DIMACS Technical Report 2005-27
September 2005

A Combinatorial Reconstruction of Sibling Relationships in Absence of Parental Data

by

Tanya Y. Berger-Wolf[1]
Center for Discrete Mathematics and Theoretical Computer Science (DIMACS)
and Department of Computer Science, University of Illinois at Chicago
tanyabw@cs.uic.edu

Bhaskar DasGupta[2]
Department of Computer Science, University of Illinois at Chicago
dasgupta@cs.uic.edu

Wanpracha Chaovalitwongse
Center for Discrete Mathematics and Theoretical Computer Science (DIMACS)
and Department of Industrial Engineering, Rutgers University
wchaoval@rci.rutgers.edu

Mary V. Ashley
Department of Biological Sciences University of Illinois at Chicago
ashley@eeb.uic.edu

[1]DIMACS Postdoctoral Fellow, suported by the National Science Foundation postdoctoral fellowship grant EIA 02-05116
[2]Supported by the National Science Foundation grants CCR-0206795, CCR-0208749 and IIS-0346973

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, such as the limit of four alleles per single locus and no more than two alleles from each potential parent per locus. We then use combinatorial optimization techniques to extract a minimum number of groups that satisfy these constraints. 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 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., [10]). 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 [3]. Most of them use statistical population parameters to find maximum likelihood clusters ([4], [12], [14], [6], [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 missing-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, known sibling groups. We use an extension of the partition distance presented in [8] 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 Definitions

A group of individuals is called *full siblings* if they have the same parents. A group of individuals is called *half siblings* if exactly one of the two parents is the same for every animal in the group. A group of individuals is called *siblings* if they are either full or half siblings.

Allele is one of the different versions of the same gene found at the same locus but in homologous chromosomes or in different individuals.

Allele frequency is the fraction of all the alleles of a gene in a population that are of one type.

Genotype is the actual alleles present in an individual; the genetic makeup of an organism.

Genetic marker is a set of alleles of genes used as experimental probes to keep track of an individual.

Locus is the location of a gene on a chromosome.

A *diploid* individual is one having two alleles (not necessarily different) for each locus.

A *homozygous* individual is one having two identical alleles at a particular genetic locus.

A *heterozygous* individual is one having two different alleles at a particular genetic locus.

1.2 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} each represent a specific allele, which may be the same or different. Mendelian inheritance laws impose two necessary (but not sufficient) constraints 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 $1 \leq j \leq l$ $|\cup_{i\in S} \{a_{ij}, b_{ij}\}| \leq 4$. 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? For example, with only one locus the set $\{(1,2),(1,3),(1,4)\}$ cannot satisfy the 2-allele property, while the set $\{(1,2),(2,3),(1,4)\}$ can, with the switch $(3,2)$ in the second animal.

We use the following **notation**: let a be the number of distinct alleles present in a given locus and R be the number of distinct alleles that appear with three different alleles in this locus or are homozygous (appear with themselves).

Theorem 1 *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*

$$a + h + R \leq 4.$$

Proof. Clearly, to verify that a set satisfies the 2-allele property (as well as the 4-allele property) it is sufficient to verify that each locus of the set satisfies it independently.

For the necessary direction, suppose there exists a series of switch operations that result in a set that satisfies the 2-allele property. Consider any locus j in the set after the switches have been performed: $\langle a_{ij}, b_{ij} \rangle_{i \in S}$. Since the 2-allele property is satisfied, each allele a_{pj} appears with no more than two different alleles b_{qj} among all the individuals and vise versa.

- Consider the loci with the number of different alleles $a = 4$. If all a_{ij} are different from all b_{ij} then there is no allele that appears with more than two other alleles, so $R = 0$ and $a + R \leq 4$ is satisfied. If some allele $a_{pj} = b_{qj}$, that is, it appears both on the left and on the right in this locus, then there is only one more allele that can be on the left side and one more allele that can be on the right side, so $a = 3$, which is a contradiction.

