INFORMS Journal on Computing
Articles in Advance, pp. 1–15
ISSN 1091-9856 | EISSN 1526-5528

DOI 10.1287/ijoc.1090.0322

New Optimization Model and Algorithm for Sibling Reconstruction from Genetic Markers

W. Art Chaovalitwongse, Chun-An Chou
Department of Industrial and Systems Engineering, Rutgers University, Piscataway, New Jersey 08854 {wchaoval@rci.rutgers.edu, joechou@rci.rutgers.edu}

Tanya Y. Berger-Wolf, Bhaskar DasGupta, Saad Sheikh
Department of Computer Science, University of Illinois, Chicago, Illinois 60607 {tanyabw@uic.edu, dasgupta@bert.cs.uic.edu, ssheik3@uic.edu}

Mary V. Ashley, Isabel C. Caballero
Department of Biological Sciences, University of Illinois, Chicago, Illinois 60607 {ashley@uic.edu, icabal2@uic.edu}

With improved tools for collecting genetic data from natural and experimental populations, new opportunities arise to study fundamental biological processes, including behavior, mating systems, adaptive trait evolution, and dispersal patterns. Full use of the newly available genetic data often depends upon reconstructing genealogical relationships of individual organisms, such as sibling reconstruction. This paper presents a new optimization framework for sibling reconstruction from single generation microsatellite genetic data. Our framework is based on assumptions of parsimony and combinatorial concepts of Mendel's inheritance rules. Here, we develop a novel optimization model for sibling reconstruction as a large-scale mixed-integer program (MIP), shown to be a generalization of the set covering problem. We propose a new heuristic approach to efficiently solve this large-scale optimization problem. We test our approach on real biological data as presented in other studies as well as simulated data, and compare our results with other state-of-the-art sibling reconstruction methods. The empirical results show that our approaches are very efficient and outperform other methods while providing the most accurate solutions for two benchmark data sets. The results suggest that our framework can be used as an analytical and computational tool for biologists to better study ecological and evolutionary processes involving knowledge of familial relationships in a wide variety of biological systems.

Key words: set covering; genetic markers; simulation; mixed-integer program; analysis of algorithms; sibling reconstruction

History: Accepted by Harvey Greenberg, Area Editor for Computational Biology and Medical Applications; received January 2008; revised August 2008, January 2009, February 2009; accepted February 2009. Published online in *Articles in Advance*.

1. Introduction

As more and more highly variable molecular markers become available for a wider range of species, investigators can increasingly characterize evolutionary, ecological, population, and demographic parameters in diverse species of plants, animals, and microbes. To effectively extract ecological and evolutionary information from these emerging genotypic data sets, computational approaches for accurate reconstruction of familial relationships need to keep pace with our ability to sample organisms and obtain their genotypes. Therefore, improved methods for reconstruction of genealogical relationships from genetic data will be extremely useful for biologists working on a wide range of such questions. For wild species, kinship and pedigrees cannot be inferred from field observations alone. Several modern tools, like codominant molecular markers such as DNA microsatellites, provide new possibilities to develop novel computational methods of establishing pedigree relationship. In studies where organisms are sampled and genotyped without information about their parents, it may be possible to identify cohorts of siblings based on microsatellite data. The sibling group identification allows inference of many interesting biological parameters, including the number of reproducing adults, their fecundity, and the average size of litters. For threatened species, knowledge of sibship relationships can be important for conservation and aid in management strategies. For studies of evolutionary genetics, sibling reconstruction can be used for assessing the heritabilities of adaptive traits and how they will respond to natural selection.

Although several methods for sibling reconstruction from microsatellite data have been previously proposed (Almudevar and Field 1999, Almudevar

2003, Beyer and May 2003, Konovalov et al. 2004, Painter 1997, Smith et al. 2001, Thomas and Hill 2002, Wang 2004), most techniques have offered very limited applications and have not been very practical (Butler et al. 2004). The main reason is that most sibling reconstruction methods use statistical likelihood models to infer genealogical relationships and are based on the knowledge of typical allele distribution and frequency, family sizes, and other information about the species (Blouin 2003). For this reason, previous techniques are often constrained by the availability of thorough population sampling. They are also heavily biased toward parentage assignment because parentage is more easily resolved.

In this study, we focus on a combinatorial approach that does not require prior genetic information about the species such as population allele frequencies. Our approach is based on combinatorial concepts of the Mendelian laws of inheritance, and for now, we are limiting our methods to diploid organisms (Berger-Wolf et al. 2005, Chaovalitwongse et al. 2007). Similar combinatorial methods have also been successful previously for closely related molecular genetics questions, such as haplotype reconstruction (Eskin et al. 2003, Li and Jiang 2003). Generally speaking, our sibling reconstruction approach uses the simple Mendelian inheritance rules to impose combinatorial constraints (referred as 4-allele and 2-allele constraints) to allow only genetically possible sibling groups to be reconstructed. Note that the main challenge of combinatorial approaches is that the actual number of sibling sets is not known a priori. We therefore employ parsimony assumptions and aim to find the smallest number of sibling groups, each satisfying the Mendelian constraints. In our previous study, we proposed an algorithm to (1) reconstruct all possible sibsets satisfying the 4-allele constraint, which is a looser version of 2-allele constraint; and (2) assign individuals to sibsets by solving a set covering problem. This algorithm was able to reconstruct sibsets with some degree of success (Chaovalitwongse et al. 2007). Most recently, we developed a heuristic approach to generate all maximal sibsets that satisfy the 2-allele constraint, which can be theoretically proven to provide only the lower bound of the real number of sibsets (Berger-Wolf et al. 2007).

