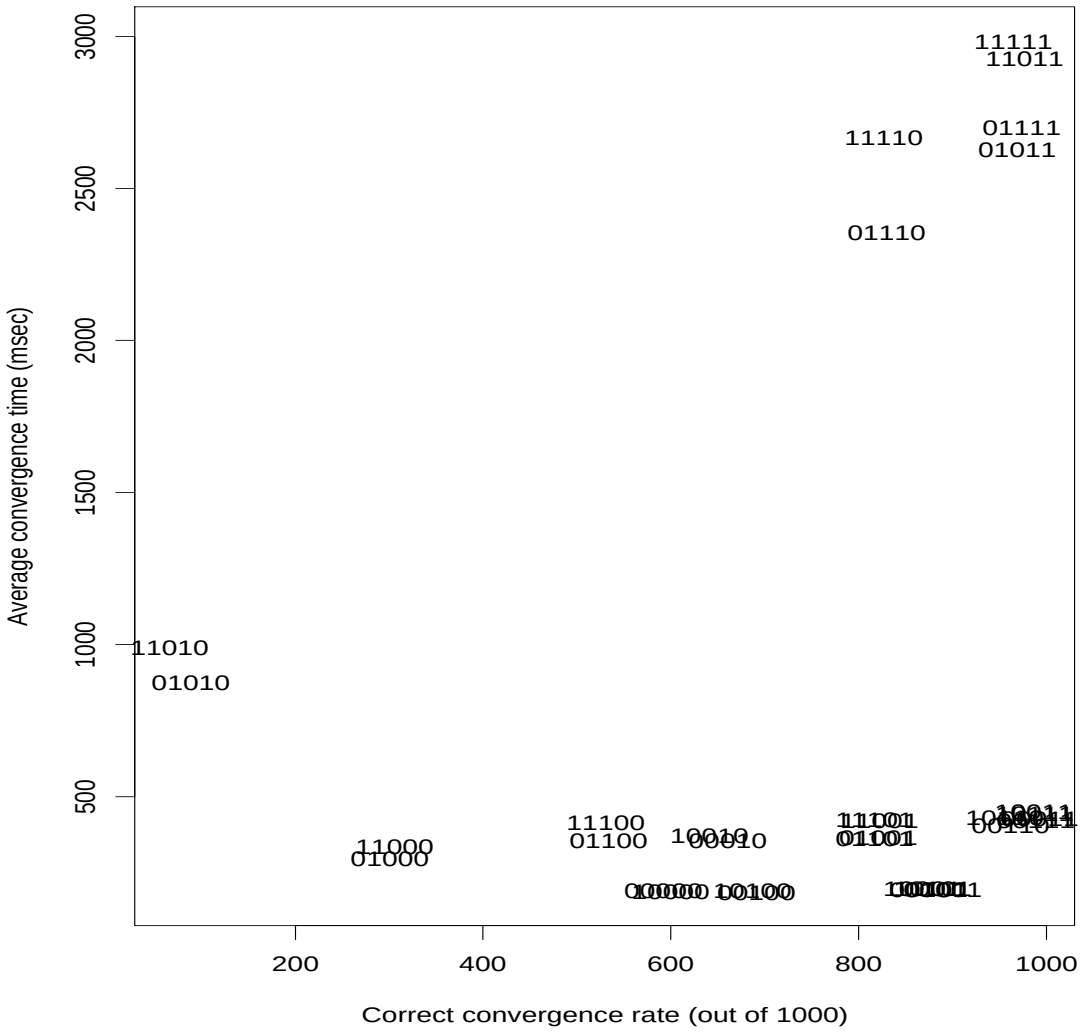


Figure 3: Plot of convergence time against correct convergence rate for experiment using step proposals for 32 configurations of control parameters. The configuration is represented by a binary string in which the digits 0 and 1 represent, in order and respectively, $\chi_0 = 0.90, 0.95$, $\delta = 0.1, 0.01$, $\epsilon = 10^{-2}, 10^{-5}$, $\beta = 0.5, 1.5$ and $N_\epsilon = 1, 3$.

