Combinatorial Reconstruction of Sibling Relationships

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

[1] Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), {tanyabw,wchaoval}@dimacs.rutgers.edu,
[2] Department of Computer Science, University of Illinois at Chicago, dasgupta@cs.uic.edu
[3] Department of Industrial Engineering, Rutgers University, wchaoval@rci.rutgers.edu,
[4] Department of Biological Sciences, University of Illinois at Chicago, ashley@uic.edu

Abstract

We present a new algorithm for reconstructing sibling relationships in a single generation of individuals without parental information, using data from codominant DNA markers such as microsatellites. We use the simple genetic constraints on the full-sibling groups, imposed by the Mendelian inheritance rules, and combinatorial optimization techniques to extract a minimum number of consistent sibling groups. The results of a simulation study of a relaxed version of the algorithm show that our approach is reasonably accurate and the full version of the algorithm should be pursued. Our algorithm does not require any a priori knowledge about allele frequency, population size, mating system, or family size distributions.

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., [8]). Assignment of individuals to full or half sibling groups in the absence of parental data is more challenging. Nonetheless, for many studies, particularly those that rely on sampling of wild populations, it is often more practical to sample cohorts of offspring rather than parent/offspring groups.

In recent years, there has been an explosion of methods that reconstruct sibling relationships without the parental data [4]. Most of them use statistical population parameters to find maximum likelihood clusters ([3], [6],[11],[12], [13]). There are two methods that incorporate a combinatorial approach to the problem. [2] uses graph clustering algorithms to form groups from pairwise likelihood distance graph and [1] enumerates all possible potential full-sibling groups based on the Mendelian inheritance rules and uses a heuristic to construct a maximal (but not necessarily optimal) partition of the individuals into those groups.

In this paper, we present a fully combinatorial optimization approach to reconstructing sibling groups based on single generation genetic data with no parental information. Our approach is the proper formalization of the algorithm in [1]. 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. Unlike [1], 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. We use an extension of the partition distance presented in [7] to compute a mathematically correct distance between our solution and the true sibling groups.

Our preliminary results show that the new combinatorial approach can be sufficiently powerful to accurately reconstruct sibling groups. Nonetheless, to validate this approach and its applicability more extensive simulations and experiments are required, as well as comparison to other known methods.

1.1 Problem Statement

We now formally state the sibling relationship reconstruction problem. 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) con-

straints on a group of diploid individuals $S \subseteq U$ to be full siblings:

Definition 1. *A set $S \subseteq U$ has the 4-allele property if for all $1 \leq j \leq l$ $|\cup_{i \in S} a_{ij} \cup b_{ij}| \leq 4$.*
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 $|\cup_{i \in S} b_{ij}| \leq 2$.

The 2-allele property is clearly stronger than the 4-allele property. Assuming the order of the parental alleles is always the same in the offspring (i.e., the maternal is always on the same side), the 2-allele property is equivalent to a biologically consistent full sibling relationship. The parental allele order, however, is not preserved and an interesting problem arises: given a set S that satisfies the 4-allele property, does there exist a series of allele switches in some loci of some individuals in S so that after those switches S satisfies the 2-allele property?

Theorem 1. *Let a be the number of distinct alleles present in a given locus and R be the number of distinct alleles that either appear with three different alleles in this locus or are homozygous (appear with itself). Then given a set of individuals with the 4-allele property there exists a series of allele switches in some of the loci resulting in a set that satisfies the 2-allele property if and only if for all the loci in the set $a + R \leq 4$.*

In general, however, we are interested in reconstructing consistent sibling groups. We model this goal by the following combinatorial optimization problem **Minimum 2-allele Set Cover**: given a collection U of n l-tuples, find a minimum number of subsets $S_1, ..., S_m$ in U that satisfy the 2-allele property and whose union is U.

The following is a simple (although not the most efficient) algorithm to solve the **Minimum 2-allele Set Cover**:

1. For each locus, independently, create all possible 2-allele sets of individuals: if there are a alleles in the locus, of which R are homozygous, then there are at most $\binom{a}{4} + \binom{R}{3} + \binom{R}{2} = O(a^4)$ sets.
2. Find the sets that are consistent with all the loci. These sets must exist, since any pair of individuals forms a consistent sibling set.
3. Find a minimum set cover of all the individuals from among the sets in previous step.

This algorithm, while biologically consistent, is computationally expensive. Therefore we use the weaker but computationally cheaper 4-allele property as a heuristic for the sibling groups reconstruction.