In this paper we present the first integrated mathematical programming model to construct and assign individuals into sibsets that satisfy the 2-allele constraint. This model is provably a true presentation of the sibling reconstruction problem under parsimony assumption. Specifically, the objective of this model is to minimize the number of reconstructed sibsets with provably true constraints equivalent to the Mendelian rules. This model is a very large-scale mixed-integer program (MIP), which is very difficult to solve. Since the model has a set covering structure, we propose a new heuristic approach based on a well-known approximation algorithm of the set covering problem to solve this large-scale optimization problem.

The rest of the paper is organized as follows. In §2, some basic background in genetics and population biology and a brief description of combinatorial concepts of the Mendelian inheritance laws are given. In §3, the complexity issues of the sibling reconstruction problem, the proposed optimization model, and the algorithmic framework of our solution approach are presented. Section 4 describes our computational experience including the characteristics of real biological data, the random data generator, a measure of solution accuracy, and the comparison of performance characteristics of our framework with those obtained by some related computational approaches in the literature. The concluding remarks and discussion are given in §5.

2. Background

In this section, we give some basic definitions of some biological terms related to the sibling reconstruction that will be used frequently in this paper. We also give some background of microsatellite data and combinatorial concepts of Mendel's inheritance laws.

2.1. Microsatellite Data

Although there are several molecular markers used in population genetics, microsatellites are the most commonly used in kinship and population studies. Microsatellites are polymorphic loci present in nuclear genomes, usually noncoding, consisting of repeating units of nucleotides. Microsatellites are short (one to six base pairs) simple repeats such as $(CA/GT)_n$ or $(AGC/TCG)_n$ that are scattered around eukaryotic genomes. They are also known as simple sequence repeats (SSRs). Microsatellites are especially useful for studying population demographics and reproductive patterns because they are neutral and co-dominant markers, and the inference of genotypes at each locus is straightforward. More importantly, microsatellites are preferred because of high numbers of alleles and heterozygosities, providing the highest resolution for identifying related individuals (Queller et al. 1993). Because of these advantages and their widespread use, we focus our development of sibling relationship (sibship) reconstruction methods to unlinked, multi-allelic, codominantly inherited markers such as microsatellites. Figure 1 shows a schematic example of microsatellites sampled at two loci and their resulting genotypes (alleles). Note that in reality the allele codings of tandem repeats at different loci may be different. This makes alleles at different loci independent of each other. For example, allele 1 at locus 1 will have a different sequence (i.e., number of tandem repeats) from that of allele 1 at locus 2.

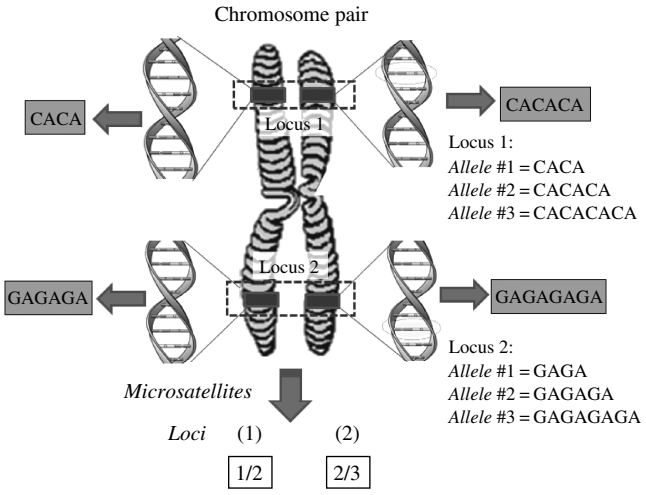


Figure 1 A Schematic Example of a Microsatellite Marker

Notes. Given a chromosome pair of an individual, two loci were sampled. At each locus, genotypes were extracted and allele encoded. In this example, the microsatellite data of this individual are (1/2), (2/3) for loci 1 and 2, respectively.

2.2. Basic Definitions

Sibset is a group of individuals that share at least one parent. When they share both parents they are called full siblings, and when they share exactly one of the parents they are called half siblings. Here, when we refer to sibling groups or sibsets we mean full siblings. *Microsatellites* are short, tandem repeats of a DNA sequence that vary in length. In the genome, microsatellites occur at a specific location on a chromosome, which is called a *locus*. In other words, a locus is a particular chromosomal location of a DNA sequence in the genome—in this case, a microsatellite DNA sequence. An *allele* is a distinct pattern of variable DNA sequences in microsatellites, which is determined by the length of the tandem repeat that occupies a given locus (position) on a chromosome. Usually, numerous alleles occur at a locus, with each allele differing in the number of repeat motifs. In diploid organisms like humans, two homologous copies of each chromosome and two alleles make up the genotype. The alleles for each locus were inherited from each parent (one from the mother and one from the father). A *homozygous* individual has two identical alleles at a particular genetic locus, whereas a *heterozygous* individual has two different alleles at a particular genetic locus.

