Special Feature for Ecological Risk Assessment of Introduced Bumblebees

Full-sib reconstruction in haplodiploid populations

Yukihiro TOQUENAGA1,* and Nozomu KOKUVO2

1 Division of Integrative Environmental Sciences, Graduate School of Life and Environmental Sciences, University of Tsukuba; Ibaraki 305–8572, Japan
2 Science and Technology Systems Department, MRI Research Associates, Inc.; Tokyo 101–0047, Japan

(Received 30 September 2008; Accepted 20 December 2008)

Abstract

Estimating colony numbers of invading Bombus terrestris in Japan is the most urgent task for eradicating the alien invasive species. Nests of bumble bees are hardly found in nature. A prospective alternative is estimating the number of nests with polymorphic genetic data of individuals foraging in the field. Full-sib reconstruction from such genetic data is NP-hard in a strict sense, but several heuristic methods have been proposed. Among them, likelihood methods have been often applied for reconstructing full-sibs in nature, but they require relatively large sample sizes for obtaining accuracy in allele frequencies. Alternative methods calculate genetic distances between paired individuals and cluster them into family groups, but there are also pros and cons for applying these methods against real data. We proposed another heuristic pairing and clustering method, Shared Loci Correspondence Analysis (SLCA), for reconstructing full-sib families of haplodiploid populations. In SLCA, genetic similarities among individuals are numbers of loci shared by the paired individuals. The similarity scores are sorted by the correspondence analysis for clustering them into full-sib families. SLCA could be applicable to quite small samples and succeeded to correctly estimate the number of bumble bee nests in the field. We also improved estimation accuracy of SLCA by blunting the similarity index among individuals. The modified SLCA performed almost perfectly against challenging problems when numbers of loci and alleles were equal or greater than eight. Future direction of improving SLCA is discussed.

Key words: Bombus terrestris; alien species; nest; haploid; Shared Loci Correspondence Analysis

INTRODUCTION

Estimating population size is the crucial first step for evaluating current states and future expansion rates of invasive alien species. In the case of Bombus terrestris intruding into Japan, simply counting the number of foragers in the field is an erroneous indicator because the effective population size of social hymenopteran species should be defined as the number of colonies. However, it is very difficult to find underground nests of bumble bees, and we can hardly count the number of colonies directly. Recent progress in molecular biology enables us to distinguish each foraging individual in the field by genetic markers, such as the DNA microsatellite. Based on the genetic polymorphic data, we can reconstruct sibships among them and estimate the number of colonies from which they come to forage in the field.

Strict full-sib reconstruction is NP-hard (Berger-Wolf et al., 2005, 2007) and heuristic methods are required for practical problems. Heuristic reconstruction methods can be roughly divided into two categories: using and without using maximum likelihood. The former ones need relatively large sample size for obtaining reliable allele frequencies. The latter ones calculate genetic distances among individuals and cluster them into groups according to the distance measurements. When we evaluate the population size of B. terrestris in Japan, we cannot always obtain sufficient sample sizes, especially at the frontline of invasion. What we need is an alternative method that can be applicable for data with small sample size. The method would be
even suitable for estimating colony numbers of endangered indigenous species that would be competitively excluded by *B. terrestris*.

In this article, we first review the previous methods for full-sib reconstruction. We start from full-sib reconstruction in a strict sense, and then review methods using likelihood measurements for understanding the fundamental structure of the sibship reconstruction. Then we review sibship reconstruction methods without using likelihood measurements, and also discuss pros and cons for those approaches.

Following the reviews of the previous studies, we introduce our new method for reconstructing full-sib relationships among haplodiploid conspecific individuals (Kokuvo et al., 2007a, b). Our method consists of two steps: assigning pairwise distance measures for constructing an individual-distance matrix and clustering individuals using the correspondence analysis. We evaluate our methods by reconstructing families from randomly generated genetic data as well as data of feral bumble bee nests. We show that our method can correctly estimate the number of nests of data for which some previous methods assign erroneous estimates. Finally we show a couple of applications of our method, such as estimating the number of colonies of *B. terrestris* in Biratori, Hokkaido (Kokuvo et al., 2008).