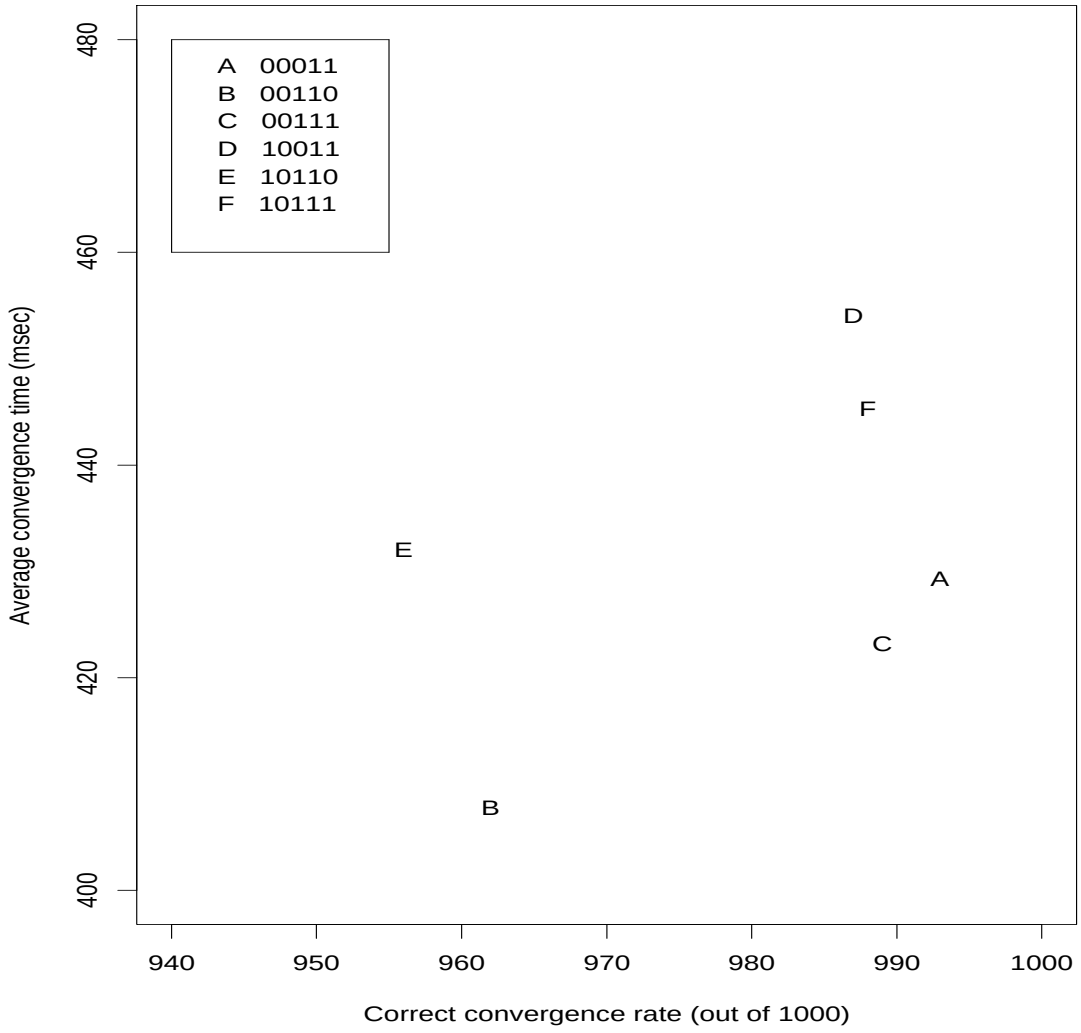


Figure 4: Closeup of low convergence time/high correct convergence region of Figure 1. The configuration is represented by a binary string in which the digits 0 and 1 represent, in order and respectively, $\chi_0 = 0.90, 0.95$, $\delta = 0.1, 0.01$, $\epsilon = 10^{-2}, 10^{-5}$, $\beta = 0.5, 1.5$ and $N_\epsilon = 1, 3$.