2.3. Combinatorial Concepts of Mendelian Inheritance Laws

Our basic framework for sibling reconstruction is built around the combinatorial concepts of Mendel's laws (Mendel 1866, Bowler 1989). The law of segregation, known as Mendel's first law, essentially concludes that the two alleles for each characteristic segregate during gamete production to preserve the population variation. In other words, offsprings inherit two alleles (one from the mother and one from the father) on each of the chromosome pairs. The law of independent assortment, known as the inheritance law or Mendel's second law, states that the inheritance pattern of one trait will not affect the inheritance pattern of another. This implies that alleles of different loci assort independently of one another during gamete formation so that there are no correlations across different loci. In short, these laws lay down a very simple rule for gene inheritance: *An offspring inherits one allele from each of its parents independently for each locus.*

Based on this simple rule, we introduce two Mendelian constraints to ensure genetically consistent sibling groups (called *full siblings*). These constraints can be mathematically defined as follows. Given a set U of $|U| = n$ individuals from the same generation, each individual $1 \le i \le n$ is represented by a genetic marker of l loci $\langle (a_{ij}, b_{ij}) \rangle_{1 \le j \le l}$. The numbers a_{ij} and b_{ij} represent a specific allele pair, denoted by front and back alleles, respectively, of individual i at locus l. The above-mentioned Mendelian laws impose the following conditions on a group of individuals $S \subseteq U$ to be full siblings (Berger-Wolf et al. 2005):

Definition 1. A set $S \subseteq U$ of l-locus individuals is said to satisfy the 4-allele condition if $|\bigcup_{i \in S} \{a_{ij}, b_{ij}\}| \le 4$ at locus j for $1 \le j \le l$;

Definition 2. Assuming there is no mutation (allele swapping) in the gene, a set $S \subseteq U$ of l-locus individuals is said to satisfy the 2-allele condition if $|\bigcup_{i \in S} a_{ij}| \le 2$ and $|\bigcup_{i \in S} b_{ij}| \le 2$ for $1 \le j \le l$.

Clearly, the 4-allele and 2-allele conditions are necessary (but not sufficient) combinatorial constraints of Mendelian inheritance laws. In other words, if one knows the maternal and paternal alleles, the offsprings' alleles in the sibset must satisfy these two conditions. However, for a group of individuals whose alleles satisfy these two conditions, they are not necessarily siblings. It is easy to see that the 2-allele condition is stronger (tighter) than the 4-allele condition.

Based on the 4-allele and 2-allele conditions, we can mathematically derive combinatorial constraints of sibsets used for sibset reconstruction (Berger-Wolf et al. 2005, Chaovalitwongse et al. 2007). It is important to note that for the 2-allele condition to hold, we need to assume that there is no front- and back-allele swapping (i.e., the order of the parental alleles is always the same side). In real life, the allele order is unknown and results in swapping alleles at a single locus. Nevertheless, we can derive combinatorial constraints, which are theoretically equivalent to the 2-allele condition, even in the case where there is allele swapping. The combinatorial 4-allele and 2-allele constraints can be defined as follows. Given a set of individuals S, we let A be a collection

of distinct alleles presented at a given locus and let R be a collection of distinct homozygous alleles (appears with itself) present at a given locus.

Definition 3. A set of individuals satisfies the 4-allele condition if $|A| \leq 4$.

Definition 4. A set of individuals satisfies the 2-allele condition if the following two conditions hold: (1) $|A| + |R| \leq 4$, and (2) each and every allele cannot appear together with more than two other alleles (excluding itself).

2.4. Sibling Reconstruction Challenges

For field biologists, familial relationships are needed to learn about a species' evolutionary potential, their mating systems and reproductive patterns, dispersal, and inbreeding. Sibling reconstruction is thus needed when wild samples consist primarily of offspring cohorts, in cases where parental samples are lacking. The real objective of sibling reconstruction is to identify a set of individuals that are siblings. Based on genetic samples of offspring cohorts alone, there is no real objective (function) of the sibling reconstruction problem since the actual pedigree and sibgroups were not known. There are two common frameworks proposed to tackle this sibling reconstruction problem. The first one is to use statistical estimates of relatedness among individuals and try to reconstruct a group containing individuals with very similar allele patterns. The second one is to use the combinatorial concepts of Mendelian rules, as mentioned in the previous section. The real challenge of the combinatorial approach is that it only imposes a rule of biologically consistent sibset but does not have a real objective function. For example, any set of two individuals can be siblings. One can simply group a pair of offspring cohorts and say that they are siblings, and all the reconstructed sibling groups always satisfy the Mendelian rules. To explain the population and its sibling groups when using the Mendelian rules, our approach uses parsimony assumptions to the smallest number of sibling groups, each satisfying the Mendelian rules. Specifically, one can, in turn, formulate the sibling reconstruction problem as a problem of minimizing the number of sibsets that contain all individuals and satisfy the Mendelian rules (i.e., 2-allele constraint). This problem is very difficult, and the complexity of enumerating all possible sibsets satisfying the 2-allele constraint is exponential. These computational challenges are addressed in this paper.