**FULL-SIB RECONSTRUCTION METHODS**

**Full-sib reconstruction is NP-hard**

Recently Burger-Wolf et al. (2005, 2007) formalized the full-sib reconstruction in terms of 2-allele and 4-allele properties. The two properties are derived from Mendelian genetics where an offspring inherits one allele from each of its parents for each locus. The 4-allele property is straightforward; the total number of distinct alleles occurring at any locus may not exceed four. The 2-allele property states that there exists an assignment of individual alleles within a locus to maternal and paternal such that the number of distinct alleles assigned to each parent at this locus does not exceed two. This is also straightforward because each parent can pass at most two different alleles to its offspring at each locus. The 2-allele property is stricter than the 4-allele property.

Full-sib reconstruction is then assigning individual alleles parsimoniously into the smallest number of (possibly overlapping) groups that satisfy the necessary 2-allele constraint. Let *R* be the number of alleles that are homozygous or appear with three other distinct alleles in a given locus. Let *A* be the total number of distinct alleles at a locus. Then a set of individuals satisfies the 2-allele property if and only if for every locus it satisfies the constraint *A* + *R* ≤ 4. However, obtaining full-sib groups satisfying this 2-allele property is NP-hard and massive computational power is needed. Practically we need some heuristic methods to reconstruct full-sib families instead of tackling the NP-hard problem of finding minimum 2-allele sets.

**Heuristic approaches**

Several methods have been proposed to assign single-generation individuals into full-sib families based on their genetic data (Painter, 1997; Almudevar and Field, 1999; Thomas and Hill, 2000; Smith et al., 2001; Beyer and May, 2003). Various approaches developed thus far can be characterized by how all the possible partitions are searched and how these partitions are ranked to determine the best-fit partition (Konovalov et al., 2005).

Some methods are based on maximizing some form of the likelihood of proposed full-sib configurations (Thomas and Hill, 2000; Smith et al., 2001; Wang, 2004). The likelihood might be represented as Eq. (1) (Thomas and Hill, 2000).

\[
L_f = \prod_i \left[ \sum_{w=1}^{h_i} \sum_{x=1}^{h_i} \sum_{y=1}^{h_i} \sum_{z=1}^{h_i} \right] \times \left[ P_{wx}^{p_{xy}} P_{yz}^{p_{yz}} L(g_{lb}) \right]
\]

where \( L_f \) is the likelihood of the observed genotypes within a hypothetical diploid full-sib family. \( L_f \) consists of products of the four-allele combinations of a pair of individuals. Each product consists of the probability of given parent genotype combination \((p_{wx}^{1}, p_{xw}^{2})\) and the genotype configuration among family members \( (\prod_{i=1}^{n} L(g_{lb})) \). One has to seek for the best \( N \) partition of families, which maximizes the likelihood equation \( (L) \). However, maximizing \( L \) is a time consuming process, and the accuracy of estimates strongly depends on the accuracy of gene frequencies, i.e. \( p_{wx}^{1} \) and \( p_{xw}^{2} \).

Smith et al. (2001) used a Markov chain Monte Carlo method for pairwise scoring (denoted as SC)
and for estimating likelihood (denoted as Likelihood). Other methods are based on the likelihood ratio of null and hypothesized kin relationships (Goodnight and Queller, 1999) and cluster the relationships by various methods (Beyer and May, 2003; Darvill et al., 2004; Konovalov et al., 2004; Knight et al., 2005). For example, Beyer and May (2003) used a graph-based algorithm (denoted as Graph), and Konovalov et al. (2004) used descending ratio (denoted as DR) search algorithms.

Alternative approach is to estimate pairwise relatedness between all individuals and graphically cluster the pairs of individuals according to the similarity among them (Blouin et al., 1996). The approach of Almudevar and Field (1999) (denoted as AF), based on the Mendelian exclusion principle, constructs and evaluates sets of feasible sibling groups. The method of Butler et al. (2004) (denoted as Simpson), based on a simple exclusion method, chooses the best partition according to the Simpson index.

