Published in final edited form as:

Theor Popul Biol. 2007 March ; 71(2): 213–229.

A Graphical Approach to Relatedness Inference

Anthony Almudevar

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, 14642, USA, Tel 585-275-6992, Fax 585-273-1031, anthony almudevar@urmc.rochester.edu

Abstract

The estimation of relatedness structure in natural populations using molecular marker data has become an important tool in population biology, resulting in a variety of estimation procedures for specific sampling scenarios. In this article a general approach is proposed, in which the detailed relationship structure, typically a pedigree graph or partition, is considered to be the object of inference. This makes available tools used in complex model selection theory which have demonstrated effectiveness. An important advantage of this approach is that it permits a fully Bayesian approach to the problem, providing a principled and accessible way to measure statistical error. The approach is demonstrated by applying the *minimum description length* principle. This technique is used in model selection to provide a rational way of comparing models of varying complexity. We show how the resulting score may be interpreted and applied as a Bayesian posterior density.

Keywords

Pedigree reconstruction; graphical models; Bayesian inference; minimum description length

1 Introduction

The collection of molecular marker data in studies of natural populations permits the statistical estimation of relatedness structure among individuals. Such information is crucial in the observation and measurement of quantities such as fitness, trait heritability, migration or effective population size. Relatedness itself may be the subject of interest in selective breeding or conservation applications. The growing importance of this type of analysis is indicated in recent surveys, including Blouin (2003), Jones and Ardren (2003), Garant and Kruuk (2005) and Thomas (2005).

Relatedness inference can vary in level of detail, ranging from aggregate measures of relatedness, pairwise estimates of relatedness type, or fully specified pedigrees (ie. family trees). Additionally, interest may be confined to a particular relationship type, in particular, parent-offspring relationships (eg Thompson and Meagher 1987; Marshall *et al* 1998; Nielsen *et al* 2001) or sibling relationships (eg Almudevar 1999; Thomas and Hill 2002; Butler 2004; Wang 2004; Konovalov 2005).

As an alternative to the categorical assignment of parentage, some authors have developed the idea of fractional assignment, in which a parentage assignment to a fixed offspring consists of an attribution of weights to each candidate father, the size of the weight corresponding to the evidence of parentage. The weights may be forced to sum to one, and so may be interpreted as

a probability distribution. This method has been developed in, for example, Devlin *et al* 1988, Smouse and Meagher (1994), Nielsen *et al* 2001 and Neff *et al* 2001. There are two rationales for this approach. The first is that because parentage assignment involves statistical error, it is appropriate to express such inference in the form of a posterior distribution, so that the uncertainty of the assignment can be assessed. Furthermore, this approach allows the introduction of auxiliary information (age, proximity or other mating attributes) in the form of a prior distribution, allowing a formal Bayesian approach. The second rationale is that when the object of study is the inference of mating or fitness patterns at the population level, individual assignment is not needed, and the aggregate measures of interest are more naturally expressed in terms of fractional paternal contributions.

Various methods for the measurement of statistical error have been proposed, in addition to the fractional assignment methods discussed above. The commonly used parentage assignment algorithm CERVUS, proposed in Marshall *et al* 1998, is based on a formal hypothesis test. A Bayesian approach for sibling relationships was proposed in Emery *et al* 2001, and a bootstrap approach to measuring statistical error in sibling partitions was proposed in Almudevar (2001a). The estimation of statistical error in relationship estimation is often accomplished by simulating from known pedigrees. However, a complete and proper inference requires an estimate of statistical error without the benefit of a known pedigree, and such methodologies remain to be developed for many of the procedures discussed above.

Given the variety of objectives and sampling regimes associated with relationship inference, it seems unlikely that any single methodology will be an optimal choice for all applications. Nonetheless, a general framework should be possible by concentrating attention on single joint relationship structures. Essentially, we define a parameter space Θ to be a set of combinatorical objects, such as pedigree graphs (as in Almudevar 2003), or partitions generated by sibling groups estimated in the various techniques described above. Principles of statistical inference may be applied, leading to rigorous inferential statements expressible in terms of Θ .

There are two significant challenges associated with this approach. First, comparing large combinatorical objects presents special difficulties when they differ in complexity. In particular, likelihood scores tend to favor more complex explanations. This is manifest, for example, in the tendency of likelihood scores to split true sibling groups (Thomas and Hill 2000; Butler 2004). Alternative scores, such as the Simpson's Index proposed in Butler (2004) (see also Konovalov 2005), favor larger sibling groups and are therefore better able to preserve such family structure. This suggests that the selection of a method which properly accounts for the sample size and relationship complexity is crucial.

The second challenge is the formation of rigorous inferential statements concerning Θ . Here, *Bayesian model averaging* (BMA) (Hoeting *et al* 1999) may be used to assign a confidence level to any model feature. Given data X , suppose we may construct a posterior density $\phi(\theta | X)$ on Θ , the space of pedigree structures. The posterior probability of any feature, for example, " a is a sibling or half-sibling of b ", is calculated by summing the posterior density over all $\theta \in \Theta$ possessing that feature. Thus, even when the objective is the estimation some aggregate feature of the pedigree, and a pedigree estimate is not strictly needed, the more general approach allows a wide variety of problems to be solved using essentially one single methodology.

In this article, we will develop this idea for pedigree graphs, using the *minimum description length* (MDL) principle due to Rissanen (1978, 1983) as a means of comparing candidate models. The objective of the MDL principle is to uncover regularity which permits the most compressed representation of the data X . This regularity is presumed to have a natural interpretation with respect to the models in Θ , so the best model is taken to be one which permits the most compression. Because the representation must also include the model itself (and more

complex models require longer representations) this method tends to avoid the inference of spurious complexity common in unadjusted likelihood methods. Of course, the data compression is not actually performed. We only need an estimate of the length of the resulting data file that would follow compression based on a specific model. Well known techniques from coding theory may be used to do this. For a recent survey of this field see Grünwald *et al* 2005.

The methods developed will be tested on simulated data from test pedigrees in Section 4, and on a single cohort of 857 North Atlantic cod larvae, supplemented with an additional simulated parental cohort in Section 5.

A version of the software used in this article may be downloaded at http://www.urmc.rochester.edu/smd/biostat/projects/help/PC/Software_Listings.htm.

2 Bayesian model averaging for graphical models

Multigenerational pedigrees possess a natural graph structure, so we will develop our methodology on that basis. We take a directed graph $\theta = (E, V)$ to be a set V of labelled objects (nodes) and a set E of directed edges, defined as an ordered pair from V . Each node corresponds to a subject (labelled $1, \dots, N_I$), and an edge $j \rightarrow i \in E$ implies a parent-offspring relationship for j and i , respectively. Note that θ must be a *directed acyclic graph* (DAG) with no more than 2 parents for any child node.

We associate with each node i an observation pair (X_i, Y_i) consisting of all data associated with that subject. The datum X_i represents heritable data, that is, data for which the conditional densities $f(X_i | X_j)$ and $f(X_i | X_j, X_k)$ are known when j, k are parents of i . The datum Y_i represents demographic data (age, sex, geographic) which can be used to test the admissibility of a mating relationship or a parent-offspring relationship. Denote the complete data sets $X = (X_1, \dots, X_{N_I})$ and $Y = (Y_1, \dots, Y_{N_I})$. The data is assumed complete in the sense that all relationships can be expressed using parent-offspring pairs (for example, the parents of sibling groups present in the sample are also present in the sample).

