Molecular Ecology Resources (2009) doi: 10.1111/j.1755-0998.2009.02562.x

COMPUTER PROGRAM NOTE

KINALYZER, a computer program for reconstructing sibling groups

M. V. ASHLEY,* I. C. CABALLERO,* W. CHAOVALITWONGSE,† B. DASGUPTA,‡ P. GOVINDAN,‡ S. I. SHEIKH‡ and T. Y. BERGER-WOLF‡

*Department of Biological Sciences, M/C 066, University of Illinois at Chicago, 845 W. Taylor Street, Chicago, IL 60607, USA, †Department of Industrial and Systems Engineering, Rutgers University, CoRE Building, 96 Frelinghuysen Road, Piscataway, NJ 08854, USA, ‡Department of Computer Science, M/C 152, University of Illinois at Chicago, 851 S. Morgan, Chicago, IL 60607, USA

Abstract

A software suite KINALYZER reconstructs full-sibling groups without parental information using data from codominant marker loci such as microsatellites. KINALYZER utilizes a new algorithm for sibling reconstruction in diploid organisms based on combinatorial optimization. KINALYZER makes use of a Minimum 2-Allele Set Cover approach based on Mendelian inheritance rules and finds the smallest number of sibling groups that contain all the individuals in the sample. Also available is a 'Greedy Consensus' approach that reconstructs sibgroups using subsets of loci and finds the consensus of the partial solutions. Unlike likelihood methods for sibling reconstruction, KINALYZER does not require information about population allele frequencies and it makes no assumptions regarding the mating system of the species. KINALYZER is freely available as a web-based service.

Keywords: combinatorial optimization, kinship, microsatellite DNA, sibgroup reconstruction, sibling

Received 13 October 2008; revision accepted 19 December 2008

Kinship reconstruction using codominant markers such as DNA microsatellites has become an important component of many investigations of wild populations (e.g. Pemberton 2008). The aim of kinship or pedigree reconstruction is to identify family groups, including parents, siblings, and higher-order relationships. Several methods and software for parentage assignment (maternity and paternity) are widely used and available (reviewed in Blouin 2003). Sibgroup reconstruction, with no or only partial parental information, is conceptually and computationally more difficult than parentage assignment, and sibling reconstruction studies have lagged behind those that use parentage assignment. More accurate and efficient approaches for sibgroup reconstruction are needed for cases where field studies sample cohorts of offspring, but obtaining samples of some or all candidate parents is less feasible. Recent examples include sampling of juvenile lemon sharks from nursery lagoons (Feldheim *et al*. 2004), brood parasitic cowbird nestlings sampled from host nests (Strausberger & Ashley 2003; Strausberger & Ashley 2005), wood duck eggs and nestlings (Roy Nielsen *et al*. 2006) and kelp bass larval cohorts (Selkoe *et al*. 2006). More studies would likely employ sibling reconstruction in data analysis if more robust approaches were widely available.

There have been several approaches taken for reconstructing full-sibling groups, although none has emerged as a clear favourite among molecular ecologists. Most sibgroup reconstruction methods use statistical likelihood models (Thomas & Hill 2000; Smith *et al*. 2001; Konovalov *et al*. 2004; Wang 2004) and, thus, rely on accurate estimates of underlying population allele frequencies, which may be difficult to obtain independently of the sample of potential siblings. The software Pedigree, available for use as on online service, employs Markov chain Monte Carlo (MCMC) methods for sib reconstruction by maximizing the joint likelihood of the entire sibship reconstruction rather than the pairwise relatedness ratio (Smith *et al*. 2001). Family Finder (Beyer & May 2003) uses a graph-based model, with edges representing pairwise sibling relationships that are weighted by the relationship likelihood (Goodnight & Queller 1999). Graph 'clusters' corresponding to sibling groups are identified by finding light edge cuts. Most of the available methods do not allow for genotyping errors or

Correspondence: Mary V. Ashley, Fax: (312) 996-9462; E-mail: ashley@uic.edu

mutations (Almudevar & Field 1999; Thomas & Hill 2000; Smith *et al*. 2001; Konovalov *et al*. 2004), yet errors are likely to occur at least at low frequencies in any large microsatellite data set. One exception is COLONY (Wang 2004). COLONY uses simulated annealing to exhaustively search for sibling reconstruction based on overall maximum likelihood, accounting for genotyping errors in the process.