3. Sibling Reconstruction Problem Under Parsimony Assumptions

We present a new optimization model for the sibling relationship reconstruction problem based on microsatellite data acquired from individuals from a single generation. The reconstruction will be based on the 2-allele constraint while we apply a parsimony-driven explanation of the sibsets. In other words, we model the objective of this optimization by assigning individuals parsimoniously into the smallest number of (possibly overlapping) groups that satisfy the necessary 2-allele constraint.

3.1. Complexity and Approximation Issues

First, we discuss the complexity and approximation issues of the sibling reconstruction problem based on the 4-allele constraint and the 2-allele constraint. We consider a set U of n different individuals, each with l loci. The 2-allele problem with parameters n and l is denoted by 2-ALLELE$_{n,l}$, and the 4-allele problem is denoted by 4-ALLELE$_{n,l}$. Let g be the parameter denoting the maximum number of individuals that can be full siblings in an instance of 4-ALLELE$_{n,l}$ or 2-ALLELE$_{n,l}$. Since no two individuals are the same (i.e., their alleles must differ at some locus), $2 \leq g \leq (\binom{4}{1} + 2 \cdot \binom{4}{2})^l = 16^l$ for 4-ALLELE$_{n,l}$. Similarly, we can derive $2 \leq g \leq 8^l$ for 2-ALLELE$_{n,l}$. Either problem has a trivial optimal solution if $g = 2$. Furthermore, if g is a constant, both 4-ALLELE$_{n,l}$ and 2-ALLELE$_{n,l}$ can be posed as a set cover problem with n elements and $\binom{n}{g} = O(n^g)$ sets, with the maximum set size being g, and thus has a $(1 + \ln g)$-approximation by using standard algorithms for the set cover problem (Vazirani 2001). For general g, since any two individuals can be put in the same sibling group, either problem has a trivial $g/2$-approximation. Next, we discuss the approximability results of 2-ALLELE$_{n,l}$ and 4-ALLELE$_{n,l}$ for $g = 3$ and any arbitrary g.

Theorem 1 (Ashley et al. 2009). *Both* 4-ALLELE$_{n,l}$ *and* 2-ALLELE$_{n,l}$ *are* $((153/152) - \epsilon)$*-inapproximable even if* $g = 3$ *assuming* $RP \neq NP$.

This theorem can be proved by providing an approximation preserving reduction from the triangle packing problem to our allele problems. The triangle packing problem requires one to find a maximum number of node disjoint triangles (3-cycles) in a graph. Conceptually, we provide a reduction from an instance of the triangle packing problem to an instance of our allele problems such that three nodes for a 3-cycle in the graph if and only if the individuals corresponding to those triangles can be covered by a sibling group. For more details, please refer to Ashley et al. (2009).

Theorem 2 (Ashley et al. 2009). *For any two constants* $0 < \epsilon < \delta < 1$ *with* $g = n^\delta$, 2-ALLELE$_{n,l}$ *and* 4-ALLELE$_{n,l}$ *cannot be approximated to within a ratio of* n^ϵ *unless* $NP \subseteq ZPP$.

This theorem can be proved by providing a reduction from the well-known graph coloring problem to

our allele problems such that there is an individual corresponding to each node in the graph, and a coloring of the graph translates to a cover by sibling groups with a constant factor blowup in the approximation. For more details, please refer to Ashley et al. (2009).

3.2. 2-Allele Optimization Model

For the sake of simplicity, we shall call the optimization model for the sibling reconstruction problem based on the 2-allele constraint 2-allele optimization model (2AOM). The objective function of 2AOM is to minimize the total number of sibsets while assigning every individual into groups satisfying the 2-allele constraint. The 2AOM problem can be mathematically formulated as follows. First, we define the following sets that will be used throughout this paper: $i \in I$ is a set of individuals, $j \in J$ is a set of reconstructed sibsets, $k \in K$ is a set of alleles, and $l \in L$ is a set of loci. In general, from biological data, we are given a set of $|L|$-locus $|I|$ individuals. Figure 2 illustrates an example of microsatellite genotypes for seven individuals scored at two loci each. We can subsequently present the data into a multidimensional 0-1 matrix format. From the input matrix, f_{ik}^l is defined as an indicator if the front allele at locus l of individual i is k, b_{ik}^l is an indicator if the back allele at locus l of individual i is k, $\bar{a}_{ik}^l = \max\{f_{ik}^l, b_{ik}^l\}$ is an indicator if allele k appears at locus l of individual i, and $\hat{a}_{ik}^l = f_{ik}^l * b_{ik}^l$ is an indicator if individual i is homozygous (allele k appears twice) at locus l. On the right, Figure 2 also shows how the markers are converted into a multidimensional 0-1 matrix representing the input variables $\bar{a}_{ik}^l$. In this example, there are a total of four distinct alleles at locus 1; therefore, we have a 7×4 matrix at locus 1. The matrix of input variables ($\hat{a}_{ik}^l$) can be constructed similarly.