Let Θ be the set of all DAGs with nodes $1, \dots, N_I$ having a maximum of two parents. For any $\theta \in \Theta$ let S_i^θ be the set of parents of i . We assume that the density of X may be written

$$f(x_1, \dots, x_{N_I}) = \prod_{i=1}^{N_I} f(x_i | X_j = x_j, j \in S_i^\theta). \quad (1)$$

This will be the case for unlinked genotypic data. Under (1), the pair (θ, f) is an example of a *Bayesian network*, a graph-based model used to describe multivariate data characterized by causal relationships (Pearl 1988). Bayesian networks have been used in a variety of fields ranging from the modelling of cell regulatory networks to the design of artificial intelligence applications (Sebastiani *et al* 2005; Jordan 1998), and an extension of such methodology to the modelling of pedigrees would seem natural.

2.1 Structure of model averaging

Inferential statements will be concerned with subsets $E \subset \Theta$, for example $E = \{\theta \in \Theta: i \text{ and } j \text{ are unrelated}\}$. Given a score function $\phi(\theta | X)$ interpretable as a posterior density on the space of graphs, we may assign a posterior probability for E

$$P(E | X) = \sum_{\theta \in E} \phi(\theta | X). \quad (2)$$

A marginal posterior density for a specific quantity, say the fitness w_i of individual i is then given by

$$\phi_{w_i}(w | X) = \sum_{\theta: w_i = w} \phi(\theta | X). \quad (3)$$

The space of graphs is generally too large to enumerate, so in practice (2) or (3) are estimated using computer generated samples from $\phi(\theta | X)$. Where independent sampling is not possible, Monte Carlo Markov chains (MCMC) are a commonly used alternative, but present certain difficulties. This method requires a definition of a graph neighborhood $N(\theta)$ specified for each $\theta \in \Theta$. A sample $(\theta_1, \dots, \theta_j, \theta_{j+1}, \dots)$ from $\phi(\theta | X)$ is generated by randomly selecting θ_{j+1} from $N(\theta_j)$ according to a probability law defined by the Hastings-Metropolis criterion (Hastings 1970). The effectiveness of such a procedure depends on the ability to define a neighborhood $N(\theta)$ which is small enough to ensure that a sizable fraction of graphs $\theta' \in N(\theta)$ have a score similar to that of θ , but large enough to allow the process to traverse a sufficiently large subset of Θ . In the case of DAGs, this is usually achieved by defining $N(\theta)$ to be those graphs obtainable by performing a simple operation on θ , such as deleting an edge, adding an edge or reversing an edge (see Madigan and York 1995). The difficulty in this approach is that such an operation on a DAG may introduce a cycle into the subsequent graph. Prohibiting such transitions has implications for the performance of such a sampler (Friedman and Koller 2003).

2.2 The advantage of order constraints

Age data can play a significant role in the design of such samplers. Suppose we are given an ordering of the nodes. Any directed graph constructed by selecting for each node an admissible set of parents of higher order will be a DAG (Almudevar 2003). The set of graphs Θ that are admissible under such an order constraint then assumes a convenient product space form, equivalent to independent selections of admissible parent sets for each node. The neighborhood $N(\theta)$ then becomes easy to define, eliminating the need to flag transitions to non-acyclic graphs. Of course, such an ordering will be imposed by age data.

Possibly, only partial age or ordering data is available, in which case it is still possible to confine the search space to those graphs conforming to the order. For example, we may be able to partition a sample into distinct generations. If a parent must come from an older generation, then the graph space Θ retains the product form of the full ordering constraint. If a parent may come from the same generation as an offspring, the product form no longer holds, since the selection of the parents of two members of the same generation cannot be made independently.

It is possible to exploit the desirable properties of the order constraint when age data is not available. If integration or maximization operations under an order constraint are considerably simplified, then we may replace the graph search space with an ordering search space, which is smaller and more tractable. This approach was used in Almudevar (2003) to determine maximum likelihood pedigrees using simulated annealing, and in Friedman and Koller (2003) to construct a Bayesian posterior density on the space of Bayesian networks. One drawback of the latter approach is that orderings do not partition the graph space, and the number of orderings to which a graph conforms varies, making it difficult to reconstruct a given posterior density. Other approaches include a data-based pruning approach (Madigan and Raftery 1994), which reduces the number of graphs to be considered. The case of partial orderings has been considered in Dash and Cooper (2004). Finding an efficient method of calculating a posterior density over graphical models remains an important area of research.

2.3 Posterior densities in product form

The construction of a posterior density is particularly convenient when φ assumes a product form and an ordering is available. We may express the ordering by associating with individual i the set of all parent sets $H_i^\theta = (s_{i,1}, \dots, s_{i,n_i})$ which conform to that ordering. If the posterior density has form

$$\phi(\theta | X) = \prod_{i=1}^{N_I} \phi_i(X_i | s_i^\theta) \quad (4)$$

for admissible θ then the parent assignments under φ are independent. Thus, to sample a pedigree θ from φ we can sample independently for each i a parent set $s \in H_i^\theta$ from distribution

$$P(s_i^\theta = s | X) = \begin{cases} \frac{\phi_i(X_i | s)}{\sum_{s' \in H_i^\theta} \phi_i(X_i | s')} & : s \in H_i^\theta, \\ 0 & ; s \notin H_i^\theta \end{cases}$$

so that we need not rely on MCMC methods. In this article we will examine two posterior densities, both of which are in product form.

3 The MDL principle

The *minimum description length* (MDL) principle was first proposed by Rissanen (1978, 1983), and is based on the relationship between data compression and data regularity. Suppose a data set X is to be coded as a binary sequence $b(X)$. The coding must be such that X may be recovered from $b(X)$ using a general purpose algorithm. An important principle of coding theory states that the more regularity exists in X , the shorter the length of $b(X)$, provided efficient coding techniques which are able to exploit the given form of regularity are used.

In a modelling selection context, if X is generated by a model $\theta \in \Theta$ then knowledge of that model can be used to further compress X . Of course, when the model is one of a set of candidate models, we need to know which model was used to generate the compression. We can denote the encoded data $b(X | \theta)$, using notation which emphasizes the dependence of the coding on the model. We also need to encode θ , as $b(\theta)$, the knowledge of which is needed to decode $b(X | \theta)$. The total code is then $b(X | \theta)b(\theta)$. Let $B(X | \theta)$ and $B(\theta)$ represent the code lengths of $b(X | \theta)$ and $b(\theta)$, giving total code length $B(\theta | X) = B(X | \theta) + B(\theta)$. The MDL principle states that the model which minimizes $B(\theta | X)$ provides the best explanation for the data. In this section we will discuss the application of the MDL principle to the problem of pedigree graph inference.

Assume that we have N_L loci of independently segregating genotypic data. We assume that population genotype frequencies are known. Let X_i denote the genotypic data for individual i . Let $n(k, i)$ be the number of candidate parent groups of size $k = 0, 1, 2$ available to offspring i . This quantity is possible affected by demographic data Y . For example if $N_I = 100$ and 50 individuals (of unknown sex) are old enough to be parents of i , then $n(0, i) = 1$, $n(1, i) = 50$, $n(2, i) = (50 \times 49)/2$. These numbers would be adjusted accordingly if sex data were available. A coding scheme conforming to the MDL principle was developed for this model (see equations (10)–(12) in Appendix), resulting in estimated code length

$$B(X | \theta) = -\log_2 L(\theta | X)$$

$$B(\theta) = \sum_{i=1}^{N_I} \log_2 (n(|S_i^\theta|, i))$$

where $L(\theta | X)$ is the pedigree likelihood function

$$L(\theta | X) = \prod_{i=1}^{N_I} P(X_i | S_i^\theta).$$

The conditional probability $P(X_i | S_i^\theta)$ is the probability of genotype X_i given parent set S_i^θ , and is easily calculated using the laws of Mendelian inheritance. We obtain an interpretable density through the exponential transformation

$$\begin{aligned} \phi(\theta | X) &= (1/2)^{B(X|\theta)+B(\theta)} \\ &= L(\theta | X) \prod_{i=1}^{N_I} \frac{1}{n(|S_i^\theta|, i)}. \end{aligned} \quad (5)$$