In a test of four methods using simulated data, Butler *et al*. (2004) conclude that none of the algorithms performed well over the range of conditions tested, which included varying number of loci and alleles, family distributions, and errors in the data. In our recent review (Ashley *et al*. 2008) testing sibling reconstruction methods, we found that among statistical methods, COLONY (Wang 2004) accurately reconstructed siblings when sufficient number of loci were sampled (at least six) and allele diversity was high. However, COLONY is limited by an assumption of one gender monogamy and is too computationally demanding for analysis of moderate to large data sets in a reasonable time.

In contrast to statistical likelihood approaches, combinatorial approaches construct potential sibling groups using only Mendelian properties (Almudevar & Field 1999; Berger-Wolf *et al*. 2007; Sheikh *et al*. 2008) and search for the most parsimonious solution, such as the smallest number of mating pairs or parents. The method of Almudevar & Field (1999) uses a heuristic approach (rather than established computational optimization methods) to find a local optimum, but is not guaranteed to find the overall best solution (i.e. the smallest number of mating pairs). Alternatively, KINALYZER uses a combinatorial approach based upon a simple rule for allele inheritance in diploid organisms: an offspring inherits one allele from each of its parents for each locus. This rule of Mendelian inheritance introduces a necessary constraint on full-sibling groups in the absence of genotyping errors or mutations: the 2-allele property (Berger-Wolf *et al*. 2007; Ashley *et al*. 2008; Sheikh *et al*. 2008). The 2-allele property states that there exists an assignment of individual alleles within a locus to maternal and paternal parents such that the number of distinct alleles assigned to each parent at this locus does not exceed two. Barring mutation or genotyping error, any sibling group must satisfy this constraint.

Formally, a diploid individual i sampled at l loci is represented by its l pairs of alleles: $i = [(a_{i1}, b_{i1}), (a_{i2}, b_{i2}), \ldots, (a_{il}, b_{il})]$. A set of individuals S in a population sample U has the 2-allele property if for each individual i in S at each locus there exists an assignment of the two alleles $a_{ij} = c_{ij}$ and $b_{ij} = \hat{c}_{ij}$ or $a_{ij} = \hat{c}_{ij}$ and $b_{ij} = c_{ij}$ such that

$$\forall 1 \le j \le l : \left| \bigcup_{i \in S} \{c_{ij}\} \right| \le 2 \text{ and } \left| \bigcup_{i \in S} \{\overline{c}_{ij}\} \right| \le 2$$

KINALYZER employs the Minimum 2-Allele Set Cover approach to find the smallest number of sibgroups $S_1, \ldots, S_m$ such that each sibgroup consists of a subset of individuals in U, the 2-allele property is satisfied for every sibgroup, and every individual is contained in at least one sibgroup ($US_i = U$). This smallest number of feasible sibgroups (that satisfy the 2-allele constraint) is found using a combinatorial optimization technique to select the fewest possible sibgroups.

Combinatorial optimization is a class of problems where the qualitative (combinatorial) structure is more important than the numerical values. Such problems are defined by structural constraints on potential solutions and a cost associated with each solution. The objective is to find a solution which optimizes (minimizes or maximizes) the cost. For the Minimum 2-Allele Set Cover, a feasible solution is any partition of individuals into groups that satisfy the 2-allele property (the structural constraint). The cost of each solution is the number of groups, and the objective is to find the solution with the smallest number of groups. Such combinatorial optimization problems are typically provably hard (computationally infeasible) and the Minimum 2-Allele Set Cover is no exception (Ashley *et al*. 2008).

