

## COMBINATORIAL RECONSTRUCTION OF HALF-SIBLING GROUPS FROM MICROSATELLITE DATA

Saad I. Sheikh \*, Tanya Y. Berger-Wolf, Ashfaq A. Khokhar  
*Department of Computer Science, University of Illinois at Chicago,  
851 S. Morgan St (M/C 152), Chicago, IL 60607, USA.  
{\*ssheikh,tanyabw,ashfaq}@cs.uic.edu*

Isabel C. Caballero, Mary V. Ashley,  
*Department of Biological Sciences, University of Illinois at Chicago,  
SEL 1031 M/C 067, 840 West Taylor Street, Chicago, IL 60607, USA.  
icabal2@uic.edu, ashley@eeb.uic.edu*

Wanpracha Chaovalitwongse, Chun-An Chou  
*Department of Industrial Engineering, Rutgers University,  
CoRE Building, 96 Frelinghuysen Rd., Piscataway, NJ 08854, USA.  
{wchaoval,joezhou}@rci.rutgers.edu*

Bhaskar DasGupta  
*Department of Computer Science, University of Illinois at Chicago,  
851 S. Morgan St (M/C 152), Chicago, IL 60607, USA.  
dasgupta@cs.uic.edu*

While full sibling group reconstruction from microsatellite data is a well studied problem, reconstruction of half sibling groups is much less studied, theoretically challenging, and a computationally demanding problem. In this paper, we present a formulation of the half-sibling reconstruction problem and prove it APX-hardness. We also present exact solutions for this formulation and develop heuristics. Using biological and synthetic datasets we present experimental results and compare them with the leading alternative software COLONY. We show that our results are competitive and allow half-sibling group reconstruction in the presence of polygamy, which is prevalent in nature.

### 1. INTRODUCTION

Kinship analysis is an important and necessary component of understanding an organism's biology and ecology. Biologists want to find out more about how organisms survive, acquire mates, reproduce, and disperse to new populations. Such information is difficult or impossible to infer from visual observation, and the establishment of kinship patterns (parentage or sibling relationships, for example) can be extremely useful. Genetic information has been used for decades by biologists to

\*Corresponding author.

make ecological and biological observations<sup>7,16</sup>. While a number of genetic markers have been introduced over the years, microsatellites (also known as STRs or SSRs) are most commonly used for wild species. Recent advances in technology have made it significantly cheaper to genotype individuals, creating a need for the techniques to process this data.

Several studies<sup>1,2,6,17,23,24</sup> including ours<sup>4,5,10,11,18,19</sup>, have recently developed computational approaches to reconstruct full-sibling groups in wild populations using genetic markers such as microsatellites.

Knowledge of the relatedness of individuals can be used to assess fecundity and mating systems, study kin selection, detect inbreeding, and to infer heritability using quantitative genetics<sup>9</sup>. While full sibling relatedness is difficult to infer, half-sibling relatedness constitutes a looser constraint on individual groupings which carries a weaker information signal and, thus, is even more difficult to reconstruct. Furthermore, monogamy, which produces only full-sibling groups, is relatively rare in nature. More common are polygamous and promiscuous mating systems where most offspring will be half-siblings (sharing only one parent), or a combination of half-sibling and full-sibling (sharing both parents) groups. Because of the ubiquity of half-sibling groups in nature, biologists need robust approaches for inferring half-sibling relationships from molecular marker data. For example, plants have flowers pollinated from many different plants, so seeds from a single plant are primarily half-siblings. Identifying these half-siblings among seedlings would allow researchers to study variation in female reproductive success among plants.

Few methods focus on half-sibling relationships, and most tend to make assumptions, e.g. monogamy<sup>24</sup>, which may not hold in practice. In Ref. 4 we formulated a parsimony-based approach aimed at minimizing the number of full-sibling groups necessary to explain the given population. In Refs. 19, 5 we presented the 2-ALLELE MIN SET COVER algorithm for achieving this objective and showed that it worked extremely well, even when the allelic information was low and outperformed other algorithms when the number of loci, or alleles per locus was low. In Ref. 18 we presented a consensus-based approach which aimed at minimizing the number of sibling groups and also exploited any available genetic information. Our approaches were designed for studies of wild populations where number of loci and alleles per locus is low (e.g. 3 loci for 1000 individuals with 10 alleles per locus) and no assumptions can be made about the underlying mating system.

In this paper, we focus on the half-sibling reconstruction problem using genetic data from cohorts of offspring rather than the breeders, since they are usually easier to sample in wild populations. The problem is not only harder to analyze theoretically, it is also more difficult to solve computationally. Our main contributions in this paper are as follows: 1) we formally define the half-sibling reconstruction problem and analyze its combinatorial properties; 2) we present a new parsimony-based formulation for the half-sibling reconstruction problem and show that it is APX-hard; 3) we develop exact algorithms for solving this hard combinatorial for-

mulation; 4) we test these algorithms using both biological and simulated datasets and compare our reconstruction results to those obtained by the leading alternative approach COLONY<sup>24</sup>.

## 2. HALF-SIBLING RECONSTRUCTION

We first define some biological concepts for the benefit of the reader not familiar with these concepts. Readers familiar with these concepts may safely skip to Section 2.3.

### 2.1. *Biological Preliminaries*

Full- and half-siblings: a group of individuals that share both parents is referred to as *full siblings*, and when they share one of the parents they are referred to as *half siblings*. In the rest of the paper, we use full-sibs and half-sibs terms to refer to these groups, respectively.