Butler et al. (2004) conducted a large-scale simulation study and compared the AF, Simpson, SC, and Likelihood approaches at various parameter settings. Their results indicated that, in the absence of genotype errors, the exclusion-based methods (AF, Simpson, and SC) could be more accurate than the likelihood-based approach, especially when a small number of loci are considered. SC requires accurate population allele frequencies for estimation. The calculation speed of AF depends primarily on the family structure of the sample, and in some cases, AF failed to produce answers within a reasonable time. Konovalov et al. (2005) compared DR, Graph, Simpson, and a modified version of Simpson (denoted as MS) and found that MS outperformed the other three. However, in the presence of unrelated individuals, MS was less accurate than DR. In addition, MS requires a heuristic parameter, and there are theoretical limits of MS for highly unrelated data sets (Konovalov et al., 2005).

COLONY, one of the maximum likelihood methods, has been currently most frequently used in reconstructing full-sib families (Carlsson, 2007; Herrmann et al., 2007; Zavodna et al., 2007; Kraus et al., 2008). However, maximum likelihood approaches require a large number of individuals because they need accurate estimation of population allele frequencies (see Eq. (1)). Most of these methods are computationally intensive and, in many cases, do not work successfully on challenging data sets in which family contributions were strongly biased (Butler et al., 2004; Jones et al., 2007). Other approaches using fractional assignment with the Bayesian statistics (Almudevar, 2007) or using molecular coancestry (Nomura, 2008) also need large samples or huge number of marker loci.

For the study of initial phase of the intrusion, the pairwise approach, such as AF, Simpson, and MS, would be one of the suitable methods for estimating family structures of foraging bumble bees with DNA data. Alternative and practical methods that neither require heuristic parameter settings nor family structure dependent accuracy are essential for the study of full-sib reconstruction. In the next section, we introduce our new method for reconstructing full-sib structures especially in haplodiploid populations.

**SHARED LOCI CORRESPONDENCE ANALYSIS**

Shared Loci Correspondence Analysis (SLCA) is another pairwise approach. SLCA starts with a list of individuals with genotypes in a target population (Kokuvo et al., 2007a). The list is then converted into a similarity score matrix between all possible pairs in the population. The similarity score is the number of loci in which at least one allele is shared between a particular pair. Thus the diagonal elements of the score matrix are the total number of examined loci. Off-diagonal elements are equal or less than the total number of loci (Fig. 1).

We then apply the correspondence analysis (CA) to the score matrix so as to sort individuals for visualizing families within the score matrix. CA is a method commonly used for categorical data collected, for instance, in social surveys or linguistics (Benzecri, 1992). Recently, it has also been applied for the study of ecology, especially for studying community structures and plant-pollinator interactions (Lewinsohn et al., 2006). With CA, multi-dimensional plots are produced showing correlations within data sets on a series of axes. According to the order of these plots, higher score pairs are ordered along the diagonal of the sorted score matrix. As the order of individuals is the same along rows and columns, the sorted matrix is always symmetri-
cal along the diagonal. Moreover, the diagonal is always highest in the matrix. In this way, we used CA to obtain square clusters of families along the diagonal axis in the sorted similarity matrix (Fig. 1).

SLCA could accurately estimate the number of full-sib families in data sets with challengingly small sample size (Kokuvo et al., 2007a), but it overestimated the number of families when the actual number of families was intermediate (Fig. 2a). This bias could not be eliminated by increasing numbers of loci and alleles. There could be contradicting two directions for improving SLCA: increasing and decreasing information of similarity scores among individuals.

Weighting each locus by allelic polymorphism is one of the way of the former direction, but it simply made the score matrices difficult to sort by the correspondence analysis. We rather chose the second direction, or blunting the similarity scores by transforming the scores into binary; if a focal pair shares at least one allele at all loci, the pair has score 1 whereas the score is set to be 0 if at least one of loci does not share any alleles between the pair. The modified SLCA (mSLCA) adopting binary scores could almost perfectly estimate wide range of challenging problems when the numbers of loci and alleles are equal to or greater than eight (Fig. 2b). Such number of loci and alleles are feasible in many full-sib reconstruction studies.