Next we define the following decision variables:

- $z_j \in \{0, 1\}$: indicates if any individual is selected to be a member of sibset j;
- $x_{ij} \in \{0, 1\}$: indicates if individual i is selected to be a member of sibset j;
- $y_{jk}^l \in \{0, 1\}$: indicates if any members in sibset j has allele k at locus l;
- $o_{jk}^l \in \{0, 1\}$: indicates if there is at least one homozygous individual in sibset j with allele k appearing twice at locus l; and
- $v_{jkk'}^l \in \{0, 1\}$: indicates if allele k appears with allele k' in sibset j at locus l.

The mathematical programming formulation of the 2AOM problem is given by the following.

Objective Function. The overall objective function in Equation (1) is to minimize the total number of sibsets:

$$\min \sum_{\forall j \in J} z_j. \tag{1}$$

(1) **Cover and Logical Constraints.** Equation (2) represents the cover constraints ensuring that every individual is assigned to at least one sibset. Equation (3) ensures that the binary sibset variables must be activated for the assignment of any individual i to sibset j.

$$\sum_{\forall j \in J} x_{ij} \geq 1 \quad \forall i \in I; \tag{2}$$

$$x_{ij} \leq z_j \quad \forall i \in I, \forall j \in J. \tag{3}$$

(2) **2-Allele Constraints.** Equation (4) ensures that the binary variable for allele indication y_{jk}^l must be activated for the assignment of any individual i to sibset j. Equation (5) ensures that the binary variable for homozygous indication o_{jk}^l must be activated for the existence of homozygous individual with allele k appearing twice at locus l in sibset j. Equation (7) restricts that the binary variable for allele pair indication $v_{jkk'}^l$ must be activated for any assignment of individual i to sibset j. Note that M_1, M_2, and M_3 are large constants, which can be defined as $M_1 = 2 * |I| + 1$ and $M_2 = M_3 = |I| + 1$. Equation (6) ensures that the number of distinct alleles and the number of homozygous alleles is less or equal to 4. Equation (8) ensures that every allele in the set does not appear with more than two other alleles (excluding itself).

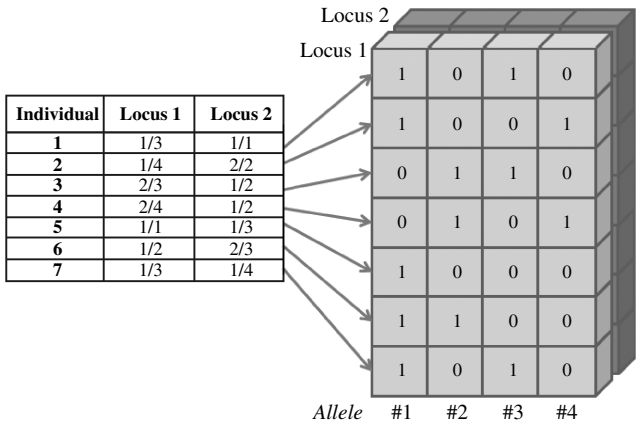
$$\sum_{\forall i \in I} \bar{a}_{ik}^l x_{ij} \leq M_1 y_{jk}^l \quad \forall j \in J, \forall k \in K, \forall l \in L; \tag{4}$$

$$\sum_{\forall i \in I} \hat{a}_{ik}^l x_{ij} \leq M_2 o_{jk}^l \quad \forall j \in J, \forall k \in K, \forall l \in L; \tag{5}$$

$$\sum_{\forall k \in K} (y_{jk}^l + o_{jk}^l) \leq 4 \quad \forall j \in J, \forall l \in L; \tag{6}$$

$$\sum_{\forall i \in I} \bar{a}_{ik}^l \bar{a}_{ik'}^l x_{ij} \leq M_3 v_{jkk'}^l \quad \forall j \in J, \forall k \in K, \forall k' \in K \backslash k, \forall l \in L; \tag{7}$$

$$\sum_{\forall k' \in K \backslash k} v_{jkk'}^l \leq 2 \quad \forall j \in J, \forall k \in K, \forall l \in L. \tag{8}$$



Individual	Locus 1	Locus 2
1	1/3	1/1
2	1/4	2/2
3	2/3	1/2
4	2/4	1/2
5	1/1	1/3
6	1/2	2/3
7	1/3	1/4

1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
1	0	0	0
1	1	0	0
1	0	1	0

Figure 2 **An Example of an Input Data Matrix ($\bar{a}_{ik}^l$) from Microsatellite Markers**

(3) **Binary and Nonnegativity Constraints.**

$$z_j, x_{ij}, y^l_{jk}, o^l_{jk} \in \{0,1\} \quad \forall i \in I, \forall j \in J, \forall k \in K, \forall l \in L.$$

The total number of discrete variables is $O(\max(|J| * |K| * |L|, |I| * |J|))$, and the total number of constraints is $O(|J| * |K|^2 * |L|)$. It is easy to see that the 2AOM problem is a very large-scale MIP problem and may not be easy to solve in large instances. Next, we will prove the correctness of our model and show that the 2AOM problem is NP-hard in the strong sense. The 2AOM problem is considered to be a generalization of the well-known set covering problem with additional constraints to satisfy the 2-allele condition.