This transformed MDL score has a natural Bayesian interpretation under which a prior density on the space of models is given by

$$P(\theta) \propto \prod_{i=1}^{N_I} \frac{1}{n(|S_i^\theta|, i)},$$

then taking $P(X|\theta) = L(\theta | X)$ we have posterior density

$$\phi(\theta | X) \propto P(\theta | X).$$

The prior and posterior densities possess a number of desirable features. First of all, they may be written in product form:

$$\phi(\theta | X) = \prod_{i=1}^{N_I} \frac{P(X_i | S_i^\theta)}{n(|S_i^\theta|, i)},$$

simplifying posterior sampling as discussed in Section 2.3. Second of all, important marginal features of the prior do not depend on N_I . For example, under the prior density, the probability of any parent set of size k for individual i is proportional to $1/n(k, i)$, hence the prior distribution of the number of parents of an individual is given by

$$P(|S_i^\theta| = k) \propto \sum_{j=1}^{n(k,i)} \frac{1}{n(k, i)} = 1,$$

that is, $P(|S_i^\theta| = k) = 1/3$. In contrast, a likelihood score corresponds to a uniform prior on Θ and so the prior distribution of parent set size would be defined by $P(|S_i^\theta| = k) \propto n(k, i)$. We would generally expect $n(k, i)$ to be an order k polynomial in N_I , so that the effect of increasing N_I would be to increase $P(|S_i^\theta| = 2)$. In other words, under a likelihood score, the prior expectation of complexity increases with N_I , introducing a complexity bias into any BMA procedure.

4 Simulation study

To evaluate the proposed methodology, a simulation study was undertaken based on genotype data simulated on known test pedigrees. These test pedigrees are generated from two base pedigrees, labelled A and B , shown in Figure 1. Base pedigree A is the union of subpedigrees $A1$ and $A2$, similarly, $B = B1 \cup B2$. The subpedigrees are taken to represent *fitness classes* in the sense that each mating in such a class always produces the same number of offspring. A *founder* is an individual of the oldest generation. The characteristics of the base pedigrees are summarized in Table 1. The simulation study will be concerned with the inference of the number of descendants of a founder.

Each base pedigree generates 4 test pedigrees, in each case using the largest number of replications resulting in a pedigree not exceeding target sizes $N_I = 200, 500, 2000, 10000$. This results in 8 test pedigrees, with sizes ranging from 152 and 150 (A replicated once, B replicated twice) to 9880 and 9975 (A replicated 65 times, B replicated 133 times). For convenience, pedigree sizes are referred to by the target sizes. The pedigrees are selected to represent distinct sampling and mating characteristics. Pedigrees $A1$ and $A2$ may be described as closed systems, in the sense that matings occur between offsprings of proximate families, and parents of all nonfounders in the sample are also in the sample. Pedigrees $B1$ and $B2$ represent open systems, in the sense that offspring of individuals in the sample mate with individuals without ancestors in the sample.

We will consider two pedigree scores, the likelihood score and the MDL score defined in (5). We assume age (*ie.* generation) data is given. Parents are always assumed to be of the previous generation. This assumption is incorporated into the definition of the number of candidate parents $n(k, i)$. No sex information is used. We assume a locus has eight alleles with allele frequencies conforming to a Zipf distribution ($p_i \propto 1/i$, $i = 1, \dots, 8$, with heterozygosity = 79.3%). The simulation study is repeated using 8, 10 and 12 loci. The posterior densities used are in product form and can be estimated using independent simulations (see Section 2.3).

For each pedigree let $\mathcal{F}$ denote the set of all founders, and let d_i denote the number of descendants for $i \in \mathcal{F}$. Let $\mathcal{F}_F^a, \mathcal{F}_F^b$ be the founders from the higher and lower fitness class respectively, and let d^a, d^b be the average number of founder descendants among the respective classes. For example, in a pedigree of type A , $\mathcal{F}_F^a$ consists of all founders of fitness 4, $\mathcal{F}_F^b$ consists of all founders of fitness 2, so that $d^a = 84$ and $d^b = 14$.

Additionally, note that a pedigree θ can be described by two parent specifications for each individual (one or both of which may be 'population' in the case of founders). A distance $D(\hat{\theta}, \theta)$ between a pedigree estimate $\hat{\theta}$ and the true pedigree θ is taken to be the number of erroneous parent specifications (0, 1 or 2 for each individual).

To summarize, we consider 8 test pedigrees, varying the number of loci $L = 8, 10, 12$ for each, for a total of 24 scenarios. For each scenario two simulation experiments are performed:

- E1.** Simulate genotypes X for test pedigree. Construct maximum/minimum score pedigree using likelihood and MDL score. Do $N = 10000$ independent trials.
- E2.** Simulate genotypes X for a single pedigree. Sample $N = 2000$ replications from the posterior density based on likelihood and MDL scores given X .

The analysis will be as follows. For experiment **E1** let $\theta(j)$ be the estimated pedigree for trial j . For each founder in $\mathcal{F}^a$ and $\mathcal{F}^b$ calculate the number of descendants based on $\theta(j)$. Report $d^a(j)$, $d^b(j)$, the average number of descendants for founders in $\mathcal{F}^a$, $\mathcal{F}^b$ respectively. Then calculate $D(j) = D(\theta(j), \theta)$, where θ is the true pedigree. We report:

$$\begin{aligned}
 \bar{d}^a &= (1/N) \sum_{j=1}^N d^a(j) \\
 \bar{d}^b &= (1/N) \sum_{j=1}^N d^b(j) \\
 RMSE^a &= \left[(1/N) \sum_{j=1}^N (d^a(j) - \bar{d}^a)^2 \right]^{1/2} \\
 RMSE^b &= \left[(1/N) \sum_{j=1}^N (d^b(j) - \bar{d}^b)^2 \right]^{1/2} \\
 D &= (1/N) \sum_{j=1}^N D(j)
 \end{aligned} \tag{6}$$

that is, the mean estimate and root mean square error (RMSE) of d^a and d^b , and the mean distance D , where $N = 10000$.

For experiment **E2** given simulated data X graphs $\theta(1), \dots, \theta(N)$ are sampled from the posterior density. Here $N = 2000$. Let $D(j) = D(\theta(j), \theta)$. For each founder $i \in \mathcal{F}^a \cup \mathcal{F}^b$ the posterior density of d_i is sampled, giving $d_i(1), \dots, d_i(N)$, which are calculated directly from $\theta(1), \dots, \theta(N)$. For each founder the following quantities are calculated:

$$\begin{aligned}
 \bar{d}_i &= (1/N) \sum_{j=1}^N d_i(j) \\
 SD_i &= \left[(1/(N-1)) \sum_{j=1}^N (d_i(j) - \bar{d}_i)^2 \right]^{1/2} \\
 QP_i &= (1/N) \sum_{j=1}^N I\{d_i(j) < d^0\} + (1/2)I\{d_i(j) = d^0\}
 \end{aligned}$$

where d^0 is the true number of descendants (that is, d^a or d^b as appropriate). Then $\bar{d}_i$ and SD_i are the mean and standard deviation of the posterior density of d_i . Additionally, QP_i is the quantile position of the true number of descendants d^0 . This definition centers the position. For example, if the frequencies of the replications are (500/2000, 1000/2000, 500/2000) for quantities (38, 39, 40), and the true value is $d^0 = 39$, then the quantile position evaluates to $QP_i = 0.5$. In general, if the posterior density gives accurate coverage, we would expect QP_i to be close to 0.5. If, in addition, SD_i is reasonably small, the posterior density becomes a useful inference tool. We then report