- If $a = 3$ and there is no allele $a_{pj} = b_{pj}$ then either $| \cup a_{ij}| = 1$ or $| \cup b_{ij}| = 1$ and each allele appears with no more than two other alleles, $R = 0$ and $a + R < 4$. If there is $a_{pj} = b_{pj}$ then this allele may appear with three other alleles in the individuals (itself and the remaining two alleles). The two other alleles appear with at most two alleles each that are different from themselves (otherwise $a = 2$). Thus, $R = 1$, and $a + R = 4$, as stated.

- If $a \leq 2$ then, clearly, $R \leq 2$ and $a + R \leq 4$.

For the sufficient direction, suppose that for all the loci $a + R \leq 4$. We consider each locus independently and, again, distinguish between the three cases, based on the value of a for a given locus j.

- If $a = 4$ then $R = 0$ and no allele appears with more than three other alleles in this locus and no allele can appear with itself. Consider some allele a_{pj}. Switch the alleles so that a_{pj} is on the same side (left) of the pair in all the individuals it appears. a_{pj} appears with at most two other alleles, therefore there must be another allele a_{qj} that does not appear with a_{pj} in any animal. Switch the alleles in the individuals containing a_{qj} so it is on the same side as a_{pj} in all the individuals. If each animal has either a_{pj} or a_{qj} then we are done and now $\cup_{i \in S} a_{ij} = \{a_{pj}, a_{qj}\}$ and since the other side contains the remaining two alleles, the 2-allele property is satisfied. However, suppose there are individuals that have neither a_{pj} or a_{qj}. Then we switch the alleles so that all the remaining individuals contain a pair (b_{sj}, b_{tj}). These alleles b_{sj} and b_{tj} could not appear with both a_{pj} and a_{qj}, since then $R > 0$. Without the loss of generality, a_{pj} appears with b_{sj}, then we can switch all the individuals that contain a_{pj} so that now b_{sj} now appears on the left. Now the union of the alleles on left side is $\{a_{qj}, b_{sj}\}$ and the union on the right side is $\{a_{pj}, b_{tj}\}$ and the 2-allele property is satisfied.

- If $a = 3$ then $R \leq 1$. Notice, that if there is an allele that appears with three other alleles it necessarily appears with itself. Let a_{pj} and a_{qj} be the alleles that do not appear with itself. We can switch the alleles so that a_{pj} is on the left and a_{qj} is on the right in all the individuals that contain them. Let b_{ij} be the third allele. Then the union on the left side is contained in $\{a_{pj}, b_{ij}\}$ while the unions on the right side is contained in $\{a_{qj}, b_{ij}\}$ and the 2-allele property is satisfied.

- If $a = 2$ then the 2-allele property is satisfied trivially.

□

Essentially, Theorem 1 shows that there are only four possible types of tuple sets for a given locus, up to the order of the alleles in a tuple and the allele numbers:

- a=4: $(1,2),(1,3),(4,2),(4,3)$, the two parents being $(1,4)$ and $(2,3)$;
- a=3: $(1,1),(1,2),(3,1),(3,2)$, the two parents being $(1,3)$ and $(1,2)$;
- a=2: $(1,1),(1,2),(2,1),(2,2)$, the two parents both being $(1,2)$;
- a=1: trivial, all the tuples are $(1,1)$.

Theorem 1 provides an easy way to check whether a given set satisfies the 2-allele property. Its proof also provides an algorithm for reconstructing a possible parental assignment of alleles by reversing possible switches. However, in the case where the set does not satisfy the 2-allele property, it does not answer the question of which individuals should be removed from the set so that it does satisfy the property. We pose this question as an optimization problem MAX 2-ALLELE SET:
Given a set S of k l-tuples

$$\left\{\langle(a_{ij}, b_{ij})\rangle_{\substack{1\le i\le k\\1\le j\le l}}\right\}$$

that satisfies the 4-allele property

$$\forall j, \quad 1 \le j \le l \ \left|\cup_{i=1}^{k} \{a_{ij}, b_{ij}\}\right| \le 4,$$

what is the minimum number of elements that need to be removed from S so it would satisfy the 2-allele property:

$$\min_{I\subseteq[k]} |I| \text{ s. t. } \forall 1 \le j \le l \ \left|\cup_{i\in I} \{a_{ij}, b_{ij}\}\right| + R_j \le 4,$$

where R_j is the number of values x such that either $(x,x) \in \cup_{i\in I}\{(a_{ij}, b_{ij})\}$ or $|\{y : (x,y) \in \cup_{i\in I}\{(a_{ij}, b_{ij})\} \vee (y,x) \in \cup_{i\in I}\{(a_{ij}, b_{ij})\}\}| = 3$.