Locus: the location of a gene on a chromosome.

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

Genetic marker: a segment of DNA that can be scored to identify individual genotypes and track inheritance.

Diploid individual is one having two alleles (not necessarily different) at each locus.

Allele frequency: the fraction of all the alleles for a gene in a population that are of a particular type.

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

### 2.2. *Microsatellite Markers*

While there are several molecular markers used in population genetics, microsatellites (also known as SSRs, STRs, SSLPs, and VNTRs) are the most commonly used markers in population biology for non-model organisms. Microsatellites are repeats of short DNA sequences distributed throughout the genome. These are co-dominant, unlinked, multi-allelic markers that offer numerous advantages for population studies. Generally, phase or haplotype information is not available for microsatellite loci in non-model organisms.

### 2.3. *Problem Statement*

The main focus of our paper is to design a method that accurately reconstructs half-sibling groups from microsatellite data. Table 1 shows an example cohort with five individuals sampled at two loci. We now formally define the problem of half-sibling reconstruction. Let  $U = \{X_1, \dots, X_n\}$ , where  $U$  is a population of  $n$  diploid individuals of the same generation, and where each individual is represented by

Table 1: Example of a cohort of five individuals sampled at two microsatellite loci with a unique full-sibling and multiple half-sibling solutions.

(a) Parental genotypes (F1)					(b) Full-sibling groups in children (F2)		
Id	Locus 1		Locus 2		Father	Mother	Offspring ids
<b>1</b>	7	8	19	20	P1	P2	1, 2
<b>2</b>	7	10	20	46	P1	P4	4, 5
<b>3</b>	5	6	19	23	P3	P2	7, 8
<b>4</b>	4	5	15	19	P3	P4	10, 11
<b>5</b>	2	10	15	19	P5	P6	13, 14
					P5	P8	15, 16
					P7	P6	17, 18
					P7	P8	19, 20

(c) *Biologically consistent* half-sibling reconstructions shown as sets of ids of offspring

$\{\{1, 2, 4, 5\}, \{7, 8, 10, 11\}\{13, 14, 15, 16\}\{17, 18, 19, 20\}\}$   
 $\{\{1, 2, 7, 8\}, \{4, 5, 10, 11\}\{13, 14, 17, 18\}\{15, 16, 19, 20\}\}$   
 $\{\{1, 2, 7, 8\}, \{4, 5, 10, 11\}\{13, 14, 15, 16\}\{17, 18, 19, 20\}\}$   
 $\{\{1, 2, 4, 5\}, \{7, 8, 10, 11\}\{13, 14, 17, 18\}\{15, 16, 19, 20\}\}$

a genetic (microsatellite) sample at  $l$  loci. That is,  $X_i = (\langle a_{i1}, b_{i1} \rangle, \dots, \langle a_{il}, b_{il} \rangle)$  and  $a_{il}$  and  $b_{il}$  are the two alleles of the individual  $i$  at locus  $l$  represented as some identifying sequence. The goal is to reconstruct half-sibling groups which is formulated as a cover<sup>a</sup> of individuals by sets  $P_1, \dots, P_m$  where individuals in the same set  $P_i$  share at least one parent. We assume no knowledge of parental information.

What complicates the half-sibling reconstruction problem is the existence of multiple half-sibling reconstructions for a given cohort. Consider the cohort of individuals in Table 1(b). The full-sibling reconstruction is clear and there is only one correct answer. However, for the same cohort, there are four different possible half-sibling reconstructions, as shown in Table 1(c). Each of these reconstructions is biologically plausible, i.e. individuals placed in a half-sibling group share exactly one parent. Every individual, and the full-sibling group it belongs to, is always in the intersection of two half-sibling groups.

<sup>a</sup>a collection of not necessarily disjoint groupings

## 2.4. Related Work

COLONY<sup>24</sup> is a widely used software for both full- and half-sibling groups reconstruction. However, it assumes that one gender mates monogamously, an assumption that may greatly limit the software's utility. In a recent paper<sup>25</sup> the problem of monogamy has been resolved, however the work is still based on simulated annealing, and therefore take days to run. Moreover, COLONY<sup>24,25</sup>, Queller et al.<sup>17</sup>, Konovalov et al.<sup>15</sup>, Almudevar *et al.*<sup>1,2</sup>, Herbingner et al.<sup>13</sup>, Wilson *et al.*<sup>26</sup>, Thomas *et al.*<sup>22</sup> all use likelihood-based approaches to reconstructing both full- and half-sibling groups. All of these approaches assume knowledge or availability of population allele frequencies or mating patterns in the given species.

## 2.5. Half-Sibs Property

In order to present a combinatorial approach based purely on parsimony we first need to translate the Mendelian genetic laws into combinatorial constraints that all half-sibling groups must obey. In Ref. 5 we presented two necessary combinatorial properties that a full-sibling group must satisfy: the 2-ALLELE property and the 4-ALLELE property. We now present a combinatorial property based on Mendelian laws that a half-sibling group must obey. This is a necessary, yet not sufficient, property for any group individuals to be a feasible half-sibling group.