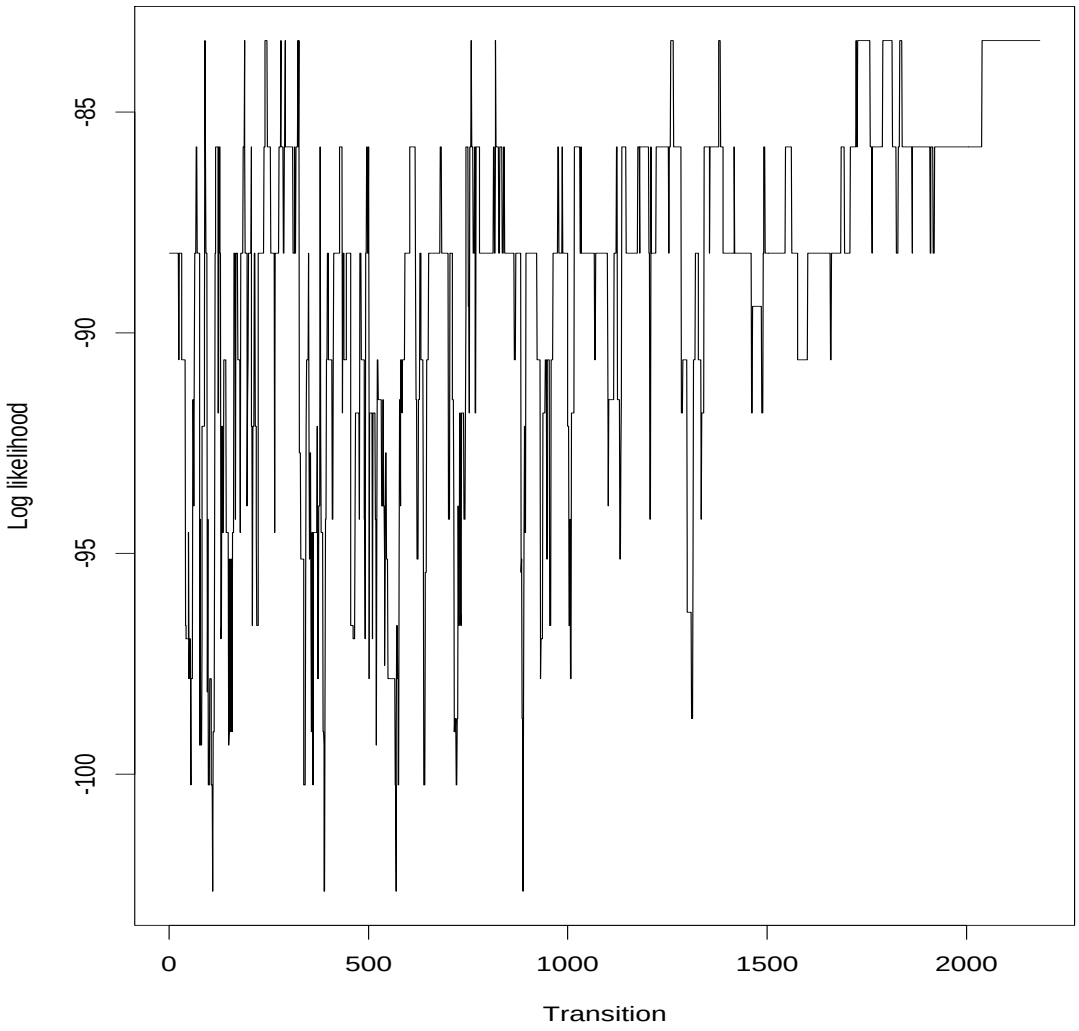


Figure 5: Trace of log likelihood by transition for trial of test pedigree 1.