Proposition 1. *The constraints in Equations* (4)–(8) *are equivalent to the* 2*-allele constraint*.

Proof. By Equation (4), $y^l_{jk} = 1$ indicates that there exists at least one member in sibset j with allele k at locus l. Therefore, in sibset j, the total number of distinct alleles at locus l is equal to $\sum_{\forall k \in K} y^l_{jk}$, which is equivalent to $|A|$ in the 2-allele theorem. By Equation (5), $o^l_{jk} = 1$ indicates that there exists at least one homozygous member in sibset j with allele k appearing twice at locus l. Therefore, in sibset j, the total number of distinct homozygous alleles at locus l is equal to $\sum_{\forall k \in K} o^l_{jk}$, which is equivalent to $|R|$ in the 2-allele theorem. This will make $|A| + |R| \le 4$ equivalent to $\sum_{\forall k \in K}(y^l_{jk} + o^l_{jk}) \le 4$. By Equation (5), $v^l_{jkk'} = 1$ indicates that allele k appears together with allele k' at locus l in sibset j. By Equation (8), we guarantee that every allele does not appear with two other allele at every locus. This completes the proof. □

Proposition 2. *If we introduce a weight or cost* c_j *to each set* z_j $\forall j \in J$ *in Equation* (1)*, the set covering problem can be reduced to the* 2*AOM problem*.

Proof. Consider a standard set covering problem: $\sum_{\forall j \in J} c_j z_j$ subject to $\sum_{\forall j \in J} a_{ij} z_j \ge 1$, $z_j \in \{0,1\}$. We can reduce this set covering problem to the 2AOM problem as follows. First, relax the constraints in Equations (5)–(8) in the 2AOM problem, and let $|K| = |L| = 1$. If $a_{ij} = 0$ in the set covering problem, define $\bar{a}^l_{ik} = M_1 + 1$; otherwise, $\bar{a}^l_{ik} = 1$, where $M_1 = |I| + 1$. Equation (4) can then be expressed by $\sum_{\forall i \in I} \bar{a}_i x_{ij} \le M_1 y_j$, which allows $x_{ij} = 1$ only if individual i can be covered by set j ($a_{ij} = 1$). Multiplying both sides of Equation (3) by x_{ij}, we obtain $(x_{ij})^2 = x_{ij} \le x_{ij} z_j$ (because $x_{ij} \in \{0,1\}$, $x^2_{ij} = x_{ij}$). Summing the expression over J, we obtain $\sum_{\forall j \in J} x_{ij} \le \sum_{\forall j \in J} x_{ij} z_j$. Combining this expression with Equation (2), we can derive the following expression: $\sum_{\forall j \in J} x_{ij} z_j \ge \sum_{\forall j \in J} x_{ij} \ge 1$, which is equivalent to the constraints $\sum_{\forall j \in J} a_{ij} z_j \ge 1$ in the set covering problem. This completes the proof. □

Proposition 3. *The* 2*AOM problem is strongly NP-hard*.

Proof. According to Ashley et al. (2009), the 2-ALLELE$_{n,l}$ problem is NP-complete. The 2AOM is an optimization version of 2-ALLELE$_{n,l}$ and a generalization of the set covering problem. Therefore, the 2AOM problem is NP-hard. □

It is very important to note that the 2AOM problem requires an initialization of the number of sibsets. If the initial number of sibsets is too small, the problem will become infeasible. If the initial number of sibsets is too large, we will have to introduce many more binary variables than needed. The proposed heuristic approach discussed next can also be used to initialize the number of sibsets as its solution can be theoretically shown to be an upper bound of the 2AOM problem.

In general, the objective and covering constraints of the 2AOM problem is rather artificial to the sibling reconstruction problem as they are built upon parsimony assumptions. In addition, the proposed heuristic approach (to be discussed in the next section) provides a sibset reconstruction solution that is in a form of set partitioning. We should therefore investigate a variant of the 2AOM problem with set partitioning constraints. Specifically, we will test this modified 2AOM problem ($2\widehat{AOM}$) by replacing the set covering constraints in Equation (2) with set partitioning constraints given by

$$\sum_{\forall j \in J} x_{ij} = 1 \quad \forall i \in I. \tag{9}$$

3.3. Heuristic Approach: Iterative Maximum Covering Set

As mentioned earlier, the 2AOM problem is a very large-scale MIP problem. In addition, based on the parsimony assumptions, the minimum number of sibsets may not give the most accurate sibling reconstruction, which is the real objective of our sibling reconstruction problem. In addition, we can only say that the optimal solution to 2AOM (the number of sibsets) is biologically a true lower bound of the real sibsets. Therefore, to solve our problem more efficiently, we herein propose a heuristic approach—namely, an iterative maximum covering set (IMCS)—which is an iterative optimization approach motivated by a widely known approximation algorithm for the set covering problem, i.e., a maximum coverage approach. The idea behind this approach is to construct one sibset maximizing the individual cover in each iteration. Essentially, in each iteration, we solve a reduced problem of 2AOM. The objective of IMCS is to maximize the total number of individuals to be covered by a sibset, which satisfies the 2-allele property. The IMCS problem can

be formally defined as follows. First, we define the following decision variables:

- $x_i \in \{0, 1\}$: indicates if individual i is selected to be a member of the current sibset;
- $y_k^l \in \{0, 1\}$: indicates if any members in the current sibset has allele k at locus l;
- $o_k^l \in \{0, 1\}$: indicates if there is at least one homozygous member in the current sibset with allele k appearing twice at locus l; and
- $v_{kk'}^l \in \{0, 1\}$: indicates if allele k appears with allele k' in the current sibset at locus l.