**HALF-SIBS PROPERTY:** For any given half-sibling group, at every locus there exists a pair of alleles  $x_j, y_j$  such that every individual in the group contains (at least) one of the two alleles. Formally, a set  $S \subseteq U$  has the HALF-SIBS PROPERTY if

$$\begin{aligned} \forall 1 \leq j \leq l: \quad & \exists \mathcal{A}_j = \{x_j, y_j\} \\ \text{s.t. } \forall i \in S \quad & a_{ij} \in \mathcal{A}_j \vee b_{ij} \in \mathcal{A}_j \end{aligned}$$

*Proof.*

By Mendelian law, two parents with loci  $\{p, q\}$  and  $\{r, s\}$  produces an offspring  $\{a, b\}$  if and only if  $|\{a, b\} \cap \{p, q\}| = 1$  **and**  $|\{a, b\} \cap \{r, s\}| = 1$ . The claim follows since a half-sibling group has at least one parent in common.  $\square$

This property is illustrated in Table 1: the first four individuals can be members of a half-sibling group because the alleles  $\{5, 7\}$  at the first locus and  $\{19, 20\}$  at the second locus satisfy the HALF-SIBS PROPERTY. Individual 5 cannot be added to this half-sibling group because there will be no set of two alleles at the first locus which will cover all five individuals.

Notice that there is no limit on the actual number of different alleles in a half-sibling group (other than the trivial  $2 + n$ ). The HALF-SIBS PROPERTY constraint is mathematically weak: for any half-sibling group that obeys this property a viable parent genotype can be constructed by using the two alleles at every locus. Furthermore, any two individuals can potentially be half-siblings. In practice, we may

also require that any individual or full-sibling group may be part of at most two half-sibling groups (one for each parent).

### 3. PARSIMONY-BASED HALF-SIBS RECONSTRUCTION

In formulating the computational problem of reconstructing half-sibling group, we choose the objective of maximum parsimony, rather than the statistical model fitting of the majority of other kinship reconstruction approaches. Our approach avoids making unnecessary assumptions about the sampled population and the genetic distributions within. The fundamental assumption we do make is that of the Mendelian genetic laws.

#### 3.1. Min-Half-Sibs *Problem Definition*

Given the assumption of Mendelian genetic laws, we find the most parsimonious collection of half-sibling groups that explain the genetics of a sampled cohort of individuals. While there are many ways to interpret the parsimony objective, in absence of other information about the population we start with the simplest: finding the *minimum* number of half-sibling groups (that obey the HALF-SIBS PROPERTY) to explain the genetics of the sample.

**Input:** A set  $U$  of  $n$  individuals, each with  $\ell$  sampled loci.

**Notation:** Let  $h_i \subseteq U$  denote a set of individuals which obey the HALF-SIBS PROPERTY.

**Valid Solutions:**  $H = \{h_0, \dots, h_m\}$  s.t.  $\cup_{h_i \in H} h_i = U$ .

**Objective:** *minimize*  $|H|$ .

#### 3.2. Min-Half-Sibs *Integer Linear Programming Formulation*

We now present an ILP formulation of optimization model to directly solve the MIN-HALF-SIBS formulation. This model is also based on HALF-SIBS PROPERTY. We first define the following decision variables:

- $x_{ij} \in \{0, 1\}$ : indicates if individual  $i$  is selected to be a member of the current half-sibling group  $j$ ;
- $z_j \in \{0, 1\}$ : indicates if the group  $j$  is non-empty;
- $w_{jk}^l \in \{0, 1\}$ : indicates if allele  $k$  appears in the current half-sibling group  $j$  at locus  $l$ .
- $a_{ik}^l \in \{0, 1, 2\}$ : counts the number of times allele  $k$  appears in individual  $i$  at locus  $l$ .

The optimization model for the half-sibling problem is formulated as follows:

HALF-SIBS MODEL:

$$\min \sum_{j \in J} z_j \quad (1)$$

$$\text{s.t. } x_{ij} \leq z_j \quad \forall i \in I, j \in J \quad (2)$$

$$\sum_{j \in J} x_{ij} \geq 1 \quad \forall i \in I \quad (3)$$

$$\sum_{k \in K} a_{ik}^l w_{jk}^l \geq x_{ij} \quad \forall i \in I, j \in J, l \in L \quad (4)$$

$$\sum_{k \in K} w_{jk}^l \leq 2 \quad \forall j \in J, l \in L. \quad (5)$$

Equation (1) is the objective to minimize the number of maximal feasible half-sibling groups. Equations (2) and (3) are the logical constraints that any individual  $i$  has to be assigned to at least one half-sibling set  $j$ . The Equations (4) and (5) are the HALF-SIBLINGS constraints that for any half-sibling group there is no more than two alleles appearing at each locus.

Note that because the half-sibling groups  $J$  is not defined yet in HALF-SIBS MODEL, an initial number of half-sibling groups is required before implementation. It is necessary to define a reasonable initial number for the half-sibling groups. Given a larger number, although it is guaranteed to solve to obtain feasible, even optimal, half-sibling groups to explain an input population, the implementation in CPLEX will be very expensive in terms of computational time. On the other hand, the smaller number may lead to infeasibility. That is, the resulting half-sibling sets cannot be enough to explain an input population.

### 3.3. Computational Complexity

We first recall some standard definitions. A  $(1 + \varepsilon)$ -approximate solution (or simply an  $(1 + \varepsilon)$ -approximation) of a minimization problem is a solution with an objective value no larger than  $1 + \varepsilon$  times the value of the optimum, and an algorithm achieving such a solution is said to have an *approximation ratio* of at most  $1 + \varepsilon$ . A problem is APX-hard if, for some constant  $\varepsilon > 0$ , the problem has no  $(1 + \varepsilon)$ -approximation under a standard complexity-theoretic assumption such as  $P \neq NP$  or  $RP \neq NP$ . Note that APX-hardness is a stronger notion than NP-hardness since it rules out arbitrarily good approximation of the problem.