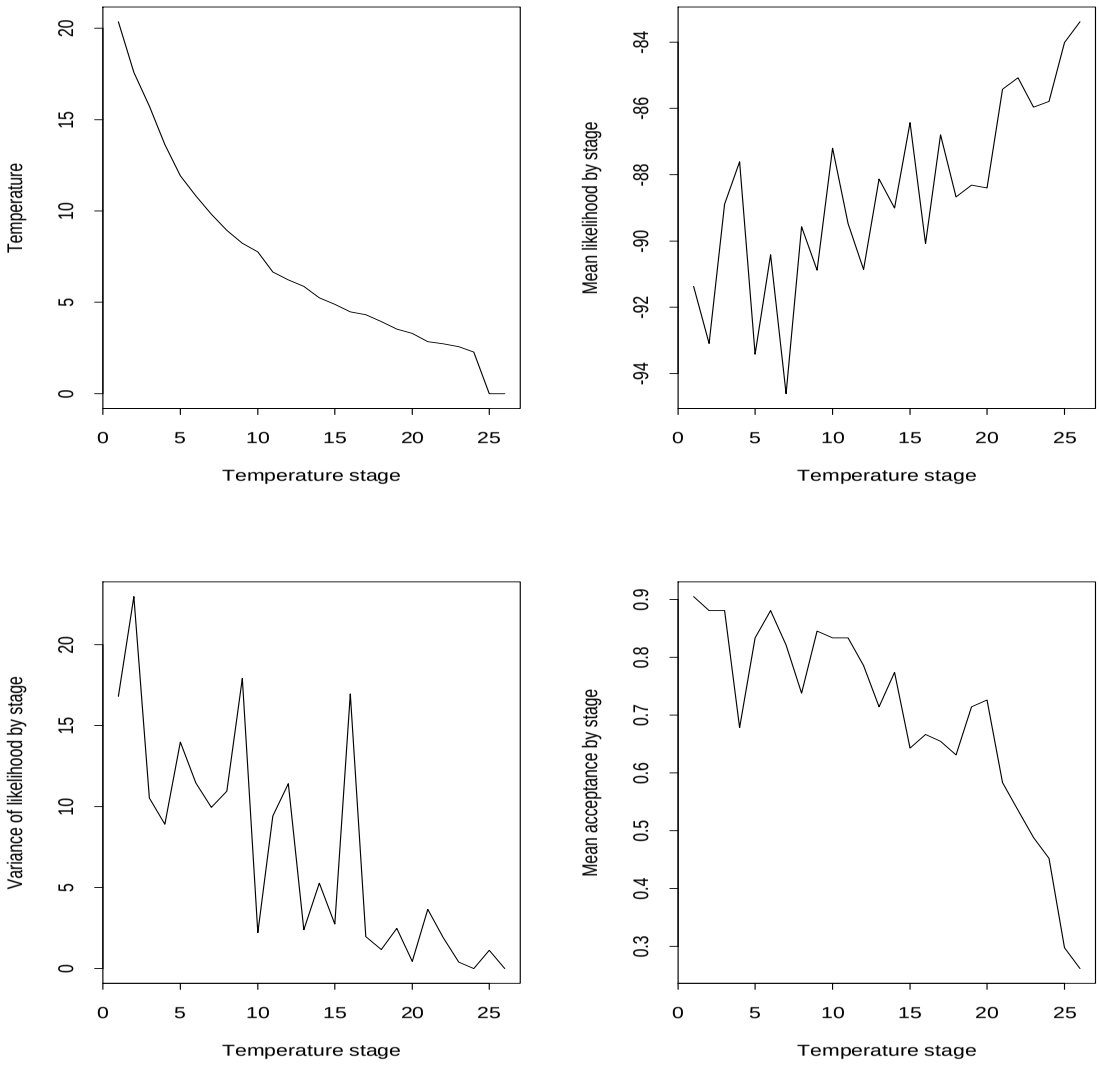


Figure 6: Diagnostic plots for trial of test pedigree 1.

