Comment on ‘Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching’

Eric C. Anderson\textsuperscript{1,2,*} and Thomas C. Ng\textsuperscript{3}

\textsuperscript{1}Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 110 Shaffer Road, Santa Cruz, CA 95060, \textsuperscript{2}Department of Applied Math and Statistics and \textsuperscript{3}Department of Biomedical Engineering, University of California, Santa Cruz, CA, USA

Associate Editor: Jeffrey Barrett

\textbf{Summary:} We show the software SOLOMON is improved by using the likelihood ratio instead of an ad hoc statistic.

\textbf{Code:} github.com/ericande/solomon/releases/tag/v0.1-bioinformatics

\textbf{Contact:} eric.anderson@noaa.gov

Received on August 7, 2013; revised on September 23, 2013; accepted on October 4, 2013

1 INTRODUCTION

In a recent \textit{Bioinformatics} article, Christie, Tennessen and Blouin (hereafter CTB) present the software SOLOMON for parentage inference from genotype data (Christie et al., 2013). They propose a method of estimating the number of true parent-offspring pairs amongst all the candidate pairs in a dataset and show it provides better results than the program CERVUS (Marshall et al., 1998) when the fraction of sampled candidates (an input to CERVUS) is misspecified.

Here we show that the performance of SOLOMON can be improved by making full use of the genotype data, replacing CTB’s statistic with the likelihood ratio as prescribed by a well-established literature (Marshall et al., 1998; Meagher and Thompson, 1987). This is particularly beneficial when using single nucleotide polymorphisms (SNPs) chosen to have minor allele frequency near 0.5, as desired for SNP-based pedigree reconstruction (Anderson and Garza, 2006).

2 METHODS

We start with an overview of SOLOMON’s methodology, adopting new notation where it permits a more complete succinct description.

2.1 SOLOMON’s methodology

Assume $L$ linked loci used to identify the true parents of $n_2$ diploid offspring from amongst $n_1$ diploid candidate parents. CTB focus first on the case where all candidate parents are of the same sex and the goal is to identify true parent–offspring pairs from amongst the $n_1 n_2$ candidate pairs. Let $z_j$ count the number of loci not compatible with Mendelian inheritance (barring a genotyping error) between the candidate parent and offspring of pair $j$. The method of CTB then proceeds in four steps.

\textbf{Step 1:} Partition all $n_1 n_2$ pairs into disjoint sets $P_0, \ldots, P_L$, where $P_i$ includes all pairs $j$ such that $z_j = i$. Denote by $\#(P_i)$ the cardinality of $P_i$.

\textbf{Step 2:} Estimate $\pi_i$, the unknown fraction of pairs in each set $P_i$ that are unrelated pairs (i.e. the two members of the pair are unrelated), with

$$\hat{\pi}_i = \min\left\{1, \frac{n_1 n_2 P(z_j = i) | \text{pair } j \text{ is unrelated}}{\#(P_i)}\right\}$$

\textbf{Step 3:} Calculate for each pair $j$ in $P_i$ with $\pi_i < 1$ the quantities:

$$T_{j}^{\text{U}} = \prod_{i < c} P(\lambda(y_{i,c}, y_{j,c}) | R_{ik} = \text{U})$$

$$T_{j}^{\text{PO}} = \prod_{i < c} f(\lambda(y_{i,c}, y_{j,c}))$$

\textbf{Step 4:} Combine $\hat{\pi}_i$ and $\lambda$ to compute a ‘posterior probability’ of parentage for each pair in each $P_i$ with $\pi_i < 1$. Let $X_j^{\text{U}}$ represent a random variable defined as the product in (2) when $s$ and $k$ are unrelated individuals simulated conditional on sharing at least one allele at every locus, and $L$ is a randomly chosen set of $L - i$ loci at which $s$ and $k$ share alleles. Let $X_j^{\text{PO}}$ be a random variable that is the product in (3) when $s$ and $k$ are a simulated true parent and offspring. CTB define the ‘posterior probability of parentage’ for a pair $j$ in $P_i$, with observed values of $T_{j}^{\text{U}}$ and $T_{j}^{\text{PO}}$, as:

$$Q(\text{PO}|\lambda) = \frac{(1 - \pi_i) P(X_j^{\text{PO}} \geq T_{j}^{\text{PO}})}{\pi_i P(X_j^{\text{U}} \geq T_{j}^{\text{U}}) + (1 - \pi_i) P(X_j^{\text{PO}} \geq T_{j}^{\text{PO}})}.$$
that half of the pairs in \( Q \)

Confidence in those assignments can still be assessed with a

loci with equifrequent alleles; and (ii) using the default 1000 replicates, requires 16.6 and 225 min. 

\[
\text{swets, 1966) for each dataset.}
\]

plotting the receiver-operating characteristic (ROC) curve (Green and

approach and SOLOMON, and accuracy was compared graphically by

SOLOMON is not making full use of the data; \( T^0_P \) and \( T^1_P \) are not

sufficient statistics.