**Theorem 1.** MIN-HALF-SIBS is APX-hard under the assumption of  $RP \neq NP$ .

*Proof.* We first need the *triangle-packing* (TP) problem which is defined as follows. We are given an undirected graph  $G$ . A triangle is a cycle of 3 nodes. The goal is to find (pack) a maximum number of *node-disjoint* triangles in  $G$ . The following result was shown in Ref. 3.

**Theorem 2.** Assuming  $RP \neq NP$ , given an instance  $G$  of TP with  $228n$  vertices there is no polynomial time algorithm that can decide, for any sufficiently small

constant  $\varepsilon > 0$ , if there is a solution of size at least  $(76 - \varepsilon)n$  of if all solutions are of size at least  $(75 + \varepsilon)n$ .<sup>3</sup>

The above theorem there shows that TP has no  $(1 + \delta)$ -approximation for  $\delta = (76/75) - \varepsilon$  under the assumption of  $\text{RP} \neq \text{NP}$ .

We now reduce an instance of TP to an instance of MIN-HALF-SIBS. For notational convenience, let the vertex set of  $G$  be  $V = \{1, 2, \dots, 228n\}$ . For every vertex  $j \in V$ , there is a ‘‘corresponding’’ individual  $j'$  in MIN-HALF-SIBS. We will now add appropriate loci to ensure that:

- (1) three individuals corresponding to a triangle of  $G$  are a possible half-sibling, and
- (2) three or more individuals that do not correspond to a triangle of  $G$  cannot be half-siblings.

Obviously any pair of individuals can be a half-sibling group. Thus, the above properties ensure that TP has a maximal solution with  $t$  triangles if and only if the instance of MIN-HALF-SIBS has  $t + \frac{228n - 3t}{2} = \frac{228n - t}{2}$  half-sibling groups. Using Theorem 2, it follows that given an instance of MIN-HALF-SIBS with  $228n$  individuals, there is no polynomial-time algorithm that can decide if the instance has a solution with at most  $\frac{228n - 75n - \varepsilon}{2} = \frac{86.5n - \varepsilon}{2}$  half-sibling groups or if all solutions must have at least  $\frac{228n - 76n + \varepsilon}{2} = \frac{86n + \varepsilon}{2}$  half-sibling groups. This shows that MIN-HALF-SIBS has no  $(1 + \delta)$ -approximation with  $\delta = \frac{86.5}{86} - \varepsilon$ .

We now show how to add loci to satisfy Properties (1) and (2). For each case, it is straightforward to use the HALF-SIBS PROPERTY discussed in the previous section to ensure that the construction is correct.

**Ruling out half-sibling groups of size 4 or more:** We ensure that no set of four or more individuals can be half-siblings. Note that it suffices to rule out *all* sets of four individuals only. There are  $\binom{n}{4} = \Theta(n^4)$  such loci, each representing each set of four individuals. Consider a set of four individuals  $i', j', k', \ell'$  corresponding to the vertices  $i, j, k, \ell$  of  $G$ . We will introduce a new loci that will disallow the individuals  $\{i', j', k', \ell'\}$  to be half-siblings, but will *not* disallow any other combinations. We insert a new locus  $t$  with six new alleles  $\langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle$  for these individuals:  $i'_t = \{x_1, x_2\}$ ,  $k'_t = \{x_3, x_4\}$ ,  $k'_t = \{x_5, x_2\}$ ,  $\ell'_t = \{x_5, x_6\}$ , and  $y'_t = \{x_1, x_5\} \quad \forall y \in V \setminus \{i, j, k, \ell\}$ .

**Ruling out non-triangles:** We next ensure that only triplets corresponding to triangles in  $G$  can be half-siblings. We introduce  $\binom{n}{3} - t = O(n^3)$  new loci, each representing a set of three elements  $i', j', k'$  such that  $i, j, k$  is *not* a triangle in  $G$ . The loci for this set of individuals will prohibit the corresponding individuals  $\{i', j', k'\}$  to be half-siblings, but all other combinations of individuals will not be affected. We insert a new locus  $t$  with six



new alleles  $\langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle$  for these individuals:  $i'_t = \{x_1, x_2\}$ ,  $j'_t = \{x_3, x_4\}$ ,  $k'_t = \{x_5, x_6\}$ , and  $y_t = \{x_1, x_5\} \quad \forall y \in V \setminus \{i, j, k\}$ . This allows any set of three individuals, other than  $\{i, j, k\}$ , to be half-siblings.

We note that all individuals are distinct, *i.e.*, for every two individuals  $i'$  and  $j'$  there is a locus in which they differ. This is because there exists a set of four individuals in which  $i'$  belongs but  $j'$  does not and the first set of loci introduced above ensures that they have non-identical loci.

It is clear from the construction of gadgets that Properties (1) and (2) are satisfied. This completes the proof.  $\square$

Since the problem is hard to approximate, providing an exact algorithm would be intractable too, and may not be meaningful biologically. Therefore, in spite of the apparent complexity problems, we proceed to present exact algorithms for this ILP formulation.