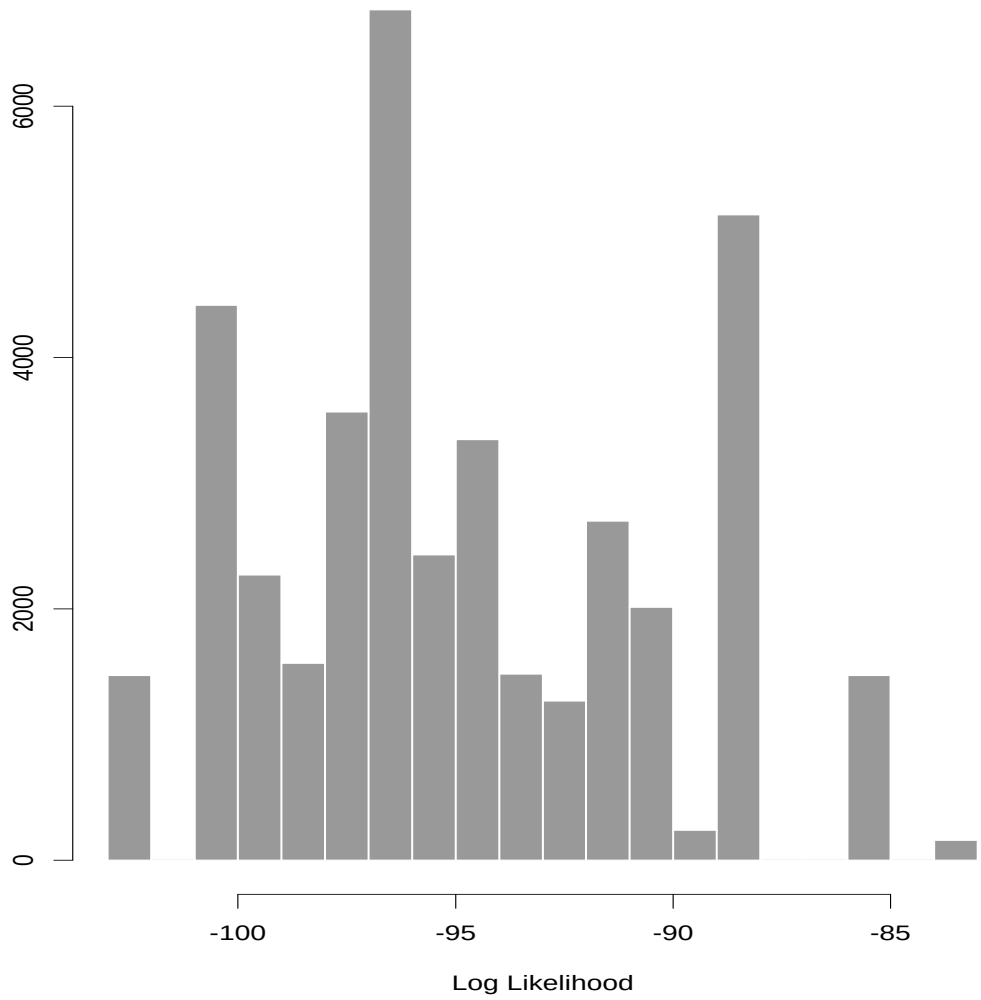


Figure 7: Histogram of log likelihood for trial of test pedigree 1.