In general, however, we are interested in reconstructing consistent sibling groups. We approximate this goal by the following combinatorial optimization problem MINIMAL 2-ALLELE SET COVER:
Given a collection U of n l-tuples

$$\left\{A_i = \langle(a_{ij}, b_{ij})\rangle_{\substack{1\le i\le n\\1\le j\le l}}\right\}$$

find a minimum number of subsets $S_1, ..., S_m$ in U that satisfy the 2-allele property and whose union is U:

$$\min m$$

such that

$$\forall t, \quad 1 \le t \le m \;\; S_t \subseteq U, \; \cup_{i=1}^{m} S_t = U, \text{ and}$$

$$\forall j, \quad 1 \le j \le l \;\; |\cup_{A_i \in S_t} \{a_{ij}, b_{ij}\}| + R_j \le 4,$$

where R_j is the number of values x such that either $(x, x) \in \cup_{A_i \in S_t}\{(a_{ij}, b_{ij})\}$ or $|\{y : (x, y) \in \cup_{A_i \in S_t}\{(a_{ij}, b_{ij})\} \vee (y, x) \in \cup_{A_i \in S_t}\{(a_{ij}, b_{ij})\}\}| = 3$.

The following is a simple (although not the most efficient) algorithm to solve the MINIMAL 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.

The difficult case to eliminate is when there are exactly four alleles present in the set and there exist at least one allele that is present with three other alleles at a locus. Since the previous algorithm, while biologically consistent, is computationally expensive, we use the weaker but computationally cheaper 4-allele and homozygot constraints as a heuristic for the reconstruction of the sibling groups.

2 The 4-allele Set Cover

We use the 4-allele and the homozygot constraints and a reduction to a specific instance of MINIMUM SET COVER problem to identify sibling groups among a given group of juveniles. We may also impose the constraint that an individual may belong to only one full sibling group. While the relationships may be promiscuous and an individual may belong to more than one *half*-sibling group, he or she is nonetheless a member of only one *full*-sibling group. First, we define the MINIMAL 4-ALLELE SET COVER problem:
Given a collection U of n l-tuples

$$\left\{ A_i = \langle (a_{ij}, b_{ij}) \rangle_{\substack{1 \le i \le n \\ 1 \le j \le l}} \right\}$$

find a minimum number of subsets $S_1, ..., S_m$ in U that

1. the number of distinct alleles plus the number of homozygot alleles present in the set is at most 4 for every locus and
2. the union of the sets is U:

$$\min m$$

such that

$$\forall t, \quad 1 \le t \le m \ \forall j, \quad 1 \le j \le l \ |\cup_{A_i \in S_t} \{a_{ij}, b_{ij}\}| + |\cup_{A_i \in S_t} \{a_{ij} : a_{ij} = b_{ij}\}| \le 4.$$

and

$$S_t \subseteq U, \cup_{i=1}^{m} S_t = U,$$

The classical Minimum Set Cover 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:

$$\min_{I \subseteq [m]} |I| \text{ such that } \cup_{i \in I} S_i = U.$$

Minimum Set Cover problem is NP-hard, moreover, it is not approximable within $(1 - \epsilon)\ln n$ unless $NP \subseteq DTIME(n^{O(\log\log n)})$ [7]. A $(1 + \ln n)$-approximation algorithm from 1974 is due to Johnson [9]. Define the $n \times m$ elements-sets matrix A by:

$$a_{ij} = \begin{cases} 1 & \text{if } i \in S_j \\ 0 & \text{otherwise} \end{cases}$$

In our solution, we use the standard integer programming formulation of the Minimum Set Cover problem:

$$\begin{array}{ll} \min & \sum_{i=1}^{m} x_i \\ & \\ \text{s.t.} & Ax \ge \overline{1} \\ & x_i \in \{0, 1\} \end{array}$$

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 the union *with repetitions* of their allele sets. That is, S_{pq} is a collection of l loci where each locus is a join of alleles of the corresponding locus for p and q. If an allele appears in both individuals or is homozygot in one of the individuals then that allele will appear *twice* in the locus. This collection of sets represents the starting point for the potential sibling sets.