### 3.4. Half-Sibs Min Set Cover *Algorithm*

We now present an *exact* algorithm to solve the MIN HALF-SIBS problem. This algorithm is similar to the 2-ALLELE MIN SET COVER algorithm we presented in Ref. 5. It consists of two stages:

- (1) Enumerate all maximal feasible half-sibling sets  $S$  in the cohort  $U$  that obey the HALF-SIBS PROPERTY.
- (2) Find the minimum number of maximal feasible sets  $C \subseteq S$  necessary to cover the entire cohort  $U$  using the Minimum Set Cover<sup>14</sup>.

#### 3.4.1. Step 1: Half-Sibling groups Enumeration Algorithm.

In order to generate all maximal half-sibling groups we exploit the fact that a set of any two alleles at a locus represents a potential parent. We first generate all maximal feasible half-sibling groups at each locus, and then intersect them to find groups that are common across loci. In order to generate maximal feasible half-sibling groups we treat every pair of alleles present in a locus as the parental genotype (for that locus) and then check which individuals inherit at least one allele. We refer to Figure 1 for details.

**Lemma 1.** *Algorithm HALF-SIBS ENUMERATION generates all maximal half-sibling groups.*

The proof is straightforward and we omit it for brevity.

This algorithm implies a straightforward upper bound on the number of half-sibling groups in a given cohort:  $O\left(\binom{2n}{2}^k\right) = O(n^{2k})$  where  $k$  is the number of loci is assumed to be a given constant. Compared to the full-sibling reconstruction problem, this tremendously increases the size of the set cover problem. However, we are able to execute this algorithm on most of the test data sets. For larger

```

input :  $U$ : individuals
output:  $\mathcal{H}$ : Set of Maximal Half-sibling groups
 $HalfSibs \leftarrow \{U\}$ ;
foreach locus  $l$  do
   $HalfSibs[l] \leftarrow \emptyset$ ;
   $Alleles[l] \leftarrow \{a \mid \text{allele } a \text{ appears at locus } l\}$ ;
  foreach  $a \in Alleles[l]$  do
     $AlleleSets[l][a] \leftarrow \{I_x \mid \text{Individual with allele } a \text{ at locus } l\}$ ;
  end
  foreach  $a_1, a_2 \in Alleles[l]$  do
     $halfsib_{a_1, a_2} \leftarrow AlleleSets[l][a_1] \cup AlleleSets[l][a_2]$ ;
     $HalfSibs[l] \leftarrow HalfSibs[l] \cup \{halfsib_{a_1, a_2}\}$ ;
  end
 $HalfSibs \leftarrow IntersectGroups(Halfsibs, Halfsibs[l])$ ;
end

```

Fig. 1: Algorithm for generating all maximal feasible half-sibling groups.

```

input :  $U$ : set of individuals,  $F$ : set of full-sib groups
output:  $H$  set of feasible half-sibling groups
 $H \leftarrow F$ ;
 $merging \leftarrow \text{true}$ ;
while  $merging$  do
   $merging \leftarrow \text{false}$ ;
  foreach  $S_i \in H$  do
    foreach  $S_j \in F$  do
       $S_{i,j} \leftarrow S_i \cup S_j$ ;
      if  $S_{i,j}$  obeys HALF-SIBS PROPERTY  $\wedge S_{i,j} \notin H$  then
         $merging \leftarrow \text{true}$ ;
         $H \leftarrow H \cup \{S_{i,j}\}$ ;
      end
    end
  end
end

```

Fig. 2: MIN FULL-/HALF-SIBLING Algorithm

data sets it is possible to prune the sets of individuals at each locus by discarding non-maximal sets.

### 3.4.2. Step 2: Min Set Cover.

The minimum set cover problem is a classical NP-complete<sup>14</sup> problem and is defined as follows: *given a universe  $U$  of elements  $X_1, \dots, X_n$  and a collection of subsets  $\mathcal{S}$  of  $U$ , the goal is to find the minimum collection of subsets  $C \subseteq \mathcal{S}$  whose union is the entire universe  $U$ .*

We use the standard integer linear program formulation of the Minimum Set Cover problem to solve it to optimality using commercial ILP solver CPLEX<sup>b</sup>.

## 4. VALIDATION METHODOLOGY

In order to set a baseline of results for comparison, we use a trivial approach that is based on the full sibling reconstruction approach presented in Ref. 5.

### 4.1. Half-sibling reconstruction from full-sibling reconstruction

Another way to interpret the parsimony objective for the half-sibling reconstruction problem is to find a reconstruction that minimizes the number of both full- and half-sibling groups. We implement this approach by first finding the minimum number of full-sibling groups necessary to explain the population using the 2-ALLELE MIN SET COVER algorithm<sup>5</sup> and then merging the full-sibling groups to obtain the minimum half-sibling groups that cover the population and are composed of full-sibling groups.

In order to determine the *minimum* number of half-sibling groups based on a full-sibling reconstruction solution we explore all possible half-sibling groups that can be generated from the given full-sibling reconstruction. The algorithm works in three steps in a similar fashion as the algorithm presented above.

#### MIN FULL-/HALF-SIBLING Algorithm

- (1) Generate a full-sibling reconstruction  $F$  using the 2-ALLELE MIN SET COVER algorithm<sup>5</sup>.
- (2) Enumerate all maximal feasible half-sibling sets  $S$  in the cohort  $U$  that obey the HALF-SIBS PROPERTY and can be obtained by merging a subset of the input full-sibling groups in  $F$ . We start generating candidate half-sibling groups by merging all pairs of full-sibling groups. We then compare each full-sibling group to these candidate half-sibling groups to determine whether additional merges can be made conforming to the HALF-SIBS PROPERTY.
- (3) Find the minimum number of maximal feasible sets  $C \subseteq S$  necessary to cover the entire cohort  $U$  using the Minimum Set Cover.