$$\begin{aligned}
\tilde{d}^a &= (1/n^a) \sum_{i \in \mathcal{J}_F^a} d_i \\
\tilde{d}^b &= (1/n^b) \sum_{i \in \mathcal{J}_F^b} d_i \\
\tilde{SD}^a &= (1/n^a) \sum_{i \in \mathcal{J}_F^a} SD_i \\
\tilde{SD}^b &= (1/n^b) \sum_{i \in \mathcal{J}_F^b} SD_i \\
\tilde{QP}^a &= \text{median}_{i \in \mathcal{J}_F^a} QP_i \\
\tilde{QP}^b &= \text{median}_{i \in \mathcal{J}_F^b} QP_i \\
\tilde{D} &= (1/N) \sum_{j=1}^N D(j)
\end{aligned} \tag{7}$$

where n^a, n^b are the numbers of founders in $\mathcal{J}_F^a, \mathcal{J}_F^b$ respectively.

The summaries $\tilde{d}^a, \tilde{d}^b, RMSE^a, RMSE^b$ defined in (6) for experiment **E1** for the case of 10 loci are shown in Figure 2 (see Table 3 for numerical summaries). For pedigree type A the MLE score results in a marginally more accurate estimate of (d^a, d^b) . For pedigree type B, the MDL score gives a significantly better estimate of (d^a, d^b) . Note the significant overestimation for the $N_I = 2000, 10000$ scenarios for the MLE score, not exhibited by the MDL score. The graph distances are given in Table 3, and similarly indicate marginally more accuracy of the MLE score for pedigree type A and significantly more accuracy for the MDL score for pedigree type B. To summarize, there is a marginal advantage to using the MLE score for pedigree type A, and a more significant advantage to the MDL score for pedigree type B, which represent precisely the scenario in which complexity control can be expected to exhibit an advantage.

We next examine how this situation changes when the amount of information is decreased to 8 loci. The summaries are given in Table 2. For pedigree A, the accuracy of the MLE has degraded to some degree from the 10 loci scenario, with estimates of d^a and d^b tending to shrink towards a common value for $N_I = 2000, 10000$ due to misassignment. The MDL score significantly underestimates (d^a, d^b) for $N_I = 10000$. This effect is to be expected, since the relative contribution of the data to the score is directly dependant on the number of loci, while the remaining terms of the score tend to penalize relationships. For pedigree type B, the MLE score tends to overestimate, (d^a, d^b) , where the MDL score tends to underestimate. Note that for pedigree type B the MDL score is more accurate by all *RMSE* and distance measures.

Table 4 gives summaries for the 12 loci scenario. This increase in information suffices to produce accurate estimates of (d^a, d^b) with either the MLE or MDL score for pedigree type A. The summaries can be seen to be very similar. The difference in accuracy for pedigree type B is more striking. While very small bias is evident using the MDL score, the overestimation previously seen for the MLE score persists. *RMSE* and distance measures indicate significantly better overall accuracy for the MDL score.

We now consider the feasibility of the fully Bayesian approach proposed in this article. A feasible posterior density provides simultaneously an estimate and a measure of model

uncertainty. We therefore expect the quantile positions $\tilde{QP}^a, \tilde{QP}^b$ to be close to 0.5, while maintaining reasonably small standard deviations $\tilde{SD}^a, \tilde{SD}^b$ (a value $\tilde{QP}^a > 0.5$ indicates negative bias).

Summaries of the various posterior densities defined in (7) are given in Table 3 for the 10 loci scenario. For pedigree type A, the posterior densities generated by the likelihood score remains feasible for all N_I . However, the posterior density for d^a generated by the MDL-based posterior for $N_I = 10000$ is not feasible, with $\tilde{QP}^a = 0.95$, so that the true value $d^a = 84$ is consistently located well into the upper tail of the individual posterior densities of d_i for $i \in \mathcal{J}_P^a$. For pedigree type B, the MDL-based posterior generates feasible posterior densities, while for the likelihood-based posterior the quantile positions are reported to be 0 for $N_I = 2000, 10000$, indicating very poor coverage of the posterior densities of d_i .

When the number of loci is reduced to 8, the posterior densities generated by the MDL score for pedigree type A result in consistently high quantile positions, rendering the densities infeasible. The likelihood-based densities remain feasible, although some bias is evident. This situation is reversed for pedigree type B, with very small quantile positions for the likelihood-based posteriors, and feasible (but biased) MDL-based posteriors.

When the number of loci is increased to 12 the MDL-based posteriors become feasible for pedigree type A, although some bias for $N_I = 2000, 10000$ is still evident. The advantage of the MDL score over the likelihood score for pedigree type B noted for the 10 loci scenario is maintained for 12 loci.

Individual posterior densities of d_i are shown in Figures 3–5, corresponding to the 8, 10 and 12 loci scenarios, for $N_I = 10000$. Two founders are selected from each of 20 base pedigree replications, one from each fitness class. Posterior densities are represented by standard boxplots constructed from the 2000 replications of the selected individual values of d_i . Consistent with the above discussion, the likelihood-based posterior performs consistently well for pedigree type A, with a clear improvement with additional loci. Similarly, the MDL-based posterior performs consistently well for pedigree type B, again with clear improvement with additional loci. We note additionally that for 8 and 10 loci, the likelihood-based posterior tends to overestimate d_i for pedigree type B, while the MDL-based posterior tends to underestimate d_i for pedigree type A. On the other hand, when 12 loci are used, the MDL-based posterior is able to produce plausible posterior densities for pedigree type A, whereas the likelihood-based posterior maintains its tendency to overestimate d_i for pedigree type B.

To summarize, for pedigree type A the likelihood score provides generally better inference, both in terms of point estimation, and the construction of feasible posterior densities. This advantage becomes marginal when more loci are used. Conversely, for pedigree type B, the MDL score produced consistently better inference than the likelihood score. This advantage was more significant, and showed less of a tendency to become marginal with increased numbers of loci. Thus, provided sufficiently informative data, the MDL score can be expected to perform well over the range of pedigree types considered here, whereas significant overfitting by the likelihood score was evident even for the 12 loci scenario for pedigree type B.

5 Example (North Atlantic Cod)

In Herlinger *et al* 1997 microsatellite markers from 6 loci were collected on a sample of 857 Atlantic cod larvae from a single cohort. The loci were assumed to be unlinked. Various

statistical hypothesis tests (Almudevar and Field 1999; Almudevar 2001b) failed to find significant evidence of sibling structure.

To demonstrate the methodology, the single cohort was supplemented with a simulated single parent generation. Some number n_t of the cohort were randomly selected. One *actual* parent for each was simulated by selecting at random one gene per locus from the offspring. The other gene is selected according to population allele frequencies estimated from the original cohort. A further $n_{spurious}$ parents were simulated by selecting genotypes according to the estimated population frequencies. The generation memberships are made available to the algorithm.

Posterior densities based on the likelihood and MDL scores were estimated as in the previous example, with attention placed on the offspring distributions of the parental generation. Two quantities were captured, first, the posterior probability of parenthood to the cohort of candidate parent i (that is, i possesses at least one offspring), α_i , and the posterior probability of a parent-offspring relationship, β_{ij} for candidate parent and offspring i and j . These quantities were estimated using equation (2) applied to the sampled pedigrees, amounting to the proportion of pedigrees possessing the feature in question.