An example demonstrates this. Consider \( L = 60 \) biallelic loci, each with

one allele at frequency of 0.501 and the other at 0.499. Clearly, little

information is gained by knowing whether the first or the second allele

is shared in a candidate parent–offspring pair, so, in effect, the statistic

\( \lambda(y_{x, r}, y_{x, l}) \) is merely recording whether or not at least one allele is shared.

An unrelated pair shares at least one allele at every locus with probability

close to \( (1 - \frac{1}{4})^L = 3.3 \times 10^{-4} \). Imagine that \( n_1 = n_2 = 500 \) and

\( \#|P_0| = 165 \). By (1) we have \( \pi_0 = (3.3 \times 10^{-4} \times 50^2)/165 = 0.5 \), so we

expect that half of the pairs in \( P_0 \) are true parent–offspring pairs.

Those true pairs cannot, however, be identified by CTB’s method because

the observed values \( T^0_P \) and \( T^1_P \) are effectively identical for every pair \( j \) in

\( P_0 \), whether it is a true parent–offspring pair or not.

We implemented CTB’s approach, but based it on the likelihood ratio:

\[
LR(y_{x, r}, y_{x, l}) = \log \frac{\prod_{l \in c} P(y_{x, r}, y_{x, l} | R_i = PO)}{\prod_{l \in c} P(y_{x, r}, y_{x, l} | R_i = U)}
\]

Kids are assigned to the candidate sire with highest \( LR(y_{x, r}, y_{x, l}) \).

Confidence in those assignments can still be assessed with a \( P \)-value, \( Q^*(PO)_{y_{x, r}, y_{x, l}} \), like that in (4), by substituting \( LR(y_{x, r}, y_{x, l}) \) for both \( T^0_P \) and \( T^1_P \) in (4) and redefining \( X^0_k \) and \( X^1_k \) to be the random variable

\( LR(y_{x, r}, y_{x, l}) \) for an \( s \) and \( k \) drawn randomly from the population conditional on \( s \) and \( k \) being either unrelated or parental, respectively, and

sharing at least one allele at a randomly determined \( L - i \) loci. We call this

the ‘likelihood ratio (LR) approach’.

To compare the performance of the LR approach to SOLOMON, we

simulated 20 datasets under two different scenarios with no genotyping

error: (i) \( n_1 = n_2 = 500 \), number of true pairs = 83 and \( L = 60 \) biallelic

loci with equifrequent alleles; and (ii) \( n_1 = n_2 = 200 \) with 50 true pairs,

and \( L = 10 \) loci each with 10 alleles and with the frequency of allele

\( v = 1, \ldots, 10 \), proportional to \( 1/v \). \( \pi_0 \) is expected to be 0.50 and 0.48

for the scenarios, respectively. Each dataset was analyzed using the LR

approach and SOLOMON, and accuracy was compared graphically by

plotting the receiver-operating characteristic (ROC) curve (Green and

Swets, 1966) for each dataset.

3 RESULTS AND CONCLUSIONS

Figure 1a shows that, as predicted, SOLOMON’s criterion for

parentage contains no extra information beyond Mendelian incompatibility in data scenario 1, with equifrequent alleles. The

ROC curves indicate SOLOMON does not rank true pairs any

higher than false pairs in \( P_0 \). The LR approach performs considerably better. Figure 1b shows that on data with 10 alleles at

frequencies that are far from equal (scenario 2), the LR approach

still outperforms SOLOMON, as expected.

Our implementation of the LR approach, using \( 10^5 \) simulation

replicates to approximate \( Q^*(PO)_{y_{x, r}, y_{x, l}} \), requires roughly 0.1 and

1 min for each dataset from scenarios 1 and 2, respectively, on a

single core from a Mac Pro running at 2.8 GHz. SOLOMON,

using the default 1000 replicates, requires 16.6 and 225 min.

ACKNOWLEDGEMENTS

The authors acknowledge helpful comments from Mark Christie

two anonymous referees.

Funding: T.N. was partially supported with funds from the US

Section of the Chinook Technical Committee of the

International Pacific Salmon Commission.

Conflict of Interest: none declared.

REFERENCES


large-scale parentage inference. Genetics, 172, 2567-2582.