Trivially, the number of maximal feasible half-sibling groups is  $O(n^{2k})$ , where  $k$  is the number of loci. It seems theoretically that we first create a relatively large input to an NP-hard problem, thus creating an impossibly large problem. However, in practice, we are able to solve most datasets in a few hours.

<sup>b</sup>CPLEX is a registered trademark of ILOG

## 4.2. Datasets

To validate and assess the accuracy of our approach, we have used datasets with known genetics and genealogy that contain half-sibling groups. However, such biological datasets containing no errors are few and we were able to obtain only two. Therefore, we test on both biological and simulated datasets.

### *Biological Datasets*

We test our approach on datasets where offspring were collected and genotyped at several microsatellite loci. Half-sibling groups were known because the offspring were collected from individual gravid females, and were thus maternally related half-siblings. As discussed above, there may be multiple correct solutions, but these datasets typically are based on configurations where the ratio of the number of fathers to the number of mothers is high, which, as we found, aids tractability of the problem.

**Cricket:** The field cricket *Grillus bimaculatus* dataset comes from a population of crickets studied in Spain<sup>8</sup>. It consists of 112 individuals from 7 wild-caught gravid females with 6 sampled loci.

**Rockfish Larvae:** The kelp rockfish *Sebastes atrovirens* dataset<sup>21</sup> consists of 672 larvae from 7 broods and 7 sampled loci. A subset consisting of 288 larvae from the first 3 broods was used due to computational inefficiencies.

### *Simulated Datasets*

To validate our approach using simulated data, we follow the same protocol as in Ref. 4. We first create random diploid parents and then generate complete genetic data for offspring varying the number of males, females, alleles, loci, number of offspring and juveniles. For a given number of females, males, loci, and a number of alleles per locus, we generate a set of diploid parents with independent identical uniform distribution of alleles in each locus. A male and a female are chosen independently, randomly, and uniformly from the parent population. For these parents, a specified number of offspring is generated. Each offspring randomly receives one allele each from its mother and father at each locus. While this is a rather simplistic approach, it is consistent with the genetics of known parents and provides a baseline for the accuracy of the algorithm since biological data are generally not uniformly random.

The simulated datasets were generated to show the effects of a degree of disproportion between the number of mothers and fathers in the breeding pairs. We used the following ratios of the number of fathers to the number of mothers<sup>c</sup>: 1:10, 1:5. The half-sibling groups based on the sex with the smaller number of breeding adults

<sup>c</sup>The genders are symmetric and the results hold for a high ratio of fathers to mothers

were chosen as the ground truth, i.e. paternal groups. We generated 10 cohorts for each set of parameters.

### 4.3. Accuracy

There is no well-accepted measure of comparing half-sibships. Moreover, as discussed above, the task is complicated by the fact that some half-sibling groups may overlap multiple times and it is not clear whether the overlap should be penalized. The absence of parental information in biological datasets implies that we cannot be sure that some half-sibling groups given by the algorithm are not representative of the half-sibling groups by other sex. We measure the error rates of algorithms using a slight modification of the *partition distance* by Gusfield<sup>12</sup>, which is the smallest number of individuals that need to be removed from the population to make two partitions equivalent. For the cases where overlap occurs we assume that the right assignment was made as long as one of the overlapping assignments is correct. For biological datasets we also report the overlap in addition to this score.

### 4.4. Half-Sibs Model *configuration*

We tested the HALF-SIBS MODEL on cricket and rockfish datasets in MATLAB with the use of the callable GAMS library with CPLEX version 10.0. The execution time limit was set to be 7 days maximum. The process was set to terminate in advance when the solution gap is less than 0.01%. We offered 20 and 10 as the initial numbers of half-sibling set for crickets and rockfish, respectively.

## 5. RESULTS

### 5.1. Half-Sibs Model *Computational Limitations*

For both datasets the HALF-SIBS MODEL was unable to prove the optimality of the solution it found in the allocated 7 days. We report the results that were obtained at the end of the period. This also shows that our HALF-SIBS MIN SET COVER approach is very efficient as it was able to solve both of the instances in less than a day.

### 5.2. *Biological Datasets*

#### *Crickets*

Our HALF-SIBS MIN SET COVER approach gives good results, the only difference with the ground truth is that two of the elements are assigned to more than one half-sibling groups. The MIN FULL-/HALF-SIBLING solution classifies 20 out of 111 individuals incorrectly. COLONY produces an accurate result. See Table 2 for details. Note that COLONY does not allow overlap between half-sibling groups because it assumes that one of the sexes is monogamous.

Table 2: Half sibling groups obtained for Crickets using four different methods ( $n = 112$ ).