2. An individual belongs to a set S_{pq} if for each locus the set of the alleles of the individual for that locus is in the the corresponding locus set of S_{pq}, including the repetitions.

3. Find a Minimum Set Cover 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.

We now state the algorithm more formally:

```
Algorithm 4-alleleSets

FOR i = 0...n-1 DO
        FOR j = i+1...n-1 DO
                S_ij = ""
                FOR t = 0...k-1 DO
                        S_ij = S_ij({a_it, b_it} ∪ {a_jt, b_jt})
FOR p = 1...n^2 - n DO
        FOR i = 0...n-1 DO
                FOR t = 0...k-1 DO
                        IF (a_it ∉ S_p ∨ b_it ∉ S_p) THEN
                            break
                animal_i ∈ S_p
Find a minimum Set Cover S = {S_1, ..., S_p}
RETURN the p sibling groups defined by S.
```

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

Proof. Each set S_{pq} is defined by a locus-by-locus union of two individuals. Each animal can have at most two alleles per locus, therefore each set has no more than four alleles per locus and the four-allele property is preserved. Every four-allele combination possible for the collection of individuals is represented by some set, since there is a set that corresponds to every pair of individuals.

In any feasible solution the set cover contains all the individuals, therefore each animal belongs to some set induced by the set cover. Each set initially does not violate the 4-allele property therefore, the entire solution 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 extended partition distance [8] 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 a of alleles per locus (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 for the number of juveniles per female. 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 for the number of juveniles per female $j = 1, 2, 5, 10$.
- 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.

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 [8]. We use an extension of the clustering distance measure described in [8].

3.2 Distance Measure

In [8] Gusfield showed that the minimum partition distance is equivalent to the MAXIMUM ASSIGNMENT problem (maximum bipartite weighted matching), which is a well known linear programming problem [5, 11]:

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

Given two collections of items $\{A_1, ..., A_n\}$ and $\{B_1, ..., B_m\}$, let the the c_{ij} in the $n \times m$ cost matrix C be the cost of the assignment of the item A_i to the item B_j. Then the Maximum Liner Assignment is an assignment of the A items to the B items of maximum cost so that each A item is assigned to at most one B item and vice versa:

$$\begin{array}{ll} \max & \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} \\ & \\ \text{s.t.} & \forall i \sum_{j=1}^{m} x_{ij} \leq 1 \\ & \forall j \sum_{i=1}^{n} x_{ij} \leq 1 \\ & x_{ij} \in \{0, 1\} \end{array}$$

It is not difficult to see that given two partitions of elements in U, $\{P_1, ..., P_n\}$ and $\{Q_1, ..., Q_m\}$ we can define $C_{ij} = |P_i \cap Q_j|$. Then the minimum number of elements that need to be deleted so that the two partitions become identical is $|U|-$(maximum assignment). Gusfield uses that number of elements to be deleted as the distance between two partitions.

4 Experiment Results

Here we present the results of the experiments with the simulated data. As stated earlier, the goal of these experiments is to assess the accuracy of the 4-alleleSets algorithm. We measure 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. Figure 1 shows selected corresponding graphs. These are representative of the data. As expected, the error increases with the number of juveniles and decreases as the number of offspring increases. Surprisingly, however, 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 and the algorithm assumes all the juveniles are siblings. At a maximum family size of 10 the algorithm reasonably well reconstructs the family groups based on only two loci and at least 5 alleles per locus (see Figure 1a). It is important to point out that in most cases, the algorithm found fewer sibling groups than there are in the population, merging the true families into a reconstructed one. This leads us to believe that the stronger algorithm 2-alleleSets will have more discriminating power to distinguish between these groups and will be more accurate.

