

Combinatorial Reconstruction of Sibling Groups

Tanya Berger-Wolf,¹ Bhaskar DasGupta,² Wanpracha Chaovalitwongse,³
Mary V. Ashley,⁴

Keywords: population genetics, sibling relationships, combinatorial optimization.

1 Introduction.

Knowledge about sibling relationships is used in genetic epidemiology, conservation biology, and animal management. For example, knowledge of the genetic relationships among individuals is critical for estimating heritabilities of quantitative characters, for characterizing mating systems and fitness, and for managing populations of endangered species.

When parental data are available, sibling groups can be established through parentage assignments (e.g., [3]). Assignment of individuals to full or half sibling groups in the absence of parental data is more challenging, however, it is often more practical. In recent years, there has been an explosion of methods that reconstruct sibling relationships without the parental data [1].

We propose the first fully combinatorial optimization approach to reconstructing sibling groups based on single generation genetic data with no parental information. We use the Mendelian inheritance rules to impose constraints on the genetic content possibilities of a sibling group. We formulate the inferred combinatorial constraints and use a provably correct algorithm to construct the smallest number of groups of individuals that satisfy these constraints. Our algorithm allows half-sibling relationships to exist in the population. The algorithm requires no prior knowledge about the allele frequency, number of loci sampled, mating system, or the size of the family groups. It can be easily extended to incorporate null-allele type errors. To assess the accuracy of our approach, we use a weaker (but computationally cheaper) version of our algorithm on simulated data that has known parents and, therefore, sibling groups.

2 Problem Statement.

Given a set of n diploid individuals of the same generation, U , the goal is to reconstruct the existing sibling relationships among them. Each individual $1 \leq i \leq n$ is represented by a genetic marker of l loci $\langle (a_{ij}, b_{ij}) \rangle_{1 \leq j \leq l}$. The numbers a_{ij} and b_{ij} represent a specific allele. Mendelian inheritance laws impose two necessary (but not sufficient) constraints on a group of diploid individuals $S \subseteq U$ to be full siblings. We say that a set $S \subseteq U$ has the *4-allele property* if for all $1 \leq j \leq l$ $|\cup_{i \in S} a_{ij} \cup b_{ij}| \leq 4$ and a set $S \subseteq U$ has the *2-allele property* if for all $1 \leq j \leq l$ $|\cup_{i \in S} a_{ij}| \leq 2$ and $|\cup_{i \in S} b_{ij}| \leq 2$. We propose provably correct algorithms based on a Set Cover approach for reconstructing groups that satisfy either property. The 2-allele property is equivalent to a biologically consistent full sibling relationship. However, it is computationally more expensive and we use the weaker 4-allele property for a preliminary experimental assessment of our approach.

¹Center for Discrete Mathematics and Theoretical Computer Science (DIMACS). E-mail: tanyabw@dimacs.rutgers.edu

²Dept. of Computer Science, University of Illinois at Chicago. E-mail: dasgupta@cs.uic.edu

³Department of Industrial Engineering, Rutgers University. E-mail: wchaoval@rci.rutgers.edu

⁴Department of Biological Sciences, University of Illinois at Chicago. E-mail: ashley@uic.edu

3 Experiment Design and Results.

We first create the adults with the full genetic information and then generate a single generation of juveniles. The parent information is retained therefore we know the true sibling groups. We then use our **4-alleleSets** algorithm to reconstruct the sibling groups. Finally, we use the extension of the [2] partition distance to measure the accuracy of the reconstruction with respect to the true sibling groups.

We compare the groups reconstructed by the **4-alleleSets** algorithm with the true sibling groups. We examine the error rate behavior as a function of the number of loci, alleles per each locus, juvenile population size, and maximum family size (number of offspring). Figure 1 shows selected corresponding graphs. These are representative of the data. As expected,

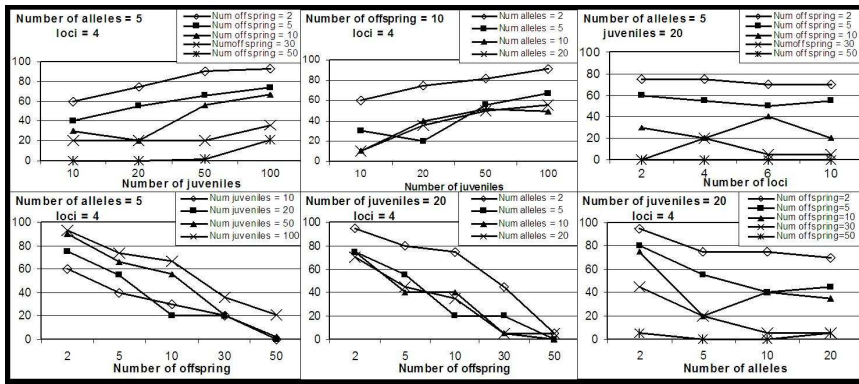


Figure 1: 4-allele algorithm error rate (percent of the number of juveniles) as a function of the number of loci, alleles per locus, juveniles, and maximum offspring per couple.

the error increases with the number of juveniles and decreases with the number of offspring per family. Surprisingly, the number of alleles per locus and the number of sampled loci are not strong factors (except when there are only 2 alleles per locus). It is important to note that in most cases the algorithm found fewer sibling groups than there are in the population, merging true families into a reconstructed one. This leads us to believe that the stronger algorithm **2-alleleSets** will have more discriminating power to separate these groups and thus be more accurate.

References

- [1] Michael S. Blouin. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *TRENDS in Ecology and Evolution*, 18(10):503–511, October 2003.
- [2] Dan Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, May 2002.
- [3] A. G. Jones and W. R. Ardren. Methods of parentage analysis in natural populations. *Molecular Ecology*, (12):2511–2523, 2003.