BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 2 2002
Pages 341–342

PATRI—paternity inference using genetic data

J. Signorovitch* and R. Nielsen

Department of Biometrics, Cornell University, Ithaca, NY 14853-7801, USA

Received on June 15, 2001; revised on July 17, 2001; accepted on August 20, 2001

ABSTRACT

Summary: PATRI is a new application for paternity analysis using genetic data that accounts for the sampling fraction of potential fathers.

Availability: Executables are available at www.biom.cornell.edu/Homepages/Rasmus_Nielsen/

Contact: jes48@cornell.edu

PaTeRnity Inference (PATRI) is a new application written in C for paternity analysis of genetic data. Paternity analysis is used extensively in molecular evolution, molecular ecology and in forensic science. The program requires genotypic, diploid data from one or more loci from mother–offspring pairs and from potential fathers. Typical data might include microsatellite markers, Restriction Fragment Length Polymorphisms (RFLPs) or Single Nucleotide Polymorphisms (SNPs). Given such genotypic data, and information about the male population size, PATRI can calculate posterior probabilities of paternity for all sampled offspring. When behavioral or ecological information can be used to divide the sampled males into different groups, PATRI can perform maximum likelihood analyses of hypotheses regarding the relative reproductive success of those groups.

The underlying statistical methodology was described in Nielsen *et al.* (2001). In brief, the method is based on calculation of the posterior probability that a male sired (fathered) a particular offspring and it is related to the fractional likelihood methods by Devlin *et al.* (1988); Roeder *et al.* (1989) and Smouse and Meagher (1994). As in previous methods, Hardy–Weinberg equilibrium and linkage equilibrium in the population is assumed (see, e.g. Weir, 1996, pp. 92 and 112). The approach by Nielsen *et al.* (2001) differs from previous approaches in that it can appropriately take into account incomplete sampling of potential fathers. Since PATRI estimates the allele frequencies from the sampled individuals and assumes Hardy–Weinberg equlibrium, PATRI is most suitable for large data sets in molecular ecology, and is not directly applicable to forensic science.

Paternity analyses in natural populations can involve hundreds of samples. For clarity, PATRI allows the user to preserve existing labels for individual samples. PATRI also checks the data for potential errors that could affect results, such as offspring genotypes that are incompatible with maternal genotypes. Once the genetic data are read, a simple interactive command-line interface guides the user through the data analysis.

For all data sets, PATRI can estimate the posterior probability that a particular male has sired a particular offspring, assuming a uniform prior among all males in the population. The male population size (N) can either be specified by the user as a fixed value, or uncertainty regarding N can be modeled using a uniform or Gaussian prior. Using a uniform prior corresponds to assuming no prior information regarding the male population size, except that an upper bound can be specified.

Posterior probabilities of paternity are printed to a table, along with the expected number of sampled offspring sired by sampled fathers, which gives an intuitive measure of power. PATRI can also produce a maximum likelihood estimate of N based solely on the parent–offspring genotypic data. The estimation of N assumes equal fecundity and unbiased sampling of males.

If sampled males can be divided into groups based on behavioral or ecological information, PATRI can be used to evaluate hypotheses regarding the relative reproductive success of these groups. For k groups the user starts with a full model containing $k-1$ parameters, $\alpha_2, \alpha_3, \ldots, \alpha_k$, where α_i, is defined as the reproductive success of group i relative to group 1. The user can then enter restrictions on these parameters. For example, the hypothesis that males from groups i and j have equal reproductive success corresponds to the restriction $\alpha_i = \alpha_j$. Given a set of restrictions, PATRI can (1) maximize the likelihood, and (2) plot a profile likelihood surface for any particular α_i. The profile likelihood surface for α_i is constructed by optimizing over all α_j, $j \neq i$ (sample output is shown in Figure 1). The maximum likelihood values are stored in a table, allowing the user to perform likelihood ratio tests of various hypotheses regarding reproductive success. This analysis can be done using a fixed value of N or by assuming N is uniformly or Gaussian distributed. All optimization is done using a quasi-Newton algorithm (Davidon–Fletcher–Powell) and numerical integration is completed using an adaptive quadrature algorithm.