2 The 4-allele Set Cover

We use the 4-allele property and a reduction to a specific instance of **Minimum Set Cover** problem to identify sibling groups among a given group of juveniles. We assume that the relationships may be promiscuous and half siblings may be both paternal and maternal. Thus, an individual animal may be in more than one sibling group. First, we define the **Minimum 4-allele Set Cover** problem: given a collection U of n l-tuples, find a minimum number of subsets $S_1, ..., S_m$ in U that satisfy the 4-allele property and whose union is U.

The **Minimum Set Cover (MSC)** problem is defined as follows: given a universe $U = \{1, 2, ..., n\}$ and a collection of sets $\mathcal{S} = \{S_1, S_2, ..., S_m\}$ such that $S_i \subseteq U$, find the smallest number of sets in $\mathcal{S}$ whose union is the universe. **MSC** problem is NP-hard. In our solution, we use the standard integer programming formulation of the **MSC** problem: given the $n \times m$ elements-sets matrix A,

$$a_{ij} = \begin{cases} 1 \text{ if } i \in S_j \\ 0 \text{ otherwise} \end{cases} \qquad \text{find} \qquad \begin{array}{l} \min \sum_{i=1}^{m} x_i \\ \text{s.t. } Ax \geq \overline{1} \\ \quad x_i \in \{0, 1\} \end{array}$$

We use the following algorithm to solve the **Minimum 4-allele Set Cover**:

1. For each pair of individuals A_p and A_q form a set S_{pq} that represents their 4-allele property. That is, S_{pq} is a collection of l loci where each locus is a union of alleles of the corresponding locus for p and q.
2. An animal belongs to a set S_{pq} if for each locus the set of the alleles of the animal for that locus is in the the corresponding locus set of S_{pq}.
3. Find a **MSC** S. For each set in S define the corresponding set of individuals covered by that set as a sibling group. Return the group structure induced by S as the answer.

Proposition 1. *Any set cover of the elements by the sets defined above is a valid collection of 4-allele groups.*

We use computer experiments on simulated data to assess the accuracy of the **4-alleleSets** algorithm.

3 Experiment Design

For this set of simulations, 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 the **4-alleleSets** algorithm to

reconstruct the sibling groups. Finally, we use the extension of the [7] partition distance to measure the accuracy of the reconstruction with respect to the true sibling groups (see section 3.2 for more details). As we have pointed out earlier, our algorithm assumes that the organisms are diploid, therefore all the simulated organisms are diploid as well.

3.1 Experiment Protocol and Parameters

We created a given number of adult males M and females F with a given number of loci l and a specified number of alleles per locus a (for this set of experiments, a is the same for all the loci). Each individual was created by randomly choosing from an independent identical uniform distribution $2l$ number of alleles from among a alleles. They are paired up into l loci. We then create the specified jF number of juveniles, where j is the factor of the number of juveniles as the number of females. A male and a female is chosen randomly, independently and uniformly from the adult population. A couple has a random number of offspring, up to a specified maximum number of offspring o (for this study, o is the same throughout the population). Each offspring randomly gets one of the mother's and one of the father's alleles per locus which are assembled randomly. There are several simplifications made in this simulation (see section 5), all of which can be addressed in the future. Nonetheless, this protocol creates a biologically consistent population of juveniles with known parents.

The parameter ranges for the study are as follows:

- The number of adult females $F = 10$ and the number of adult males $M = 10$.
- The number of loci sampled $l = 2, 4, 6, 10$.
- The number of alleles per locus $a = 2, 5, 10, 20$.
- The factor of the number of juveniles as the number of females $j = 1, 2, 5, 10$.
- The maximum number of offspring per couple $o = 2, 5, 10, 30, 50$.

We use the **4-alleleSets** algorithm on the juvenile population to find the smallest number of 4-allele sets and designate them as the full sibling groups. While the **MSC** problem is NP-hard, modern Mixed Integer Programming (MIP) solvers can solve our simulation instances optimally. We formulated the MSC as a MIP problem and used a commercial MIP solver from CPLEX 9.0[1] to obtain an optimal solution (the minimum number of the 4-alle sets). All the instances of the set cover problem in our simulation were solved optimally in about 10 seconds.

[1] CPLEX is a registered trademark of the ILOG,Inc.

We compare the groups reconstructed by the **4-alleleSets** algorithm with the true sibling groups. In the past, several methods have been used to compare the true groups and the reconstructed groups. However, they are mathematically inconsistent. We use an extension of the clustering distance measure described in [7] and implemented in [9].

3.2 Error Measure

In [7] Gusfield showed that the minimum distance between two set partitions is equivalent to the **Maximum Assignment** problem (maximum bipartite weighted matching) – a well known linear programming problem [5, 10]. The minimum number of elements that need to be deleted so that the two partitions become identical is (total number of elements)–(maximum assignment). Gusfield uses this number as the distance between two partitions.