(a) Original	(b) HALF-SIBS MIN SET COVER	(c) MIN FULL-/HALF-SIBLING
Set(1) 0 - 15	Set(1): 0 - 15	Set(1): 0 - 15 <u>33</u>
Set(2) 16 - 31	Set(2): 16 - 31 <b>110</b>	Set(2): <u>13 32 34 73 80 96 109</u>
Set(3) 32 - 47	Set(3): 32 - 47	Set(3): 16 - 31 <b>80 81 82 - 85 87 89 90 - 95</b>
Set(4) 48 - 63	Set(4): 48 - 63	Set(4): 35 - 47
Set(5) 64 - 79	Set(5): 64 - 79	Set(5): 48 - 63
Set(6) 80 - 95	Set(6): 80 - 95	Set(6): 64 - 72 74 - 79
Set(7) 96 - 111	Set(7): <b>73</b> 96 - 111	Set(7): 80 81 86 88 89 96 - 111

  

(d) COLONY	(e) HALF-SIBS MODEL ILP
Set(1): 0 - 15	Set(1): 0 - 15
Set(2): 16 - 31	Set(2): 16-31 <u>110</u>
Set(3): 32 - 47	Set(3): 32 - 47
Set(4): 48 - 63	Set(4): 48 - 63 <b>44 96</b>
Set(5): 64 - 79	Set(5): 62 - 79 <u>102</u>
Set(6): 80 - 95	Set(6): 97 - 101 103 - 109 <u>111</u>
Set(7): 96 - 111	Set(7): 80 - 95

**Rockfish Larvae Subset**

Three approaches: HALF-SIBS MIN SET COVER , MIN FULL-/HALF-SIBLING, and COLONY - produced 100% accurate assignments. See Table 3 for details. HALF-SIBS MIN SET COVER solution had an overlap of 4 out of 288 individuals.

**5.3. Simulated Datasets**

As expected, the ratio of the numbers of fathers to the number of mothers is the major factor in the accuracy of reconstruction. When the number of fathers and mothers are comparable, it is possible to pick many alternative parsimony-based reconstructions, thus the accuracy was low for such scenarios. Table 4 presents the results of the reconstruction by the three methods.

Table 3: Half sibling groups (sets) obtained by four different methods from a Rockfish Larva Subset ( $n=288$ ). Bold-face numbers and underlined numbers represent overlaps and misassignments, respectively.

(a) Original		(b) HALF-SIBS MIN SET COVER				(c) MIN FULL-/HALF-SIBLING		
Set(1)	0 – 95	Set(1)	0 – 95	<b>125</b>		Set(1)	0 – 95	
Set(2)	96 – 191	Set(2)	96 – 191			Set(2)	96 – 191	
Set(3)	192 – 287	Set(3)	<b>111</b>	<b>147</b>	<b>182</b>	192 – 287	Set(3)	192 – 287

  

(d) COLONY		(e) HALF-SIBS MODEL ILP	
Set(1)	0 – 95	Set(1):	0 - 95
Set(2)	96 – 191	Set(2):	96 - 191
Set(3)	192 – 287	Set(3):	192 - 287

Table 4: Accuracy results for MIN-HALF-SIBS, MIN FULL-/HALF-SIBLING and COLONY algorithms for the simulated datasets.

Fathers	Mothers	Loci	Alleles	Families	Offspring	MIN-HALF-SIBS		MIN FULL-/HALF-SIBLING		COLONY	
						$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$
2	20	6	5	40	2	100	0	66.3	13.99	96.15	10.1
2	20	6	10	40	2	100	0	47.5	7.07	99.8	1.99
2	20	10	10	20	2	100	0	60.45	15.48	99.9	0.99
2	10	6	10	2	10	80	24.49	80	24.49	90	20
2	10	6	15	2	5	70	24.49	70	24.49	75	25

$\mu$ : Mean of partition distance

$\sigma$ : Standard Deviation of partition distance

## 6. CONCLUSIONS

We have developed new intuitive formulations for reconstructing half-sibling relationships from microsatellite markers. We make no assumptions about the data or mating patterns other than parsimony and Mendelian genetics. We have also discussed the complexity of the proposed formulations and provided exact algorithms to solve these. Unfortunately, the resulting optimization problems are APX-hard and the approaches are computationally intense in practice.

We have presented both algorithmic and Integer Linear Programming solutions

to the problem. We showed that our HALF-SIBS MIN SET COVER algorithm provides a solution that is more efficient than the ILP when solved on a state-of-the-art ILP solver, CPLEX.

The HALF-SIBS MIN SET COVER method correctly reported all the half-sibling groups on biological data and on simulated data when the number of mothers or fathers was much larger than the other. While this approach is not very efficient, we are currently working on techniques to make this approach more efficient and scalable. Depending upon the number of alleles per locus and the number of loci, the HALF-SIBS MIN SET COVER approach tends to be as efficient as COLONY. However, the algorithm presented here can be easily parallelized by applying domain and functional decomposition using the methodology described in Ref. ?. While the MIN FULL-/HALF-SIBLING approach was not very accurate, it is more efficient as it explores a much smaller space of solutions.

As discussed in Ref. 5, for wild and endangered populations parsimony seems to be the only assumption we can make since any judgments about allele frequencies, mating patterns, and family sizes may be invalid. We argue that our methodology is superior as it gives accurate results without the assumptions made by other methods. We have avoided the assumption of monogamy and emphasized the problems it can raise. Our results (Table 4) show that our approach is capable of reconstructing half sibling groups in populations where neither of the sexes is monogamous.

Clearly, the proposed approaches, including COLONY, are not computationally scalable in practice. However, our work lays the foundation for understanding the computational structure of the half-sibling problem. We consider our methods as a starting point for developing viable practical solutions for half-sibship reconstruction. We have recently introduced the first high-performance approach to full sibling reconstruction<sup>?</sup>, which adapted to make the HALF-SIBS MIN SET COVER approach more scalable.