The binary transformation of mSLCA may be a special expression of the 2-allele property for haplodiploid populations. In haplodiploid organisms, a female offspring (workers and queens in bumble bees) inherits the perfect set of genome from its father, and hence, the binary scores should be 1s if paired individuals share a common father. In other words, mSLCA divides the target population into patrilines rather than matrilines. A crucial assump-
tion for applying mSLCA to estimate nest numbers of hymenopteran populations is that queens are monandrous. Fortunately, this assumption is fulfilled in majority of hymenopteran species (Hughes et al., 2008). In the following section, we show some applications of SLCA and mSLCA for bumble bee populations in the field.

APPLICATION OF SLCA IN THE FIELD

Using six polymorphic DNA microsatellite loci, SLCA could accurately estimate the number of bumble bee colonies with data of challengingly small sample size (Kokuvo et al., 2007a, b). For the data sets, KINSHIP (Goodnight and Queller, 1999) erroneously split a single colony of the data into many families, but MS (Konovalov et al., 2005) could also estimate the number of colonies correctly (Toquenaga, personal comment). mSLCA improved the estimation power, and could accurately estimate the number of colonies of even commercial bumble bees where genetic variation would be more decreased than populations in the field (Fig. 3).

Based on the above examination, We applied SLCA to estimate the colony number of *B. terrestris* in Biratori, Hokkaido, Japan where the alien species has been heavily used for pollination of tomatoes in green houses (Kokuvo et al., 2008). We used six highly variable polymorphic DNA microsatellite loci that had respectively 10, 4, 8, 5, 4, and 10 alleles.

Contrasting to maximum likelihood methods, SLCA is a full-sib reconstruction for a given sample rather than estimating population means. Thus we apply Jackknife resampling to estimate the population mean of the number of colonies of escaping *B. terrestris*. A total of 28.82 ± 1.62 families are assigned to the 39 collected queens. With this data set, KINSHIP (Goodnight and Queller, 1999) returned almost the same number of estimated families, and consistent results between SLCA and KINSHIP indicated that those queens came from at least 27 colonies (pairs from 6 colonies and singles from 21 colonies). Quite a few *B. terrestris* queens flue freely in the wild in Biratori although we could not distinguish whether each queen came from commercial or feral nests.

As we mentioned above, SLCA relies on monandry of queens and actually divides a target pop-}

ulation into patrilines. Thus, we also applied mSLCA to estimate the number of male parents of a colony for four Japanese indigenous bumble bee species (Kokuvo et al., 2009). Results clearly showed that monoandry is the rule for the four Japanese bumble bee species. Reconstructing patrilines is also helpful to understand interactions within and between patrilines in polygynous social insects (Kudô et al., 2007).

CONCLUSION

Full-sib reconstruction is and will be a hot topic in basic sciences as well as applied fields. We succeeded to propose a heuristic method, i.e., SLCA, but its application is now restricted to monandrous haplodiploid species. SLCA consists of two parts: scoring similarity among individuals in a focal population and clustering the individuals according to the scores. Binary scoring and the correspondence analysis are good partners for monandrous hymenopteran species, but it would not be in general. We have to modify either of or the both parts for applying it to diploid and multiple-mating populations. Sometimes answers might not be the most parsimonious ones. Bayesian statistics and other brand-new novel approaches are awaited.
ACKNOWLEDGMENTS

This work was supported in part by Grants-in-Aid for Scientific Researches 3 (No. 17405004 to Yukihiko Toquenaga and Dr. Koichi Goka) from the JSPS and by Research Project for 4 Utilizing Advanced Technologies in Agriculture, Forestry and Fisheries (No. 51701 to Dr. Koichi Goka and Yukihiko Toquenaga).

REFERENCES


