COMBINATORIAL RECONSTRUCTION OF HALF-SIBLING GROUPS

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

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

Wanpracha Chaovalitwongse
Department of Industrial Engineering, Rutgers University,
CoRE Building, 96 Frelinghuysen Rd., Piscataway, NJ 08854, USA.
Email: wchaoval@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.
Email: 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 computationally intense problem. In this paper, we present two different formulations of the half-sib reconstruction problem and prove their NP-hardness. We also present exact solutions for these formulations and develop heuristics. Using biological and synthetics data sets we present experimental results and compare them with the leading alternative software COLONY. We show that our results are computationally superior and in terms of quality allow half-sib group reconstruction in the presence of polygamy (unlike COLONY), which is prevalent in nature.

1. Introduction

Several studies [1–6], including ours [7–12], have recently developed computational approaches to reconstruct full-sibling groups of wild populations from genetic markers such as microsatellites. Few methods focus on half-sibling relationship. However, half-sib reconstruction has many applications in the study of animal mating systems which are polygamous or promiscuous, and where cohorts of offspring can be more easily sampled than the adult breeders. In this paper, we focus on the half-sibling reconstruction problem. The problem is not only harder to analyze theoretically, it is also much 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 two new parsimony-based formulations for the half-sibling reconstruction problem and show that they are NP-complete; 3) we develop exact algorithms for solving these hard combinatorial formulations; 4) we test these methods using both biological and simulated datasets and compare our reconstruction results to those obtained by the leading alternative approach COLONY[3].

2. Half-Sibling Reconstruction

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 [13]. While full sibling relatedness is difficult to infer, half-sibling

*Corresponding author.

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-sib and full-sib (sharing both parents) groups. Because of the ubiquity of half-sib groups in nature, population biologists need robust approaches to inferring half-sibling relationships from molecular marker data. For example, most plants have flowers pollinated from many different plants, so seeds from a single plant are primarily half-sibs. Identifying these half-sibs among seedlings would allow researchers to study variation in female reproductive success among plants.

In order to formally define the half-sib reconstruction problem, we first establish some basic terminology and describe the genetic markers.

2.1. Definitions

Half and Full Siblings: a group of individuals that shares both parents is referred to as *full siblings*, and when they share at least 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.

Homozygous (heterozygous) individual is one having two identical (different) alleles at a particular genetic 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 $U = \{X_1, \ldots X_n\}$, where U is a population of n diploid individuals of the same generation, and where each individual is represented by a genetic (microsatellite) sample at l loci. That is, $X_i = (\langle a_{i1}, b_{i1}\rangle, \ldots, \langle a_{il}, b_{il}\rangle)$ and a_{ij} and b_{ij} are the two alleles of the individual i at locus j represented as some identifying string. The goal is to reconstruct half-sib groups which is formulated as a cover of individuals by sets $P_1, \ldots P_m$ where individuals in the same set P_i share at least one parent. We assume no knowledge of parental information.

What complicates the half-sib problem is the existence of multiple half-sib reconstructions for a given cohort. Consider the cohort of individuals in Table 2b, the full-sib reconstruction is clear and there is only one correct answer. However, for the same cohort, there are as many as four different half-sib potential reconstructions, as shown in Table 2c. Each of these reconstructions is biologically plausible, i.e. individuals placed in a half -sib group share exactly one parent. Every individual, and the full-sib group it belongs to, is always in the intersection of two half-sib groups.

2.4. Related Work

COLONY [3] is a widely used software for both the full and half-sibs reconstruction. However, it as-

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

Id	Locus 1		Locus 2	
1	7	8	19	20
2	7	10	20	46
3	5	6	19	23
4	4	5	15	19
5	2	10	15	19

Æ(a) Sampled Data

Father	Mother	Children
P1	P2	1, 2
P1	P4	4, 5
P3	P2	7, 8
P3	P4	10, 11
P5	P6	13, 14
P5	P8	15, 16
P7	P6	17, 18
P7	P8	19, 20

Æ(b) Full Sibs

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

Æ(c) *Biologically consistent* half-sib reconstructions

sumes that one gender mates monogamously, an assumption that may greatly limit the software's utility. COLONY, Almudevar *et al.* [4, 5], Herbinger el al. [14], Wilson *et al.* [15], Thomas *et al.* [16] all use likelihood-based approaches to reconstructing both full- and half-sib groups. All of these approaches assume knowledge or availability of allele distributions or mating patterns in the given species.