In the future, we intend to extend this work to handle data with genotyping errors using consensus methods, similar to our work for full-sibling groups<sup>18</sup>. We will also try to make detailed comparisons to the recent version of COLONY<sup>25</sup>. Furthermore, we will investigate whether the half-sibling group information obtained through the techniques presented in this paper can help understand other kinship relationships.

## ACKNOWLEDGMENTS

This research is supported by the following grants: NSF IIS-0612044 (Berger-Wolf, Ashley, Chaovalitwongse, DasGupta), Fulbright Scholarship (Saad Sheikh), NSF CAREER CCF-0546574 (Chaovalitwongse), NSF CAREER IIS-0747369 (Berger-Wolf), NSF DBI-0543365 (DasGupta) and NSF IIS-0346973 (DasGupta). We are grateful to the people who have shared their data with us: Amanda Bretman, University of East Anglia, Susan M. Sogard and Eric C. Anderson, National Marine Fisheries Service. We are also grateful to our collaborator Priya Govindan for her



support.

## References

1. A. Almudevar. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, 63, 2003.
2. 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.
3. M. Ashley, T. Berger-Wolf, P. Berman, W. Chaovalitwongse, B. DasGupta, and M.Y. Kao. On approximating four covering and packing problems. *Journal of Computer and System Science*, 75:287–302, 2009.
4. T. Y. Berger-Wolf, B. DasGupta, W. Chaovalitwongse, and M. V. Ashley. Combinatorial reconstruction of sibling relationships. In *Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics (CBGI 05)*, pages 1252–1255, Utah, 2005.
5. T. Y. Berger-Wolf, S. I. Sheikh, B. Dasgupta, M. V. Ashley I. C. Caballero, W. Chaovalitwongse, and S. P. Lahari. Reconstructing sibling relationships in wild populations. *Bioinformatics*, 23(13):49–56, 2007.
6. J. Beyer and B. May. A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, 12:2243–2250, 2003.
7. M. S. Blouin. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *TRENDS in Ecology and Evolution*, 18:503–511, 2003.
8. A. Bretman and T. Tregenza. Measuring polyandry in wild populations: a case study using promiscuous crickets. *Molecular Ecology*, 14:2169–2179, 2005.
9. V. D. E. Castele, T., and E. Matthysen. Natal dispersal and parental escorting predict relatedness between mates in a passerine bird. *Molecular Ecology*, 15:2557–2565, 2006.
10. W. A. Chaovalitwongse, C.-A. Chou, T. Y. Berger-Wolf, B. Dasgupta, S. Sheikh, M. V. Ashley, and I. C. Caballero. New optimization model and algorithm for sibling reconstruction from genetic markers. *INFORMS Journal On Computing*, (to appear), 2009.
11. W. Chaovalitwongse, T. Y. Berger-Wolf, B. Dasgupta, and M. V. Ashley. Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, 22(1):11 – 24, 2007.
12. D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, 2002.
13. C. M. Herbinger, P. T. O’Reilly, R. W. Doyle, J. M. Wright, and F. O’Flynn. Early growth performance of atlantic salmon full-sib families reared in single family tanks versus in mixed family tanks. *Aquaculture*, 173(1–4):105–116, 1999.
14. R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.
15. D. A. Konovalov, C. Manning, and M. T. Henshaw. KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, 4(4):779–82, 2004.
16. D. C. Queller and K. F. Goodnight. Estimating relatedness using genetic markers. *Evolution*, 43(2):258–275, 1989.
17. D. C. Queller and K. F. Goodnight. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, 8(7):1231–

- 1234, 1999.
18. S. I. Sheikh, T. Y. Berger-Wolf, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, and B. DasGupta. Error-tolerant sibship reconstruction in wild populations. In *Proceedings of 7th Annual International Conference on Computational Systems Bioinformatics (CSB)*, 2008.
  19. S. I. Sheikh, T. T. Berger-Wolf, W. Chaovalitwongse, and M. V. Ashley. Reconstructing sibling relationships from microsatellite data. In *Proceedings of the European Conf. on Computational Biology (ECCB)*, 2007.
  20. S. I. Sheikh. Combinatorial Methods in Kinship Analysis *Ph.D. Thesis*
  21. S.M. Sogard, E. Gilbert-Horvath, E. C. Anderson, R. Fisher, S. A. Berkeley, and J. Carlos Garza. Multiple paternity in viviparous kelp rockfish, *Sebastes atrovirens*. *Environmental Biology of Fishes*, 81:7–13, 2008.
  22. S. C. Thomas and W. G. Hill. Estimating Quantitative Genetic Parameters Using Sibships Reconstructed From Marker Data. *Genetics*, 155:1961–1972, 2000.
  23. S. C. Thomas and W. G. Hill. Sibship reconstruction in hierarchical population structures using markov chain monte carlo techniques. *Genetics Research*, 79:227–234, 2002.
  24. J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166:1968–1979, 2004.
  25. J. Wang and A. W. Santure. Parentage and sibship inference from multi-locus genotype data under polygamy. *Genetics*, 181:1579–1594, 2009.
  26. A. J. Wilson, G. Mcdonald, H. K. Moghadam, C. M. Herbinger, and M. M. Ferguson. Marker-assisted estimation of quantitative genetic parameters in rainbow trout, *Oncorhynchus Mykiss*. *Genetics Research*, 81:145–156, 2003.