There are a wide variety of computational techniques that solve combinatorial optimization problems (Cook *et al*. 1997; Papadimitriou & Steiglitz 1998). While any technique applied to any particular combinatorial problem may take a long, even infeasible, time to find a solution, such solution, when found, is *guaranteed* to be optimal. Combinatorial optimization in KINALYZER is based on the implementation of CPLEX,[1] a commercially available optimization software package. CPLEX employs multiple optimization algorithms including simplex, cutting plane, interior point, barrier, and branch-and-bound to solve difficult combinatorial optimization problems and is guaranteed to find the overall optimum. Note that while every optimization technique will find the overall optimum, some may take longer than others on any given problem. The main advantage of CPLEX is that the most efficient optimization algorithm is used based on the structure of the problem.

The computational objective of minimizing the number of sibling groups is formally equivalent to minimizing the number of mating pairs, and provides the most parsimonious reconstruction goal. The solution satisfies Mendelian rules of inheritance and is *guaranteed* to be the optimal solution *if* the underlying objective of the smallest number of matings is correct. As we develop computational methods with different biological objectives, such as minimizing the number of fathers or maximizing family size, these will be added to the KINALYZER software suite.

KINALYZER also includes a consensus-based approach ('Greedy Consensus') that discards individual loci one at a time and reconstructs solutions using the remaining loci. The final solution output is a consensus of the partial solutions. The consensus is calculated by first computing the

[1] CPLEX™ is a registered trademark of ILOG.

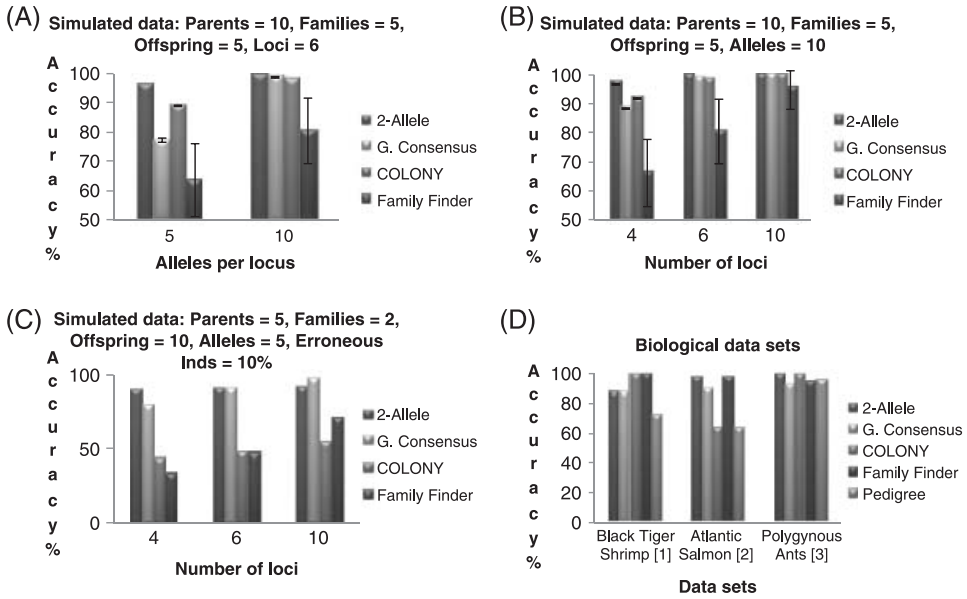

Fig. 1 Comparison of accuracy of different sib reconstruction approaches. A–C show results using simulated data. Simulated data was generated by first randomly generating parent pairs based on population parameters (alleles per locus, number of loci), and then randomly generating their offspring. The number of male/female parents, families and the offspring per family were varied as indicated to generate the simulated populations. Each algorithm was run on simulated data sets created with the specified parameters until the mean and standard deviation of error rates were stable for 10 consecutive iterations. Accuracy was calculated by the Gusfield Partition Distance (Gusfield 2002) between the algorithm's reconstruction and the known sibling relationships (see Ashley *et al*. 2008 for further details on simulations). '2-Allele' refers to the minimum set cover implemented in KINALYZER. 'G. Consensus' (Sheikh *et al*. 2008) refers to the consensus approach described in the text that is also available in KINALYZER. D shows analysis of real data where sibling relationships (from controlled crosses) were known for tiger shrimp *Penaeus monodon* (Jerry *et al*. 2006), Atlantic salmon (Herbinger *et al*. 1999) and the polygynous ants *Leptothorax acervorum* (Hammond *et al*. 1999).

groups that are in common and then greedily (taking the best immediate, or local, solution) merging the nearest pair of groups iteratively. Distance is computed based on costs associated with errors and allelic information shared (see Sheikh *et al*. 2008 for details).