In Figure 6, histograms of all nonzero values of α_i are constructed, separately for actual and spurious parents i , for each posterior density. For the likelihood-based posterior, the values are concentrated near 1 for both actual and spurious parents, whereas for the MDL-based posterior the histograms exhibit significant separation. We would naturally adopt a protocol under which parenthood is not assumed for candidate parents for which $\alpha_i = 0$, leaving open the question of when to infer parenthood for $\alpha_i > 0$. The separation in α_i induced by the MDL score suggests that a better specificity to sensitivity tradeoff is possible using that score.

To see this, we adopt a threshold rule under which parenthood is assumed for candidate parents for which $\alpha_i \geq p$. Figure 7 shows the estimates of the number of candidate parents assigned parenthood, for varying values of p , categorized as true or false according to whether the parent is actual or spurious. The estimates are given for both posterior densities. As can be seen, the MDL curve is strictly below the likelihood curve, meaning that the MDL-based posterior induces fewer false parenthood estimates given equal true parenthood estimates, compared to the likelihood-based posterior. The estimates induced by the maximum scoring pedigree is also indicated, which lie approximately on the characteristic curves. This suggests that there is no benefit to comparing the maximum scoring pedigree separately, since it merely represents one possible estimate along a continuous tradeoff of sensitivity to specificity.

A similar analysis was carried out for β_{ij} . In Figure 8, nonzero values of β_{ij} are represented in histogram form, separately for actual parent-offspring pairs; spurious pairs when i is an actual parent; and spurious pairs when i is a spurious parent. In each case no difference in the distribution of β_{ij} among spurious pairs involving actual or spurious parents is evident, so the values are pooled for further analysis.

Similarly to the values of α_i , a greater separation between the distributions of β_{ij} for actual and spurious pairs is evident for the MDL-based posterior. The same threshold rule analysis was carried out for the values of β_{ij} , with the objective of estimating the total number of parent-offspring pairs. Results are given in Figure 9. Again, the MDL-based posterior provides a better sensitivity to specificity tradeoff. Note also that the estimates obtained from the maximum scores are located approximately on the characteristic curves.

6 Conclusion

In this article, the various advantages of a model based approach to relationship inference, in which all relationships in a sample are represented by a single combinatorial object were

considered. Principally, these are the ability to control for spurious estimates of complexity, and a natural method of accounting for statistical uncertainty by using a computationally intensive fully Bayesian approach.

As a specific application, the inference of a pedigree graph using a score based on the *minimum description length* (MDL) principle was developed. Age data was used to simplify the construction of a Bayesian posterior density. Data assumed the form of genotypic data from varying numbers of loci.

The procedure was applied to a variety of simulated pedigree data. The MDL-based procedure was compared to a likelihood-based procedure. It was found that for relatively complete pedigrees in which all individuals, other than a small set of founders, have both parents in the sample the likelihood approach was marginally more accurate than the MDL approach, although this advantage became negligible with an increase in the number of loci simulated. For less complete pedigrees, the MDL score exhibited significantly better accuracy, which persisted when the number of loci were increased.

Additionally, the proposed method of constructing posterior densities yielded generally acceptable results, in the sense that coverage tended to include true values in cases for which accuracy was also acceptable.

In addition the procedure was applied to a single cohort of 857 North Atlantic cod larvae, supplemented by a simulated parental cohort consisting of a mix of actual and spurious parents. Individual posterior probabilities of parenthood and parent-offspring status using both posteriors were compared, demonstrating a better sensitivity to specificity tradeoff using the MDL-based procedure. More generally, it was shown that the calculation of such posterior probabilities was feasible, and offer the possibility of a more refined inference procedure than that based on the calculation of a single maximum score pedigree.

While it is important to establish the feasibility of such larger scale modelling techniques for relationship inference, we also must note that the assumptions adopted here are somewhat restrictive in a number of ways. We have not accounted for genotype errors, which are regularly observed in large scale genotype screens (Blouin 2003). These errors can have disproportionate effects, since a single change in gene value can rule out by exclusion a true parent-offspring pair. However, likelihoods can be modified to include error models (Marshall *et al* 1998; Wang 2004), which could then be extended to any scoring procedure.

Additionally, age data may be incomplete (or non-existent). In this case, a score function in product form (4) could be optimized using the simulated annealing approach of Almudevar (2003). This method does not extend naturally to the problem of sampling from a posterior density, however. In this case, MCMC methods on graph spaces may be substituted, and may incorporate any partial age data that is available, thus reducing the search space. Applying simulated annealing or MCMC is generally more challenging than independent sampling, and the application to samples numbering in the thousands remains largely untested.

The procedure proposed here depends on the availability of allele frequencies. These may be estimated from the data, although independent estimates would be preferable, since estimation of heterozygosity of the allele distributions may be confounded with the presence of relatedness. One alternative is to use Bayesian priors in place of estimated frequencies (Painter 1997).

In principle, a likelihood function can incorporate linkage among typed loci, but the product form of the likelihood would no longer apply. As in the case of incomplete age data, the problem

may be solved by using MCMC on graph spaces, at the cost of foregoing the advantages of the product form.

Acknowledgements

This work was supported by NIH grant GM075299.

References

1. Almudevar A, Field C. Estimation of single generation sibling relationships based on DNA markers. *JABES* 1999;4:136–165.
2. Almudevar A. A bootstrap assessment of variability in pedigree reconstruction based on DNA markers. *Biometrics* 2001a;57:757–763. [PubMed: 11550925]
3. Almudevar A. Most powerful permutation invariant tests for relatedness hypotheses based on genotypic data. *Biometrics* 2001b;57:1080–1088. [PubMed: 11764247]
4. Almudevar A. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology* 2003;63:63–75. [PubMed: 12615491]
5. Blouin MS. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution* 2003;18:503–511.
6. Butler K, Field C, Herbinger CM, Smith BR. Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Molecular Ecology* 2004;13:1589–1600. [PubMed: 15140101]
7. Dash D, Cooper GF. Model averaging for prediction with discrete Bayesian networks. *Journal of Machine Learning Research* 2004;5:1177–1203.
8. Devlin B, Roeder K, Ellstrand NC. Fractional paternity assignment: theoretical development and comparison to other methods. *Theor Appl Genet* 1988;76:369–380.
9. Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR. Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Molecular Ecology* 2001;10:1265–1278. [PubMed: 11380883]
10. Friedman N, Goldszmidt M. Learning Bayesian networks with local structure. In: Jordon, MI., editor. *Learning in Graphical Models*. MIT press; Cambridge, Massachusetts: 1998. p. 412-459.
11. Friedman N, Koller D. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* 2003;50:95–125.
12. Garant D, Kruuk LEB. How to use molecular marker data to measure evolutionary parameters in wild populations. *Molecular Ecology* 2005;14:1843–1859.
13. Grünwald, PD.; Myung, IN.; Pitt, MA. *Advances in minimum description length theory and applications*. MIT Press; Cambridge, Massachusetts: 2005.
14. Hamming, RW. *Coding and Information Theory*. Prentice-Hall, Englewood Cliffs; New Jersey: 1986.
15. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;57:97–109.
16. Herbinger C, Doyle RW, Taggart CT, Lochmann SE, Brooker AL, Wright JM, Cook D. Family relationships and effective population size in a natural cohort of Atlantic cod (*Gadus morhua*) larvae. *Canadian Journal of Fisheries and Aquatic Sciences* 1997;54:11–18.
17. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical Science* 1999;14:382–417.
18. Jones AG, Ardren WR. Methods of parentage analysis in natural populations. *Molecular Ecology* 2003;12:2511–2523. [PubMed: 12969458]
19. Jordan, M. *Learning in Graphical Models* (Adaptive Computation and Machine Learning. MIT Press; Cambridge, MA: 1998.
20. Konovalov DA, Bajema N, Litow B. Modified SIMPSON O(n³) algorithm for the full sibship reconstruction problem. *Bioinformatics* 2005;21:3912–3917. [PubMed: 16118259]
21. Madigan D, York J. Bayesian graphical models for discrete data. *International Statistical Review* 1995;63:215–232.