3. Minimum Half-Sib Reconstruction

One way to interpret parsimony for half-sib reconstruction is to find the minimum number of half-sib groups necessary to explain the cohort. We will formulate the problem and discuss its complexity and an algorithmic solution. In order to do so we first need to define a combinatorial property that all half-sib groups must obey.

3.1. Half-Sibs Property

In Ref. 10 we presented two necessary combinatorial properties that a full-sib 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-sib group must obey. This is a necessary, yet not sufficient property for any feasible half-sib group.

HALF-SIBS PROPERTY : For any given half-sib 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

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

Proof. Recall that a half-sib group is a cohort that shares at least one parent. By Mendelian laws of inheritance, if a group of individuals shares a parent then they must inherit one of two alleles from the parent at each locus. Thus, there must exist at each locus a pair of alleles from which every individual must inherit one. □

In Table 1 the first four individuals can be members of a half-sib 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-sib 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-sib group. The HALF-SIBS PROPERTY constraint is mathematically weak: for any half-sib group that obeys this property a parent can be constructed by using the two alleles at every locus. Furthermore, any two individuals can potentially be half-sibs. In practice, we may also require that any individual or full-sib group may be part of at most two half-sib groups.

3.2. Min-Half-Sibs$_{n,\ell}$ Problem Definition

Input: A set U of n individuals, each with ℓ 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* $|H|$.

3.3. Computational Complexity

Theorem 3.1. MIN-HALF-SIBS$_{n,\ell}$ *is NP-hard.*

Proof. We reduce from the EXACT COVER BY 3-SETS (X3C) problem. X3C is known to be NP-complete [17] and is defined as follows: *given* $\langle n, S_1, S_2, \dots, S_t \rangle$*, where* $n = 3q$ *for some* $q \in \mathbb{Z}^+$ *and* $S_1, S_2, \dots, S_t$ *are a collection of 3-element subsets of* $[n] = \{1, 2, ..., n\}$*, is there a collection of* q *subsets from* $S_1, S_2, \dots, S_t$ *such that their union is* $[n]$*?*

Given an instance $\langle n, S_1, S_2, \dots, S_t \rangle$ of X3C we create an instance of MIN-HALF-SIBS$_{n,\ell}$. For every $j \in [n]$, there is an individual j'. We now describe the gadgets necessary to ensure some structural properties.

Type I Gadgets: These gadgets ensure that no set of four individuals can be half-sibs. There are $\binom{n}{4} = \Theta(n^4)$ such gadgets, each representing a set of four elements ensuring they cannot be half-sibs. Consider a set of four elements $a, b, c, d \in [n]$. The gadget for this set of individuals will disallow the individuals for these items $\{a', b', c', d'\}$ to be half-sibs, but will allow any other combination. We insert a new locus i with six new alleles $\langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle$ for these individuals: $a'_i = \{x_1, x_2\}$, $b'_i = \{x_3, x_4\}$, $c'_i = \{x_5, x_2\}$, $d'_i = \{x_5, x_6\}$, and $e'_i = \{x_1, x_5\}$ $\forall e \in U - \{a, b, c, d\}$.

Type II Gadgets: These gadgets ensure that only the valid sets can be half-sibs. There are $\binom{n}{3} - t = \Theta(n^3)$ such gadgets, each representing a set of three elements that is not one of $S_1 \dots S_t$. Suppose one such set is $\{a, b, c\}$. The gadget for this set of individuals will prohibit the corresponding individuals $\{a', b', c'\}$ to be half-sibs, but all other combinations are allowed. We insert a new locus i with six new alleles $\langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle$ for these individuals: $a'_i = \{x_1, x_2\}$, $b'_i = \{x_3, x_4\}$, $c'_i = \{x_5, x_6\}$, and $e'_i = \{x_1, x_5\}$ $\forall e \in U - \{a, b, c\}$. This allows any set of size three, other than $\{a, b, c\}$, to be half-sibs.