Using simulated and real data with known sibling relationships, we have compared available sibling reconstruction software (Berger-Wolf *et al*. 2007; Ashley *et al*. 2008; Sheikh *et al*. 2008). An example of a comparison of KINALYZER to three of the commonly used statistical methods, Pedigree (Smith *et al*. 2001), COLONY (Wang 2004) and Family Finder (Beyer & May 2003) is shown in Fig. 1. Error rates were calculated using the Gusfield partition distance (Gusfield 2002), the minimum number of individuals to remove in order to make the two partitions (the reconstructed sibgoups and the actual sibgroups) equivalent. Overall, KINALYZER performed as well or better than other methods on a wide range of data set parameters. It remains robust even when the allelic diversity is low (Fig. 1A), the number of loci sampled is small (Fig. 1B), and there are genotyping errors (Fig. 1C). It also performed well on three different biological data sets tested, while three other available methods were less consistent (Fig. 1D) (Berger-Wolf *et al*. 2007; Ashley *et al*. 2008).

KINALYZER is a web-based program that requires an input file comprising the individuals and genotypes to analyse. Because many users will already be familiar with Kinship (Goodnight & Queller 1999), KinGroup (Konovalov *et al*. 2004), GERUD (Jones 2005) or Cervus (Marshall *et al*. 1998) input files, we have preserved these for KINALYZER with the exception that no population allele frequencies are needed to run the program. To upload the genotype data file, the user logs into the website and provides their name and e-mail address (Fig. 2) which are necessary to deliver the results. Currently KINALYZER only accepts .csv input files from Excel. Three- or two-digit coded alleles are automatically recognized by the program and missing data should be coded as –1 (failure to do so will prompt the program to display a message to correct the format). The columns should correspond to the identity of individuals and name of loci. The input file may contain extra columns not used by KINALYZER (i.e. sex, locality, group, etc.); the software has an option to disable them prior to uploading the file. There is no limit to sample sizes or number of loci. The web address for KINALYZER is http://kinalyzer.cs.uic.edu/.

Because this is a web-based program, the user will be given an input file number and the results are delivered via

Fig. 2 Screenshot of KINALYZER software interface, showing user login, data upload and formatting windows, and confirmation of submissions with information on receiving the results.

e-mail. The time to analyse the data will depend on how many jobs the server is processing at that time. Users can find out about the status of the queue job online at any time (using the input file number). The output file will show individuals divided into full-sib groups (sets). Each one of these sets will list the individuals by the identification name (or number) that was provided in the input file.

Because sibling reconstruction is still a developing field, we recommend that investigators try different approaches, and select an appropriate procedure based on their study systems and the assumptions and limitations of currently available methods. No single method is guaranteed to provide the correct answer, but we favour the 2-allele method implemented in KINALYZER because of the available methods, it makes the fewest number of assumptions and performs well over a wide range of data parameters. It is, therefore, a good general method, especially when few loci are sampled or the allelic diversity is low (Fig. 1). The 'Greedy Consensus' method was found to be highly accurate in tests using benchmark data, especially when allelic variation was low, and was highly tolerant of genotyping errors and mutations (Sheikh *et al*. 2008). As mentioned above, other reconstruction objectives will also be added to KINALYZER as they are developed.

Acknowledgements

The development of KINALYZER was supported by NSF IIS-0612044 and IIS-0611998 (Berger-Wolf, Ashley, Chaovalitwongse, DasGupta), NSF CCF-0546574 (Chaovalitwongse), NSF IIS-0747369 CAREER (Berger-Wolf), NSF DBI-0543365 (DasGupta), NSF IIS-0346973 (DasGupta), DIMACS special focus on Computational and Mathematical Epidemiology (DasGupta) and a Fulbright Scholarship (Sheikh). Numerous people have shared their data for testing KINALYZER, including Jeffrey Connor, the Atlantic

Salmon Federation, Dean Jerry and Stuart Barker. We also thank Anthony Almudevar, Bernie May, and Dmitry Konovalov for sharing their software.