22. Madigan D, Raftery E. Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association* 1994;89:1335–1346.
23. Marshall TC, Slate J, Kruuk LEB, Pemberton JM. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* 1998;7:639–655. [PubMed: 9633105]
24. Neff BD, Repka J, Gross MR. A Bayesian framework for parentage analysis: the value of genetic and other biological data. *Theoretical Population Biology* 2001;59:315–331. [PubMed: 11560451]
25. Nielsen R, Mattila DK, Clapham PJ, Palsbøll PJ. Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics* 2001;157:1673–1682. [PubMed: 11290722]
26. Painter I. Sibling reconstruction without parental information. *JABES* 1997;2:212–229.
27. Pearl, J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers; San Mateo, California: 1988.
28. Rissanen J. Modeling by the shortest data description. *Automatica* 1978;14:465–471.
29. Rissanen J. A universal prior for integers and estimation by minimum description length. *Annals of Statistics* 1983;11:416–431.
30. Sebastiani, P.; Abad, M.; Ramoni, MF. Bayesian networks for genomic analysis. In: Dougherty, ER.; Shmulevich, I.; Chen, J.; Wang, ZJ., editors. *Genomic Signal Processing and Statistics*, EURASIP Book Series on Signal Processing and Communications. Hindawi Publishing Corporation; 2005.
31. Smouse PE, Meagher TR. Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) gray (Liliaceae). *Genetics* 1994;136:313–322. [PubMed: 8138167]
32. Thomas SC. The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society B* 2005;360:1457–1467.
33. Thomas SC, Hill WG. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* 2000;155:1961–1972. [PubMed: 10924488]
34. Thomas SC, Hill WG. Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research* 2002;79:227–234. [PubMed: 12220130]
35. Thompson EA, Meagher TR. Parental and sib likelihoods in genealogy reconstruction. *Biometrics* 1987;43:585–600. [PubMed: 3663818]
36. Wang JL. Sibship reconstruction from genetic data with typing errors. *Genetics* 2004;166:1963–1979. [PubMed: 15126412]

7 Appendix

In this appendix we develop a pedigree score based on the MDL principle, under the assumptions discussed in Section 3. The resulting score is not provably optimal (in the sense of representing the maximum possible data compression), but rather relies on an intuitively reasonable approach to the problem.

7.1 Design of efficient codes

Much of the application of the MDL principle depends on two basic results from coding theory, which we introduce in this section. Suppose we have an *alphabet* $A = \{a_1, \dots, a_k\}$, and define a *word* to be any finite string of symbols from A . It is possible to encode a word X of length $n = |X|$ into a decodable binary string $b(X)$ using $n \log_2 k$ bits (letting $|x|$ represent the length, or number of characters in, a word x). This is done by constructing a *fixed length code*, which assigns a unique binary string to each character in A , of length $\log_2 k$. Note that this property is approximate if k is not a power of 2. To encode X we simply concatenate the binary strings assigned to the characters in X . The fixed length code allows us to place an upper bound on the code length needed to encode arbitrary words.

Now suppose we know in advance the frequencies of the symbols from A which appear in X (in the English language, "e" appears more frequently than "z"). Denote these frequencies P

$= (p_1, \dots, p_k)$, that is, symbol a_i occurs with frequency p_i . Then there exists an decodable coding scheme which approximately achieves a code length of $nH(P)$, where $H(P)$ is the entropy of P ,

$$H(P) = - \sum_{i=1}^k p_i \log_2 p_i$$

This may be done by assigning a unique binary string of length $-\log_2 p_i$ to symbol a_i . The Huffman code is commonly used to achieve this (Hamming 1986). Furthermore, this is the maximum possible compression under the stated conditions. We can then interpret $H(P)$ as the average number of bits per character needed to encode X .

Suppose we model X as a sample from a multinomial distribution on A . Anticipating the frequencies P we would choose to encode X using a Huffman code, which would yield a code length of

$$|b(X)| = - \sum_{i=1}^k n_i \log_2 p_i \quad (8)$$

where n_i is the frequency of symbol a_i in X . If P is correctly specified we would have $n_i/n \approx p_i$, hence

$$\begin{aligned} |b(X)| &= n \sum_{i=1}^k \frac{n_i}{n} \log_2 p_i \\ &\approx n \sum_{i=1}^k p_i \log_2 p_i \\ &= nH(P) \end{aligned}$$

so that $b(X)$ would achieve close to the minimum achievable code length. Note that (8) is exactly the negative log likelihood function for the multinomial model, that is:

$$\text{likelihood} = (1/2)^{\text{code length}} \quad (9)$$

Now suppose there is a model $\theta \in \Theta$ which explains X . One consequence of this is that knowledge of θ can be used to simplify X , which will result in a shorter code, say, $b(X | \theta)$. Of course, we will also need to code θ , say $b(\theta)$, so that X can be decoded. The complete code then becomes $b(X | \theta)b(\theta)$. We let $B(X | \theta)$ and $B(\theta)$ denote the code lengths of $b(X | \theta)$ and $b(\theta)$ respectively.

According to the MDL principle, we select as the model that θ which results in the smallest code length $B(\theta | X) = B(X | \theta) + B(\theta)$. Similar to the maximum likelihood principle, the MDL principle seeks the model which best explains the data. Note, however, that $b(\theta)$ will be longer when θ is more complex, which controls spurious complexity.

7.2 Coding of data based on model

We will now consider the problem of how to encode genotypic data from a pedigree. Suppose for the moment that X consists of a single locus of genotypic data for n individuals. We can encode X with a code length of approximately $N_I H(P^g)$ for N_I individuals, where P^g is the population genotype frequency distribution. For example, if there are 5 alleles of equal probability, under Hardy-Weiberg equilibrium there are 10 heterozygotes with frequency $2/25$ and 5 homozygotes with frequency $1/25$, resulting in entropy of

$$H(P^g) = 10 \times [- (2/25)\log_2 (2/25)] + 5 \times [- (1/25)\log_2 (1/25)] = 3.84,$$

so we need between 3 and 4 bits per genotype. But further compression can be achieved by exploiting model θ .

First consider the problem of coding genes inherited from two parents present in the sample. Suppose θ assigns individual i parents j and k . Given the parental genotypes, we can construct a set G_{jk} of genotypes (all genotypes resulting from a selection of 1 gene from each parent), which must include the offspring genotype (we would not consider a genetically excluded model θ). This set will range in size from 1 to 4. If it is of size 2, then the genotype of offspring i can be encoded in a single bit. This can be done by devising some canonical ordering of genotypes, and assigning a bit value of 0 to the lowest ranked genotype in G_{jk} . In general, if the size of G_{jk} is 1, 2 or 4, we can code the offspring genotype with $\log_2 |G_{jk}|$ bits. It happens that $|G_{jk}| = 3$ when both parents share the same heterozygous genotype ab . In this case, offspring genotypes ab , aa and bb may be assigned codes binary codes 0, 10, 11 respectively. This gives a direct link between code length and the likelihood function. When $|G_{jk}|$ is 1, 2 or 4, Mendel's rule assigns a likelihood of $1/|G_{jk}|$ to the offspring genotype. When $|G_{jk}| = 3$ Mendel's rule assigns a likelihood of $1/2$, $1/4$ and $1/4$ to offspring genotypes ab , aa and bb respectively. Thus, in each case we have the relationship