In our case, since the the relationships are not necessarily monogamous, the set of full sibling groups does not induce a partition on the individuals and the formula has to be slightly adjusted. Given two collections of non-disjoint sets $\{P_1, ..., P_n\}$ and $\{Q_1, ..., Q_m\}$ of elements in U the minimum distance between the two set collections is the minimum number of elements needed to be removed so that the remaining set collections are identical.

4 Experiment Results

As stated, the goal of our experiments is to assess the accuracy of the **4-allele-Sets** algorithm. We define the error in reconstruction as the distance between the 4-allele sets and 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, 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.

5 Future Work and Extensions

Of course, our results are preliminary and more research is needed. We need to investigate the computa-

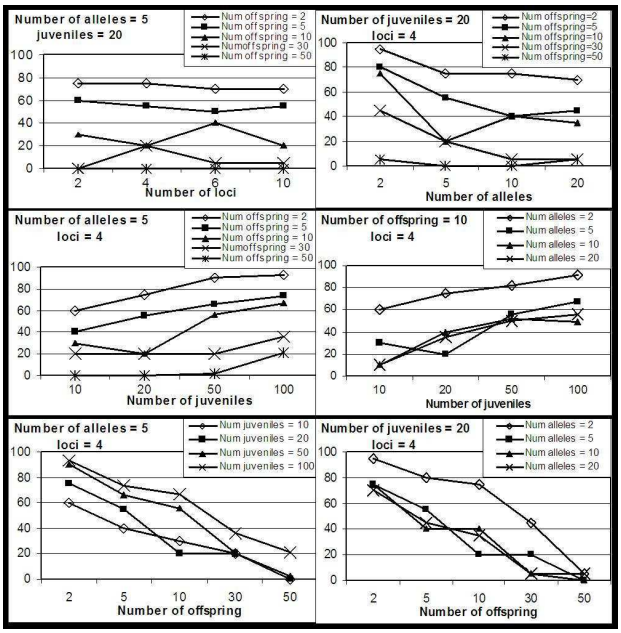

Fig. 1. 4-allele algorithm error rate (percent of the number of juveniles) as a function of the number of loci, alleles per locus, juveniles, and maximum offspring per couple.

tional complexity and better algorithmic solutions to the **Minimum 2-allele Set Cover** and **Minimum 4-allele Set Cover** problems. We need to conduct simulations with the **Minimum 2-allele Set Cover** algorithm and run these for a wider range of parameters and parameter distributions, as well as allow for errors in the data. We need to validate the results on biological datasets, especially where the sibling groups have been established using other methods. Finally, we need to compare the performance of our method to other methods of sibling reconstruction.

6 Conclusions

We have presented a fully combinatorial method for reconstructing sibling relationships in the absence of parental data. Unlike other existing methods, it does not use any statistical estimates of the relatedness among the individuals, but rather a direct Mendelian constraint on the possible genetic content of a sibling group. A simple such constrain turns out to be sufficiently powerful to reconstruct the sibling relationships fairly accurately in our simulations. The stronger version of this constraint, we believe, has the potential to accurately reconstruct sibling groups without any prior knowledge of the population structure and its genetic characteristics.

Acknowledgments

This research is supported by the following grants: EIA 02-05116, CCR-0206795, CCR-0208749 and IIS-0346973. We thank ExxonMobil Research and Engineering for the use of the CPLEX solver license and the computer cluster.

References

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

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

[3] M. Blouin, M. Parsons, V. Lacaille, and S. Lotz. Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*, 5:393–401, 1996.

[4] M. 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.

[5] W. Cook, W. Cunningham, W. Pulleyblank, and A. Schrijver. *Combinatorial Optimization*. Wiley-Interscience Publications, 1998.

[6] S. C.Thomas and W. G.Hill. Sibship reconstruction in hierarchical population structures using markov chain monte carlo techniques. *Genet. Res., Camb.*, 79:227–234, 2002.

[7] D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, May 2002.

[8] A. G. Jones and W. R. Ardren. Methods of parentage analysis in natural populations. *Molecular Ecology*, (12):2511–2523, 2003.

[9] D. A. Konovalov, B. Litow, and N. Bajema. Partition distance via the assignment problem. *Bioinformatics*, March 2005. Advanced Access Publication.

[10] E. Lalwer. *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart and Winston, New York, USA, 1976.

[11] I. Painter. Sibship reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:212–229, 1997.

[12] B. R. Smith, C. M. Herbinger, and H. R. Merry. Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, 158:1329–1338, 2001.

[13] J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166:1968–1979, April 2004.