References

Almudevar A, Field C (1999) Estimation of single-generation sibling relationships based on DNA markers. *Journal of Agricultural Biological and Environmental Statistics*, **4**, 136–165.

Ashley MV, Berger-Wolf TY, Caballero IC, Chaovalitwongse W, DasGupta B, Sheikh SI (2008) Full sibling reconstruction in wild populations from microsatellite genetic markers. In: *Computational Biology: New Research*. Nova Science Publishers, Hauppauge, New York.

Berger-Wolf TY, Sheikh SI, DasGupta B, Ashley MV, Caballero IC, Chaovalitwongse W, Putrevu SL (2007) Reconstructing sibling relationships in wild populations. *Bioinformatics*, **23**, I49–I56.

Beyer J, May B (2003) A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, **12**, 2243–2250.

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, **18**, 503–511.

Butler K, Field C, Herbinger CM, Smith BR (2004) Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Molecular Ecology*, **13**, 1589–1600.

Cook WJ, Cunningham WH, Pulleyblank WR, Schrijver A (1997) *Combinatorial Optimization*, 1st edn. John Wiley & Sons, New York.

Feldheim KA, Gruber SH, Ashley MV (2004) Reconstruction of parental microsatellite genotypes reveals female polyandry and philopatry in the lemon shark, *Negaprion brevirostris*. *Evolution*, **58**, 2332–2342.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**, 1231–1234.

Gusfield D (2002) Partition-distance: a problem and class of perfect graphs arising in clustering. *Information Processing Letters*, **82**, 159–164.

Hammond RL, Bourke AFG, Bruford MW (1999) Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum*. *Molecular Ecology*, **10**, 2719–2728.

Herbinger C, O'Reilly P, Doyle R, Wright J, O'Flynn F (1999) Early growth performance of Atlantic salmon full-sib families reared in single family tanks or in mixed family tanks. *Aquaculture*, **173**, 105–116.

Jerry D, Evans B, Kenway M, Wilson K (2006) Development of a microsatellite DNA parentage marker suite for black tiger shrimp *Penaeus monodon*. *Aquaculture*, 542–547.

Jones AG (2005) GERUD 2.0: a computer program for the reconstruction of parental genotypes from half-sib progeny arrays with known or unknown parents. *Molecular Ecology Notes*, **5**, 708–711.

Konovalov DA, Manning C, Henshaw MT (2004) KinGroup: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, **4**, 779–782.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

Papadimitriou CH, Steiglitz K (1998) *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, Mineola, New York.

Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society Series B*, **275**, 613–621.

Roy Nielsen C, Gates R, Parker P (2006) Intraspecific nest parasitism of wood ducks in natural cavities: Comparisons with nest boxes. *Journal of Wildlife Management*, **70**, 835–843.

Selkoe KA, Gaines SD, Caselle JE, Warner RR (2006) Current shifts and kin aggregation explain genetic patchiness in fish recruits. *Ecology*, **87**, 3082–3094.

Sheikh SI, Berger-Wolf TY, Chaovalitwongse W, Ashley MV (2008) Error-tolerant sibship reconstruction in wild populations *7th Annual International Conference on Computational Systems Bioinformatics*.

Smith BR, Herbinger CM, Merry HR (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.

Strausberger BM, Ashley MV (2003) Breeding biology of brood parasitic brown-headed cowbirds (*Molothrus ater*) characterized by parent-offspring and sibling-group reconstruction. *Auk*, **120**, 433–445.

Strausberger BM, Ashley MV (2005) Host use strategies of individual female brown-headed cowbirds *Molothrus ater* in a diverse avian community. *Journal of Avian Biology*, **36**, 313–321.

Thomas SC, Hill WG (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, **155**, 1961–1972.

Wang JL (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.