$$\text{likelihood of offspring genotype} = (1/2)^{\text{code length}}.$$

Now suppose individual i has a single parent j represented in sample X . At least one gene in i is represented in j . If both are (and i is heterozygous), we may use an arbitrary rule to designate one as having been inherited from j . If j is homozygous, we would assign a likelihood of 1 to that gene, and $1/2$ otherwise. Accordingly, we would need 0 or 1 bit to code a gene inherited from a homozygote or heterozygote, respectively, giving

$$\text{likelihood of offspring gene} = (1/2)^{\text{code length}}.$$

This leaves the problem of coding genes that are not inherited from other members of the sample. First, consider the genotypes belonging to founders. These occur in proportions approximately equal to the population genotype distribution hence each genotype can be coded with an average of $H(P^g)$ bits. Next, consider individuals with one parent in the sample. Coding of the gene inherited from the present parent has been discussed above. The remaining gene is essentially sampled from the population and will therefore occur in frequencies given by P^a , the population frequency distribution of the alleles. This gene can therefore be coded with an average of $H(P^a)$ bits.

For all genotypes, therefore, the relationship in (9) is preserved, and the code length is given by

$$B(X | \theta) = - \log_2 L(\theta | X) \quad (10)$$

where $L(\theta | X)$ is the likelihood of θ given data X . For multilocus data $X = (X(1), \dots, X(N_L))$ we add the code lengths attained for each locus, that is,

$$B(X | \theta) = - \sum_{j=1}^{N_L} \log_2 L(\theta | X(j)). \quad (11)$$

7.3 Coding pedigree structure

The method of coding the data given a model can be satisfactorily resolved, leaving the question of how best to code θ . While we can't claim a method which is provably optimal for any given criterion, we can develop a reasonable solution which will later be shown to possess desirable properties.

To do this we adapt an approach taken in Friedman and Goldszmidt (1998). For each individual we code the number of parents, one of 0, 1 or 2, which may be done using 2 bits. The number of bits is the same for each individual, so plays no role in model selection, and may be left out of our code length estimate. We then code the parent subset. Here, we assume that age and sex data (or any form of exclusionary information) is available in the form of Y . Let $n(k, i)$ represent the number of possible parent sets of size k of individual i permitted by Y . If Y is empty, we would have $n(1, i) = N_I - 1$ and $n(2, i) = (N_I - 1)(N_I - 2)/2$. We necessarily have $n(0, i) = 1$. If the number of parents is k we would require $\log_2 n(k, i)$ bits to code the parent set, using a fixed length code. Note that we need 0 bits to encode founders. We then have

$$B(\theta) = \sum_{i=1}^{N_I} \log_2 \left(n(k_i | H_i^\theta, i) \right). \quad (12)$$

Figure 1.
Test base pedigrees

Figure 2.

Summary of experiments **E1** for 10 loci simulations. Left column contains plots of d^a (○) and d^b (Δ) for pedigree types *A*, *B*. Right column contains plots of $RMSE^a$ (○) and $RMSE^b$ (Δ) for pedigree types *A*, *B*. MDL score is shown in bold lines, MLE score is shown in dotted lines. True values d^a , d^b are indicated by arrows.

Figure 3.

Estimated posterior densities of individual descendent numbers for 20 founders (experiment **E2**, $N_I = 10000$, 8 loci). Two founders are selected from each of 20 base pedigree replications, one from each fitness group. Each score and pedigree is separately represented.

Figure 4.

Estimated posterior densities of individual descendent numbers for 20 founders (experiment E2, $N_I = 10000$, 10 loci). Two founders are selected from each of 20 base pedigree replications, one from each fitness group. Each score and pedigree is separately represented.

Figure 5.

Estimated posterior densities of individual descendent numbers for 20 founders (experiment E2, $N_I = 10000$, 12 loci). Two founders are selected from each of 20 base pedigree replications, one from each fitness group. Each score and pedigree is separately represented.

Figure 6. Histograms of values of nonzero α_i , the posterior probabilities of parenthood, among n_t true parents and n_f spurious parents, for likelihood-based and MDL-based posterior densities.

Figure 7.

Threshold posterior estimates of number of parents, categorized as true or false. This is given as the total number of n_t and n_f actual and spurious parents which are estimated as parents by applying a threshold rule of the form $\alpha_i \geq p$, allowing p to vary from .04 to 1. Both likelihood-based and MDL-based posterior densities are employed. The maximum score estimate is also indicated.

Figure 8. Histograms of values of nonzero β_{ij} , the posterior probabilities of parent-offspring relationship, given by true pairs among n_t actual parents; false pairs among n_t actual parents; and false pairs among n_f spurious parents, for likelihood-based and MDL-based posterior densities.

Figure 9.

Threshold posterior estimates of number of parent-offspring pairs, categorized as true or false. This is given as the total number of pairs which are estimated as parent-offspring pairs by applying a threshold rule of the form $\beta_{ij} \geq p$, allowing p to vary from .04 to 1. Both likelihood-based and MDL-based posterior densities are employed. The maximum score estimate is also indicated.

Table 1

Base pedigree characteristics.

	Founders	Fitness	Founder Descendants	Total Individuals
<i>A1</i>	8	4	84	120
<i>A2</i>	8	2	14	32
$A = A1 \cup A2$	16	-	-	152
<i>B1</i>	2	3	39	53
<i>B2</i>	2	2	14	22
$B = B1 \cup B2$	4	-	-	75

Table 2Summary of results (6) and (7) for experiments **E1** and **E2** applied to 8 test pedigrees using 8 loci.

8 Loci			MLE				MDL			
Type/ Exp	Quantity	d^a/d^b	200	500	$N_I =$ 2000	10000	200	500	$N_I =$ 2000	10000
A E1	$\hat{d}^a$	84	83.57	83.28	81.67	75.56	83.56	83.11	78.22	48.61
	$\hat{d}^b$	14	14.07	14.38	16.06	22.27	14.06	14.36	15.18	12.16
	$RMSE^a$	84	2.21	4.93	11.64	27.66	2.23	5.03	12.86	40.37
	$RMSE^b$	14	1.44	2.93	7.26	19.17	1.44	2.93	6.41	8.33
	$\bar{D}$	-	14.15	48.55	348.37	4332.98	14.16	49.33	403.03	6255.24
B E1	$\hat{d}^a$	39	41.19	45.92	57.35	72.65	40.32	41.64	41.54	28.95
	$\hat{d}^b$	14	16.27	20.41	31.04	48.16	15.37	16.70	18.12	14.59
	$RMSE^a$	39	4.48	11.19	25.48	44.62	3.35	5.84	8.47	14.42
	$RMSE^b$	14	5.05	10.91	24.20	43.84	3.70	5.69	8.17	7.51
	$\bar{D}$	-	9.00	52.36	537.38	6421.52	7.03	33.95	388.21	6225.14
A E2	$\hat{d}^a$	84	83.86	83.41	81.04	74.58	82.32	79.70	68.33	43.16
	$\hat{d}^b$	14	13.99	14.21	16.69	23.17	13.90	13.53	13.96	13.29
	$\sim^a$	84	1.03	1.69	6.91	14.86	3.03	4.68	10.07	12.51
	$SD \sim^b$	14	0.25	0.92	4.26	9.99	0.70	1.86	4.79	7.35
	$SD \sim^a$	84	0.53	0.60	0.67	0.72	0.81	0.93	0.97	1.00
	$QP \sim^b$	14	0.50	0.50	0.28	0.20	0.54	0.55	0.60	0.63
	$QP \sim^a$	14	0.50	0.50	0.28	0.20	0.54	0.55	0.60	0.63
	$\bar{D}$	-	10.76	60.70	387.54	5061.43	12.78	77.46	627.02	7726.81
	$\bar{D}$	-	10.76	60.70	387.54	5061.43	12.78	77.46	627.02	7726.81
B E2	$\hat{d}^a$	39	46.58	42.43	56.48	73.08	41.75	38.23	39.36	32.15
	$\hat{d}^b$	14	15.89	20.14	33.76	48.50	15.28	16.21	18.89	18.28
	$\sim^a$	39	2.15	3.23	7.23	16.85	3.48	4.20	6.38	10.21
	$SD \sim^b$	14	0.54	3.16	7.94	15.39	1.22	3.46	6.24	8.34
	$SD \sim^a$	39	0.01	0.20	0.00	0.01	0.18	0.67	0.53	0.78
	$QP \sim^b$	14	0.50	0.06	0.00	0.00	0.55	0.38	0.31	0.37
	$QP \sim^a$	14	0.50	0.06	0.00	0.00	0.55	0.38	0.31	0.37
	$\bar{D}$	-	17.08	56.60	636.32	7029.86	16.59	63.42	699.16	7864.75
	$\bar{D}$	-	17.08	56.60	636.32	7029.86	16.59	63.42	699.16	7864.75