Type III Gadgets: These gadgets ensure that the individuals are distinct. There are $O(n^2)$ such gadgets, each gadget ensuring that a pair of individuals is distinct, while allowing any subset of individuals to be in a half-sib group. Suppose one such pair of individuals is $\{a, b\}$. We insert a new locus i with two new alleles $\langle x_1, x_2 \rangle$ for these individuals: $a'_i = \{x_1, x_2\}$, $b'_i = \{x_1, x_1\}$, and $e'_i = \{x_2, x_2\}$ $\forall e \in U - \{a, b\}$. This ensures that for a pair of individuals a' and b' are unique. However, this locus does not prevent any half-sib groups.

Using these gadget, we can now reduce any instance of the X3C problem to an instance of MIN-HALF-SIBS$_{n,\ell}$ by generating a corresponding individual j' for every element j in X3C, ensuring that a minimum half-sibs solution will automatically give us a solution to X3C. □

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. 10. It consists of two stages:

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

3.4.1. *Step 1: Half-Sibs Enumeration Algorithm.*

In order to generate all maximal half-sib groups we exploit the fact that any two alleles at a locus rep-

resent a potential parent. We first generate all maximal feasible half-sib groups at each locus, and then intersect them to find groups that are common across loci. In order to generate maximal feasible half-sib groups we treat every pair of alleles as the parental genotype (on that locus) and then check which individuals inherit at least one allele from the pair of alleles. We refer to Figure 2 in Appendix A for details.

Lemma 3.1. *Algorithm* Half-Sibs Enumeration *generates all maximal half-sib groups.*

The proof is straight forward and we omit it for brevity.

This algorithm also implies us an upper bound on the number of half-sib groups in a given cohort: $O(\binom{2n}{2}^k) = O(n^{2k})$. Compared to the full-sib reconstruction problem, this tremendously increases the size of the set cover problem. However we are able to execute this algorithm on most of the data sets. For larger 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 [18] problem and is defined as follows: *given a universe U of elements $X_1, \ldots, X_n$ 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[a].

4. Minimum Full-Sib/Half-Sib Reconstruction

Another way to interpret the parsimony objective for the half-sib reconstruction problem is to find a reconstruction that minimizes both full- and half-sib groups. We implement this approach by first finding the minimum number of full-sib groups necessary to explain the population using the 2-Allele Min Set Cover [10] and then merging the full-sib groups to obtain the minimum half-sib groups that cover the population and are composed of full-sib groups. Note, that a reverse Half-Sib to Full-Sib parsimony approach may benefit full-sibling reconstruction, but it is beyond the scope of this paper.

4.1. Half-Sibs From Full-Sibs

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

(1) Generate a full-sib reconstruction F using the 2-Allele Min Set Cover algorithm.
(2) Enumerate all maximal feasible half-sib sets C in the cohort U that obey the Half-Sibs Property and can be obtained by merging a subset of the input full-sib groups F. We start by generating candidate half-sib groups by using all pairs of full-sib groups and then comparing all full-sib groups to all candidate half-sib groups to see if they can be merged conforming to the Half-Sibs Property .
(3) Find the minimum number of maximal feasible sets $S \subseteq C$ necessary to cover the entire cohort U using the Minimum Set Cover.

5. Validation Methodology

5.1. Datasets

To validate and assess the accuracy of our approach, we have used datasets with known genetics and genealogy. 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-sib groups were known because the offspring were collected from individual pregnant or gravid females, and were thus maternally related

[a] CPLEX is a registered trademark of ILOG

input : U: individuals
output: $\mathcal{H}$: Set of Maximal Half-sib groups
$HalfSibs \leftarrow \{U\}$;
foreach *locus* l **do**
  $HalfSibs[l] \leftarrow \emptyset$;
  $Alleles[l] \leftarrow \{a|$allele a appears at locus$l\}$;
  foreach $a \in Alleles[l]$ **do**
    $AlleleSets[l][a] \leftarrow \{I_x|$ Individual with allele a 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

ÆFig. 1.: Algorithm for generating all maximal feasible half-sib groups.