5 Future Work and Extensions

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

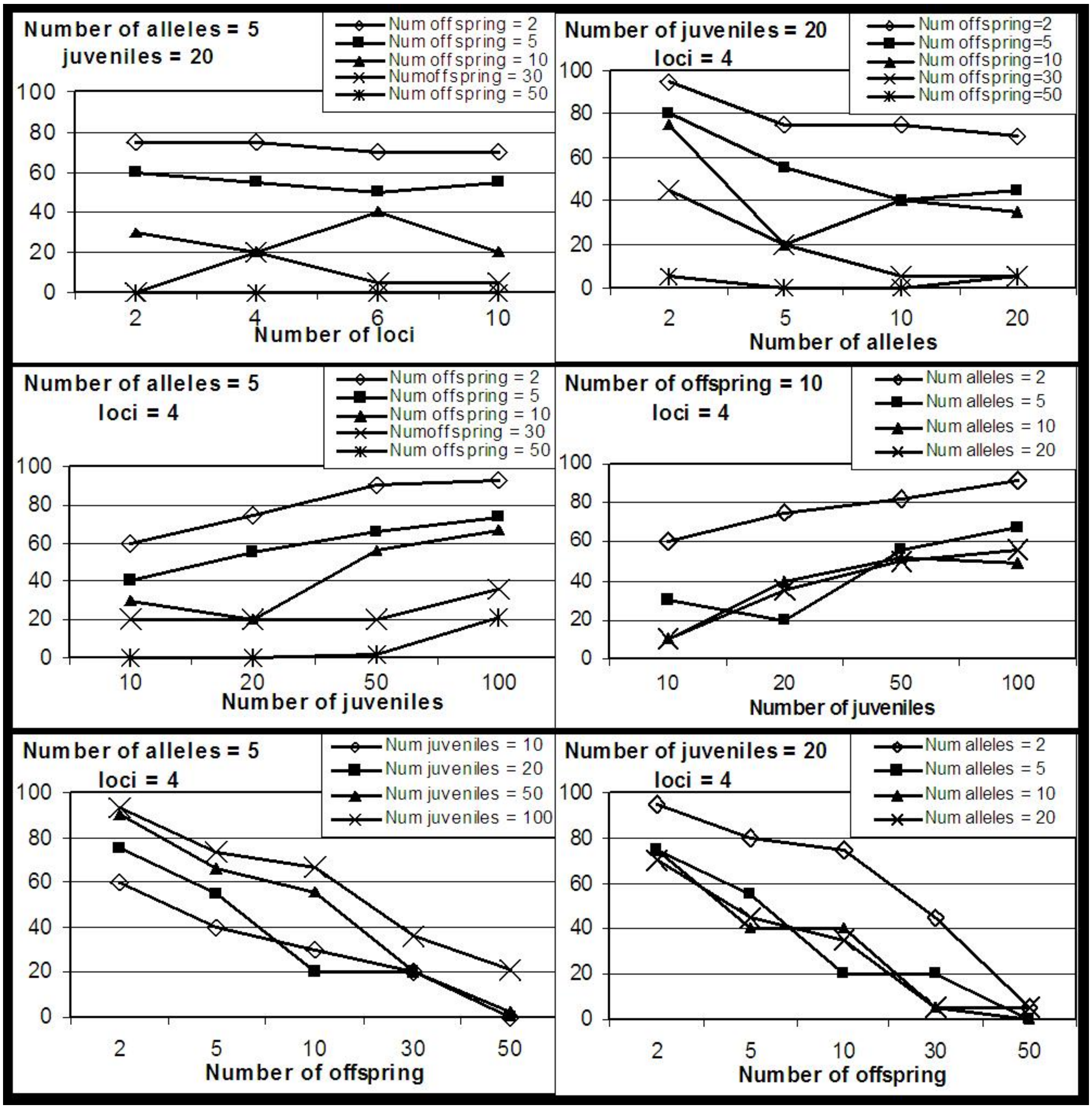

Figure 1: 4-allele algorithm error rate as a function of the number of loci, alleles per locus, juveniles, and maximum offspring per couple. The y-axis is the error rate, measured as the percent of the number of juveniles.

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: National Science Foundation grant EIA 02-05116 (Tanya Berger-Wolf), CCR-0206795, CCR-0208749 and IIS-0346973(Bhaskar DasGupta).

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] 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.

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

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

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

[7] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45:634–652, 1998.

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

[9] D. S. Johnson. Approximation algorithms for combinatorial problems. *J. Comput. System Sci.*, 9:256–278, 1974.

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

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

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

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

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