Table 3Summary of results (6) and (7) for experiments **E1** and **E2** applied to 8 test pedigrees using 10 loci.

10 Loci			MLE				MDL			
Type/ Exp	Quantity	d^a/d^b	$N_I =$				$N_I =$			
			200	500	2000	10000	200	500	2000	10000
A E1	$\hat{d}^a$	84	83.94	83.89	83.64	82.43	83.94	83.88	83.49	79.31
	$\hat{d}^b$	14	14.00	14.07	14.30	15.52	14.00	14.06	14.28	14.77
	$RMSE^a$	84	0.80	1.78	4.02	9.62	0.80	1.79	4.11	10.88
	$RMSE^b$	14	0.46	1.14	2.36	5.84	0.46	1.14	2.34	5.13
	$\bar{D}$	-	5.66	17.97	99.76	1039.27	5.66	18.00	102.78	1285.95
B E1	$\hat{d}^a$	39	39.81	41.72	48.33	60.59	39.57	40.39	41.81	41.64
	$\hat{d}^b$	14	14.77	16.30	22.03	33.24	14.57	15.21	16.57	17.74
	$RMSE^a$	39	2.38	5.50	14.66	29.04	1.98	3.50	5.87	7.95
	$RMSE^b$	14	2.60	5.22	13.45	26.77	2.16	3.32	5.46	7.50
	$\bar{D}$	-	3.23	18.64	212.45	2552.43	2.70	12.13	103.71	1535.95
A E2	$\hat{d}^a$	84	83.73	83.91	83.11	81.91	83.52	83.61	80.34	71.76
	$\hat{d}^b$	14	14.00	14.08	14.82	16.04	13.95	13.99	14.53	14.10
	$\sim a$	84	0.54	0.22	2.76	5.65	1.11	1.17	4.61	8.37
	$SD \sim b$	14	0.08	0.13	1.05	3.14	0.34	0.46	1.54	3.79
	$SD \sim a$	84	0.63	0.50	0.60	0.62	0.66	0.61	0.85	0.95
	$QP \sim b$	14	0.50	0.50	0.50	0.34	0.52	0.50	0.51	0.58
	$QP \sim a$	14	0.50	0.50	0.50	0.34	0.52	0.50	0.51	0.58
	$\bar{D}$	-	8.17	12.17	117.66	1274.07	8.44	13.94	172.05	2293.95
B E2	$\hat{d}^a$	39	38.99	43.02	48.66	60.87	38.84	40.49	42.00	41.14
	$\hat{d}^b$	14	16.30	15.24	20.01	32.11	15.97	14.83	16.07	18.33
	$\sim a$	39	0.22	1.46	3.59	8.51	0.70	1.89	3.58	6.84
	$SD \sim b$	14	0.56	1.08	2.20	7.60	1.04	1.22	2.72	5.37
	$SD \sim a$	39	0.50	0.25	0.00	0.00	0.55	0.38	0.36	0.42
	$QP \sim b$	14	0.10	0.35	0.04	0.00	0.15	0.45	0.42	0.29
	$QP \sim a$	14	0.10	0.35	0.04	0.00	0.15	0.45	0.42	0.29
	$\bar{D}$	-	8.00	22.68	190.05	2671.88	7.70	18.01	158.77	2665.30

Table 4Summary of results (6) and (7) for experiments **E1** and **E2** applied to 8 test pedigrees using 12 loci.

12 Loci			MLE				MDL			
Type/ Exp	Quantity	d^a/d^b	$N_I =$				$N_I =$			
			200	500	2000	10000	200	500	2000	10000
A E1	$\tilde{d}^a$	84	83.99	83.98	83.95	83.78	83.99	83.98	83.94	83.63
	$\tilde{d}^b$	14	14.00	14.01	14.04	14.21	14.00	14.01	14.04	14.18
	$RMSE^a$	84	0.17	0.69	1.51	3.34	0.17	0.69	1.52	3.44
	$RMSE^b$	14	0.08	0.35	0.87	1.97	0.08	0.35	0.87	1.95
	$\tilde{D}$	-	2.45	7.63	34.75	250.28	2.45	7.63	34.86	264.39
B E1	$\tilde{d}^a$	39	39.28	40.03	42.88	51.27	39.22	39.61	40.55	41.98
	$\tilde{d}^b$	14	14.26	14.83	17.20	24.28	14.21	14.50	15.28	16.59
	$RMSE^a$	39	1.37	2.95	7.47	18.67	1.25	2.13	3.77	5.98
	$RMSE^b$	14	1.32	2.73	6.91	16.81	1.21	2.02	3.47	5.48
	$\tilde{D}$	-	1.21	7.00	85.11	1183.67	1.08	4.92	41.33	443.28
A E2	$\tilde{d}^a$	84	84.00	84.00	83.74	83.72	83.95	83.93	83.37	81.84
	$\tilde{d}^b$	14	14.00	14.00	14.22	14.27	14.00	13.99	14.18	14.01
	$\sim^a$	84	0.12	0.06	0.55	1.38	0.41	0.49	1.42	3.36
	$SD \sim^b$	14	0.04	0.01	0.22	0.55	0.05	0.12	0.45	1.19
	$SD \sim^a$	84	0.50	0.50	0.50	0.50	0.51	0.52	0.60	0.75
	$QP \sim^b$	14	0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.52
	$QP \sim^a$	14	0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.52
	$\tilde{D}$	-	1.55	5.94	51.13	312.29	1.58	6.33	57.81	501.35
B E2	$\tilde{d}^a$	39	39.00	39.66	43.40	52.36	38.99	39.56	41.32	42.75
	$\tilde{d}^b$	14	14.00	14.00	16.26	25.21	14.00	13.99	14.68	17.43
	$\sim^a$	39	0.01	0.05	0.76	3.95	0.09	0.40	1.50	4.06
	$SD \sim^b$	14	0.04	0.00	0.34	3.43	0.05	0.12	0.84	3.00
	$SD \sim^a$	39	0.50	0.50	0.01	0.00	0.51	0.50	0.23	0.26
	$QP \sim^b$	14	0.50	0.50	0.50	0.00	0.50	0.50	0.50	0.30
	$QP \sim^a$	14	0.50	0.50	0.50	0.00	0.50	0.50	0.50	0.30
	$\tilde{D}$	-	0.18	2.43	86.18	1239.54	0.25	2.70	57.20	787.11