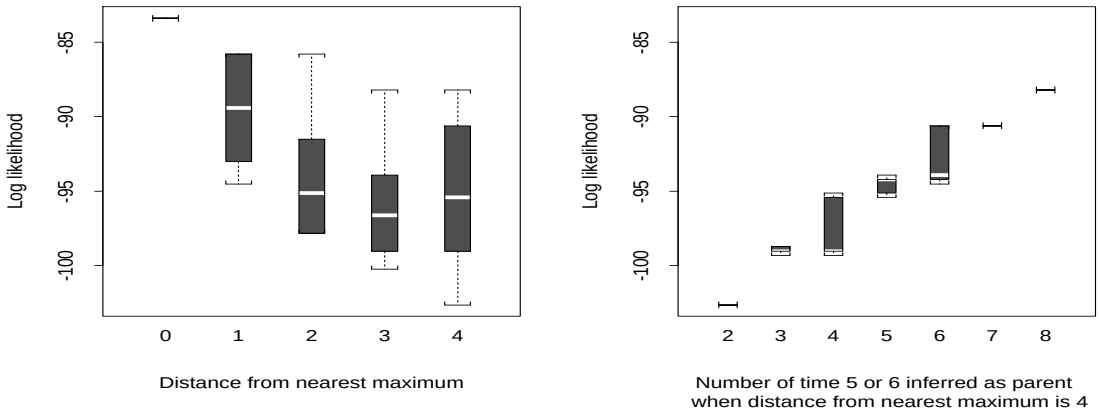


Figure 8: First figure is boxplot of log likelihood for trial of test pedigree 1 grouped by permutation distance from nearest global maximum state. The second figure shows boxplots of log likelihood among states a distance of 4 steps from a global maximum grouped by the number of times individual 5 or 6 appears as a parent in the corresponding constrained maximum likelihood pedigree.