We mathematically formulate the IMCS problem at each iteration as follows.

Objective Function. The overall objective function in Equation (10) is to maximize the total number of individuals to be selected as members of the current sibset:

$$\max \sum_{\forall i \in I} x_i. \tag{10}$$

(1) **2-Allele Constraints.** Equation (11) ensures that the binary variables for allele indication must be activated for the assignment of individual i to the current sibset. Equation (12) ensures that the binary variables for homozygous indication must be activated for the existence of homozygous individual, with allele k appearing twice at locus l in the current sibset. Equation (13) restricts that the binary variables for allele pair indication $v_{kk'}^l$ must be activated for the selection of individual i. Note that M_1, M_2, and M_3 are large constants, which can be defined as $M_1 = 2 * |I| + 1$ and $M_2 = M_3 = |I| + 1$. Equation (14) ensures that the combination of the number of distinct alleles and the number of homozygous alleles in the current sibset is less or equal to 4. Equation (15) ensures that every allele of each individual does not appear together with more than two other alleles (excluding itself).

$$\sum_{\forall i \in I} \bar{a}_{ik}^l x_i \le M_1 y_k^l \quad \forall k \in K, \forall l \in L; \tag{11}$$

$$\sum_{\forall i \in I} \hat{a}_{ik}^l x_i \le M_2 o_k^l \quad \forall k \in K, \forall l \in L; \tag{12}$$

$$\sum_{\forall i \in I} \bar{a}_{ik}^l \bar{a}_{ik'}^l x_i \le M_3 v_{kk'}^l \quad \forall k \in K, \forall k' \in K \backslash k, \forall l \in L; \tag{13}$$

$$\sum_{\forall k \in K} (y_k^l + o_k^l) \le 4 \quad \forall l \in L; \tag{14}$$

$$\sum_{\forall k' \in K \backslash k} v_{kk'}^l \le 2 \quad \forall k \in K, \forall l \in L. \tag{15}$$

(2) **Binary and Nonnegativity Constraints.**

$$x_i, y_k^l, o_k^l \in \{0, 1\} \quad \forall i \in I, \forall k \in K, \forall l \in L.$$

It is easy to see that the IMCS problem is much more compact than the 2AOM problem and it does not require an initialization in terms of the total number of sibsets. The total number of discrete variables in the IMCS problem is $O(\max(|K| * |L|, |I|))$, and the total number of constraints is $O(|K|^2 * |L|)$.

Iterative Procedure. The idea of iterative procedure of the proposed heuristic approach is motivated by Khuller et al. (1999). This heuristic approach is required to solve the IMCS problem in multiple iterations (m), where m is the final number of sibsets at the termination of our approach. In each iteration, the solution to the IMCS problem gives a list of individuals to be assigned to the current sibset. Then we remove the assigned individuals from the set I and repeat this procedure until there are no individuals in set I. The procedure of the iterative maximum covering set approach is given in Figure 3. The overall approach is viewed as solving an assignment problem rather than solving the set covering problem, because an individual belongs to only one set. We note that this approach is very fast and scalable and can be used for very large-scale sibset reconstruction problems. This is because after every subsequent iteration, the IMCS problem becomes significantly smaller as we remove the largest possible group of assigned individuals and alleles that do not appear in the remaining individuals.

3.4. Solution Perturbation of Iterative Maximum Covering Set

Because the mathematical programming formulation of IMCS has a combinatorial objective function, it is very likely that multiple or alternate optimal solutions exist (called degeneracy). It might be worthwhile investigating the accuracy of alternate optimal and second-best solutions. We note that the IMCS approach is greedy-based; the reconstructed sibsets are very much dictated by the sibset constructed in the first (and possibly second) iteration. Here, we can perturb the reconstructed sibsets by exploring alternate optimal or second-best solutions in the first iteration only, and in both first and second iterations.

To obtain alternate optimal and/or second-best solutions, we first optimally solve the IMCS problem in the first (or second) iteration, use a cut constraint to delete the optimal solution from the feasible space, and resolve the IMCS problem with the additional cut constraint. We subsequently repeat the steps of

Iterative Maximum Covering Set approach

Input: Set I of unassigned individuals
Output: The number of sibsets and sibset assignment for individual set I, allele set K, locus set L

WHILE $I \neq \varnothing$ **DO**
—Solve the IMCS problem
IF optimal solution shows $x_i = 1$ **THEN**
—Remove individual i from set I
IF there is no individual in set I having allele k at any loci in L **THEN**
—Remove allele k from set K