*To whom correspondence should be addressed.

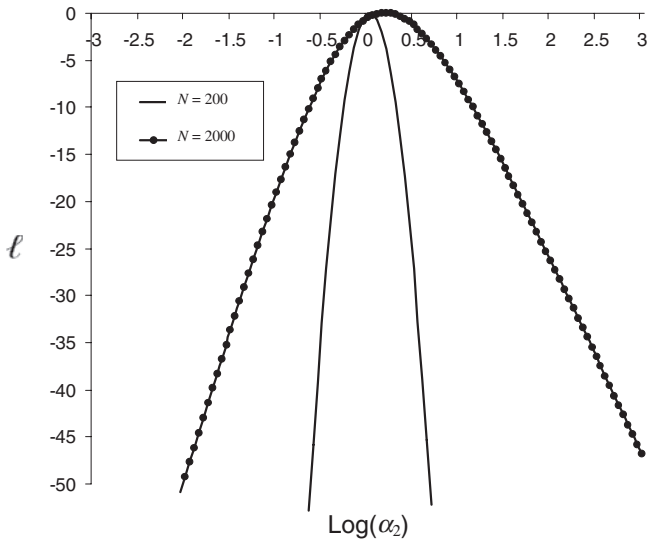

Fig. 1. The likelihood surface for α_2 for two simulated data sets. The true value of α_2 is 1.0.

The following example illustrates how PATRI can be used to evaluate the hypothesis that one group of males has higher reproductive success than another. Two data sets were created by simulating diploid genotypes for 200 females at 10 loci and $N = 200$ males (100 in each of two groups) and $N = 2000$ males (1000 in each of two groups), respectively. Linkage equilibrium, Hardy–Weinberg equilibrium and 10 alleles at equal frequencies at each locus were assumed. For each female, a mate was chosen according to his probability of siring an offspring: $1/(\alpha_2+1)$ for males in group 1 and $\alpha_2/(\alpha_2+1)$ for males in group 2. We set $\alpha_2 = 1$. Offspring genotypes were then sampled from their parents assuming Mendelian segregation and independence among loci. Finally, for each data set, 100 male genotypes were sampled from each group of potential fathers, giving sampling fractions of 1 and $1/10$ for the two data sets, respectively.

Given these data sets, PATRI found maximum likelihood estimates of $\alpha_2 \approx 1.0$ and $\alpha_2 \approx 1.6$, respectively using a uniform prior for the male population size with an upper bound at 10 000. PATRI also calculated the likelihood surfaces for α_2 shown in Figure 1. Assuming that the likelihood ratio test statistic follows a χ^2 distribution, approximate 95% confidence intervals for α_2 for the two data sets are (079, 1.41) and (0.66, 3.81). The two values are centered around 1 and the (true) hypothesis of $\alpha_2 = 1.0$ cannot be rejected.

ACKNOWLEDGEMENT

This research was supported by NSF grant DEB-0089487.

REFERENCES

Devlin,B., Roeder,K. and Ellstrand,N.C. (1988) Fractional paternity assignment—theoretical development and comparison to other methods. *Theor. Appl. Genet.*, **76**, 369–380.

Nielsen,R., Mattila,D.K., Clapham,P.J. and Palsbøll,P.J. (2001) Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic Humpback Whale. *Genetics*, **157**, 1673–1682.

Roeder,K., Devlin,B. and Lindsay,B.G. (1989) Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, **45**, 363–379.

Smouse,P.E. and Meagher,T.R. (1994) Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae). *Genetics*, **136**, 313–322.

Weir,B.S. (1996) *Genetic Data Analysis II*. Sinauer, Sunderland, MA.