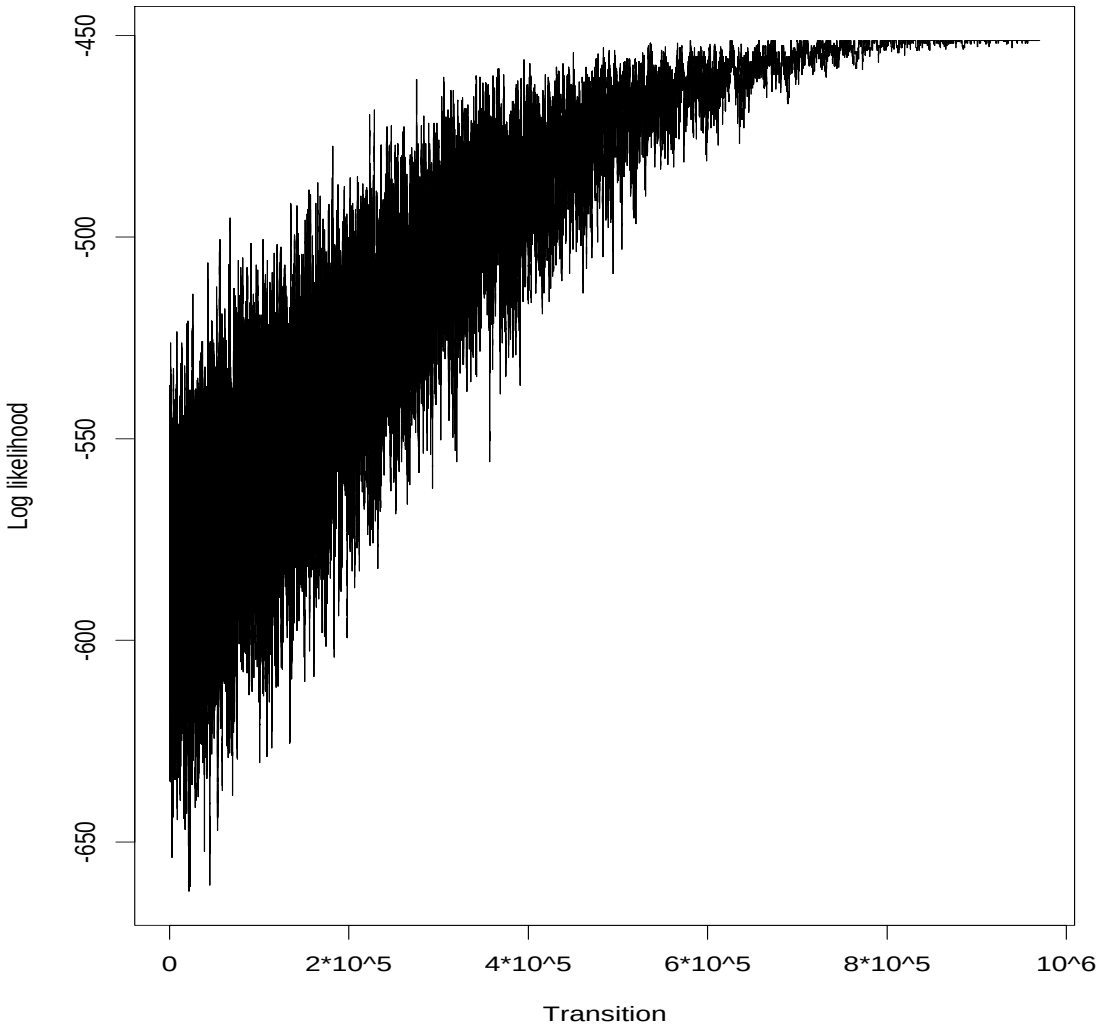


Figure 9: Trace of log likelihood by transition for trial of test pedigree 2.