input : S_1, S_2 sets of individuals
output: S sets in common
$S \leftarrow \emptyset$;
foreach $s \in S_1$ **do**
  foreach $t \in S_2$ **do**
    $S \leftarrow S \cup \{s \cap t\}$;
  end
end

ÆFig. 2.: *IntersectGroups*: Algorithm for intersecting sets.

input : U: set of individuals, F: set of full-sib groups
output: H set of feasible half-sib groups
$H \leftarrow F$;
merging $\leftarrow$ true;
while *merging* **do**
  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 $\leftarrow$ true ;
        $H \leftarrow H \cup \{S_{i,j}\}$;
      end
    end
  end
end

ÆFig. 3.: Algorithm for determining Minimum Half-sibs from full-sibs

half-sibs. 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.

Crickets: The field cricket *Grillus bimaculatus* dataset comes from a population of crickets studied in Spain [19]. It consists of 112 individuals from 7 wild-caught gravid females with 6 sampled loci.

Rockfish Larvae: The kelp rockfish *Sebastes atrovirens* dataset [20] 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 random data, we follow the same protocol as in Ref. 8. 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 random and uniform.

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: $1:10, 1:5, 1:3, 1:1$. The half-sib groups based on the sex with the smaller number of breeding adults were chosen as the ground truth, i.e. paternal groups. We generated 10 cohorts for each set of parameters.

5.2. 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-sib groups may overlap multiple times and it is not clear whether the overlap should be penalized. The absence of paternal information implies that we cannot be sure that some half-sib groups given by the algorithm are not representative of the half-sib groups by other sex. We measure the error rates of algorithms using a slight modification of the Gusfield Partition Distance [21]: 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.

6. Results

6.1. Biological Datasets

Crickets

Our Min Half-Sibs approach gives good results, the only difference with the ground truth is that two of the elements are assigned to more than one half-sib groups. The Min-Full-Sib/Half-Sib solution classifies $\frac{20}{111}$ elements incorrectly. COLONY produces an accurate result. See Table 2 for details. Note that COLONY does not allow overlap between half-sib groups because it assumes that one of the sexes is monogamous.

Rockfish Larvae Subset

All three approaches: Min-Half-Sibs, Min Full-Sib/Half-Sib and COLONY produces 100% accurate assignments. See Appendix B: Table 3 for details. Only Min-Half-Sibs produces an overlap of $\frac{4}{288}$ individuals.

6.2. Simulated Datasets

As expected, the ratio of the number of fathers to the number of mothers is the major factor in the accuracy of reconstruction. When the number of father and mothers is comparable, it is possible to pick a different parsimony-based reconstruction, thus the accuracy was low for such scenarios. We were only

Æ**Table 2.**: Results for Crickets by different methods

Set(1):	0 – 15
Set(2):	16 – 31
Set(3):	32 – 47
Set(4):	48 – 63
Set(5):	64 – 79
Set(6):	80 – 95
Set(7):	96 – 111

Æ(a) Original

Set(1):	0 – 15
Set(2):	16 – 31 **110**
Set(3):	32 – 47
Set(4):	48 – 63
Set(5):	64 – 79
Set(6):	80 – 95
Set(7):	**73** 96 – 111

Æ(b) Min Half Sibs

Set(1):	0 – 15 33
Set(2):	13 32 34 73 80 96 109
Set(3):	16 – 31 **80 81** 82 – 85 87 **89** 90 – 95
Set(4):	35 – 47
Set(5):	48– 63
Set(6):	64 – 72 74 – 79
Set(7):	80 81 86 88 89 96 – 111

Æ(c) Min Full-Sib/Half-Sibs

Set(1):	0 – 15
Set(2):	16 – 31
Set(3):	32 – 47
Set(4):	48 – 63
Set(5):	64 – 79
Set(6):	80 – 95
Set(7):	96 – 111

Æ(d) COLONY

Æ**Table 3.**: Results for Rockfish Larvae Subset by different methods

Set(1):	0 – 95
Set(2)	96 – 191
Set(3)	192 – 287

Æ(a) Original

Set(1):	0 – 95 **125**
Set(2)	96 – 191
Set(3)	**111 147 182** 192 – 287

Æ(b) Min Half Sibs

Set(1):	0 – 95
Set(2)	96 – 191
Set(3)	192 – 287

Æ(c) Min-Full-Sibs/Half Sibs

Set(1):	0 – 95
Set(2)	96 – 191
Set(3)	192 – 287

Æ(d) COLONY

Æ**Table 4.**: Accuracy of the Min Half-Sibs, Main Full-Sibs/Half-Sibs, and COLONY algorithms on simulated datasets.

