Available online at www.sciencedirect.com
ScienceDirect
Theoretical Population Biology 71 (2007) 213–229
Theoretical Population Biology
www.elsevier.com/locate/tpb

A graphical approach to relatedness inference

Anthony Almudevar*

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA

Received 11 April 2006
Available online 27 October 2006

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 (i.e. family trees). Additionally, interest may be confined to a particular relationship type, in particular, parent–offspring relationships (e.g. Thompson and Meagher, 1987; Marshall et al., 1998; Nielsen et al., 2001) or sibling relationships (e.g. Almudevar and Field, 1999; Thomas and Hill, 2002; Butler et al., 2004; Wang, 2004; Konovalov et al., 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

*Fax: +1 585 273 1031.
E-mail address: anthony_almudevar@urmc.rochester.edu.

0040-5809/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2006.10.005

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 et al., 2004). Alternative scores, such as the Simpson's Index proposed in Butler et al. (2004) (see also Konovalov et al., 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/people/faculty/almudevar.html`.

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 $1, \ldots, N_I$), and an edge $j \to i \in E$ implies a parent–offspring relationship for j and i, respectively. Note that θ must be a directed acyclic graph (DAG) with no more than two 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 used to test the admissibility of a mating relationship or a parent–offspring relationship. Denote the complete data sets $X = (X_1, \ldots, X_{N_I})$ and $Y = (Y_1, \ldots, Y_{N_I})$. The data are 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 $1, \ldots, N_I$ 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