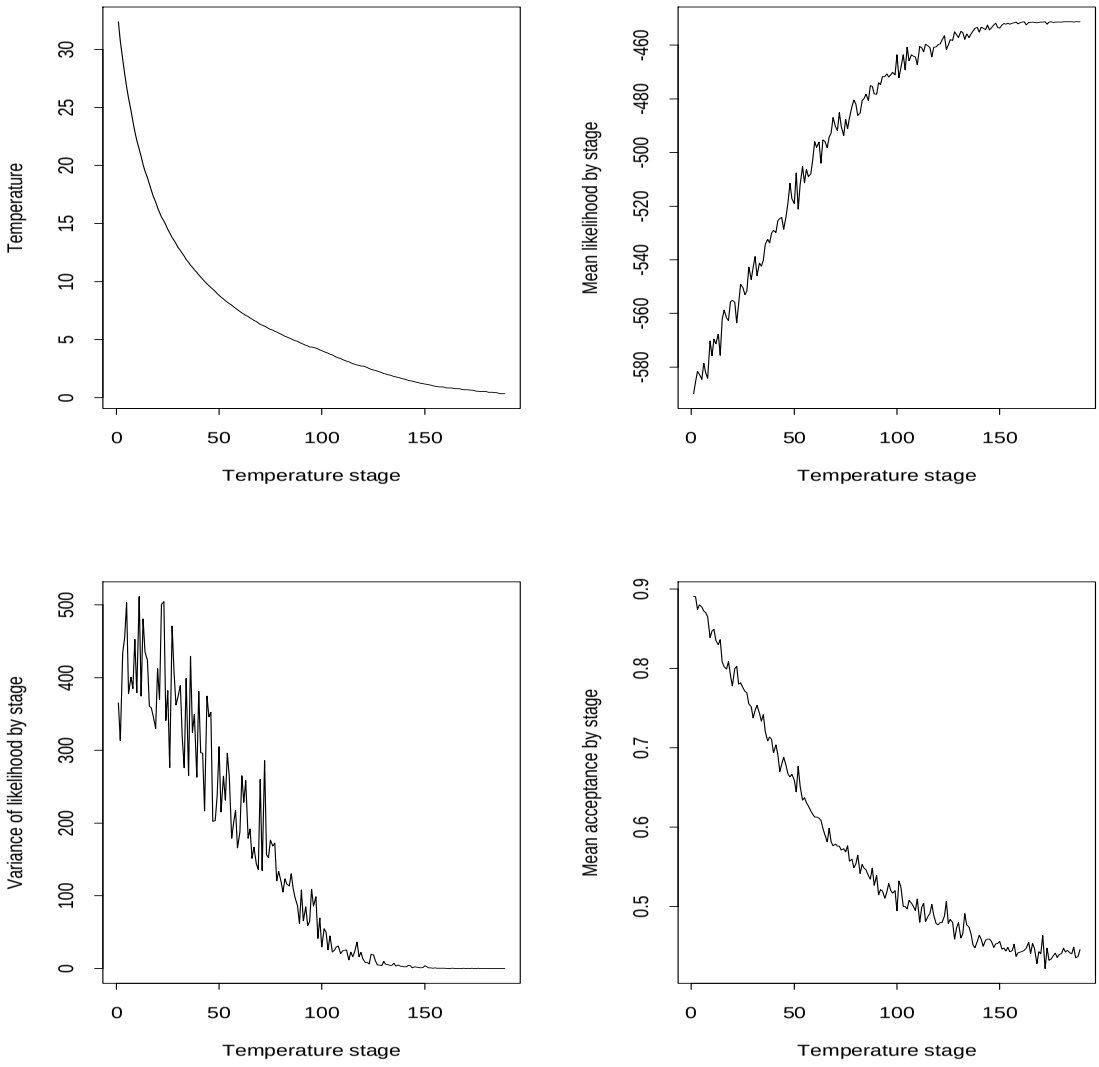


Figure 10: Diagnostic plots for trial of test pedigree 2.

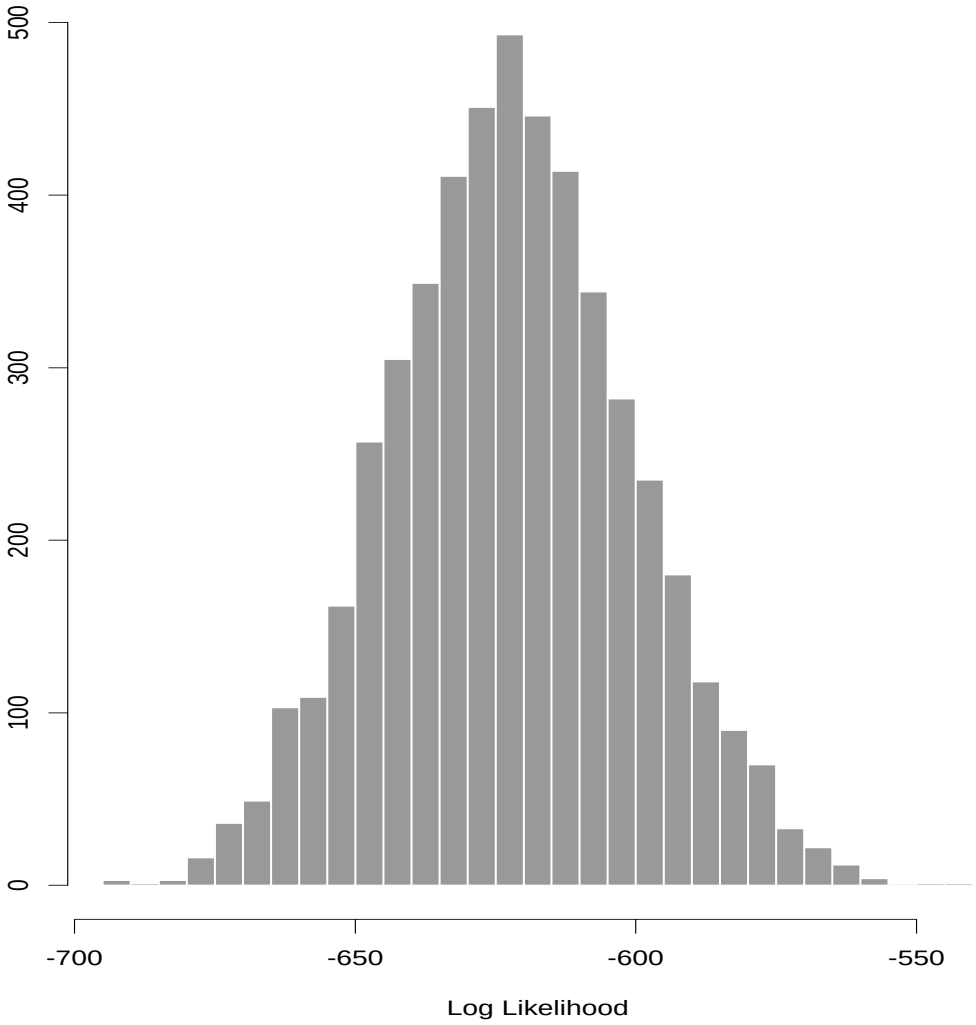


Figure 11: Histogram of log likelihood for trial of test pedigree 2.