Fathers	Mothers	Loci	Alleles	Families	Offspring	Min-Half-Sibs		Min Full-Sibs/Half-Sibs		COLONY	
						μ	σ	μ	σ	μ	σ
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

able to compare our results to those of COLONY on datasets where the monogamy assumption is not violated, that is the number of the fathers is really small compared to the mothers. Table 4 presents the results of the reconstruction of the three methods.

7. Conclusions

We have developed new intuitive formulations for reconstructing half-sib relationships from microsatellite markers. We make no assumptions about the data or mating patterns other than parsimony. We have also discussed the complexity of both formulations and provided exact algorithms to solve these formulations. Unfortunately, both problems are NP-Hard and the approaches are computationally intense in practice.

The Min-Half-Sibs method correctly reported all the half-sib groups. While the Min Full-Sibs/Half-Sibs approach was not very accurate, it is more efficient as it explores a much smaller space of solutions. Unlike COLONY, our methods allow for both maternal and paternal half-sibs, which are possible and likely in many natural systems.

As discussed in Ref. 10, 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 supe-

rior as it gives accurate results without the assumptions made by other methods.

7.1. Future Work

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-sib computational problem. We consider our methods as a starting point for developing viable practical solutions for half-sibship reconstruction.

In the future, we intend to extend this work to handle data with genotyping errors using consensus methods, similar to our previous work for full-sibs[11]. Furthermore, we will explore the reconstruction of full-sibling relationships from the paternal and maternal half-sibling groups using the Min-Half-Sibs to obtain the required half-sib groups.

Acknowledgments

This research is supported by the following grants: NSF IIS-0612044 (Berger-Wolf, Ashley, Chaovalitwongse, DasGupta), Fulbright Scholarship (Saad Sheikh), NSF CCF-0546574 (Chaovalitwongse), and NSF CAREER IIS-0747369 (Berger-Wolf). 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.

References

1. 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, July 1999.
2. Jen Beyer and B. May. A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, 12:2243–2250, August 2003.
3. J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166:1968–1979, April 2004.
4. 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.
5. A. Almudevar. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, 63, 2003.
6. 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.
7. Wanpracha Chaovalitwongse, Tanya Y. Berger-Wolf, Bhaskar Dasgupta, and Mary V. Ashley. Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, 22(1):11 – 24, February 2007.
8. 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, July 2005.
9. W. 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 of Computing*, to appear.
10. Tanya Y. Berger-Wolf, Saad I. Sheikh, Bhaskar Dasgupta, Mary V. Ashley Isabel C. Caballero, Wanpracha Chaovalitwongse, and Satya P. Lahari. Reconstructing sibling relationships in wild populations. *Bioinformatics*, 23(13):49–56, July 2007.
11. 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) (to appear)*, 2008.
12. S. I. Sheikh, T. Y. Berger-Wolf, W. Chaovalitwongse, and M. V. Ashley. Reconstructing sibling relationships from microsatellite data. In *Proceedings of the European Conf. on Computational Biology (ECCB)*, January 2007.
13. T. Van de Casteele, P. Galbusera, and E. Matthysen. A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology*, 10(6), JUN 2001.
14. 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, March 1999.
15. 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(02):145–156, 2003.
16. S. C. Thomas and W. G. Hill. Estimating Quantitative Genetic Parameters Using Sibships Reconstructed From Marker Data. *Genetics*, 155(4):1961–1972, 2000.
17. M. R. Garey and D. S. Johnson. *Computers and Intractability - A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.
18. Richard 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.

19. A. Bretman and T. Tregenza. Measuring polyandry in wild populations: a case study using promiscuous crickets. *Molecular Ecology*, 14(7):2169–2179, 2005.
20. 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.
21. D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, May 2002.