Figure 3 Pseudo-Code of the IMCS Approach

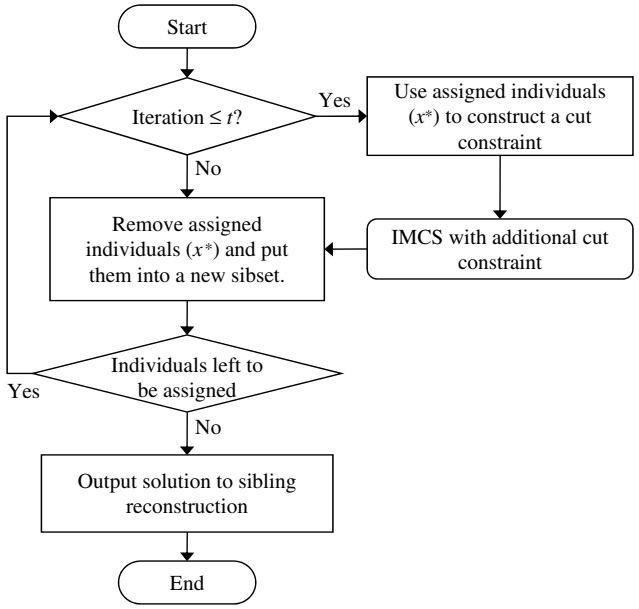


Figure 4 Flowchart of Solution Perturbation of IMCS by Applying a Cut Constraint

Note. t is the number of iterations where the cut constraint is applied ($t = 1, 2$).

IMCS approach in other iterations as usual. Figure 4 illustrates the flowchart of the alternate optimal and second best solution procedure. Here, we use two types of cut constraints. The first cut constraint, called Opt Cut, is a simple constraint to ensure that the optimal solution is deleted from the feasible space, which is given by $\sum_{i' \in K} x_{i'} \leq |K| - 1$, where K is a subset of I whose $x_i^* = 1$ in the previous optimal solution. It may be possible that this cut will only delete one individual from the sibset in the first iteration, and the reconstructed sibsets might be very similar to the ones without the cut. The second cut constraint, called Complementary Cut, is proposed to ensure that the selected individuals in the first sibset will be somewhat different from the previous optimal solution. In other words, we want to ensure that at least one of the individuals that was not selected in the previous optimal solution must be selected in the perturbed solution. The second cut constraint, in fact, ensures that the set of complements is covered, which is given by $\sum_{i' \in I \setminus K} x_{i'} \geq 1$. Although adding the second cut constraint cannot guarantee that the new solution is an alternate optimal or second-best solution, the constraint gives more diversification to the solutions; also note that the second cut constraint is tighter than the first cut constraint.

4. Computational Experience

This section describes the characteristics of our data set (both real biological and simulated data) used in this study to evaluate the performance of the proposed 2AOM, $2\widehat{AOM}$, and IMCS approaches. 2AOM and IMCS approaches are made available at http://kinalyzer.cs.uic.edu. The performance was assessed by the accuracy of reconstructed sibsets with respect to the real (known) sibling groups.

4.1. Data Set

We used both real biological data and randomly generated data to assess the performance of our optimization model and algorithm. Some of the real data used in this study have been previously used for sibling reconstruction (Almudevar and Field 1999). These data were considered benchmark data because the true sibling relationships were known. However, because of the limitations of the real data, including scoring errors and missing alleles, we developed a random population (problem) generator used to control the characteristics of the data set to validate our approaches.

4.1.1. Real Biological Data. In this study, we used four real biological data sets of microsatellite genotypes scored from individuals whose true sibling groups were known. Although the data sets were obtained from wild species (animals and plants), they came from controlled crosses because true sibling groups are typically not known in wild populations. Most data sets analyzed in this study were imperfect because of the technical errors in acquiring and scoring microsatellite data, which resulted in missing alleles and/or genotyping errors. There are several possible and relatively common causes of imperfect data, including allelic dropout and null alleles. In this study, any missing alleles or detected genotyping error was replaced by a wildcard (∗) to indicate the missing information. When checking for genetic feasibility of membership of a new individual in a sibling group, a wildcard could correspond to any allele. Generally speaking, if the data sets were complete and the sample sizes were large enough, one should be able to reconstruct very accurate sibsets (although one cannot guarantee a perfect reconstruction). Given that we had almost complete allele information for the salmon and shrimp data sets, we expected to obtain more reliable and accurate sibset results than those obtained in the radish and fly data sets. The characteristics of the data sets are shown in Table 1.

Table 1 Characteristics of Real Biological Data Sets

Species	No. of individuals	No. of sibsets	No. of loci	Avg. no. of alleles/locus	Percentage of missing alleles
Salmon*	351	6	4	7.8	0.00
Radish*	531	2	5	3.0	3.99
Shrimp	59	13	7	14.9	0.06
Fly	190	6	2	7.0	37.89

*Some known sibsets in the data set are biologically inconsistent because of genotyping errors during the data collection.

Salmon. The Atlantic salmon *Salmo salar* data set was acquired from the genetic improvement program of the Atlantic Salmon Federation (Herbinger et al. 1999). We used a truncated sample of microsatellite genotypes of 250 individuals from five families with four loci per individual. The data did not have missing alleles at any locus. This data set was a subset of one used by Almudevar and Field (1999) to illustrate their technique.

Radish. The wild radish *Raphanus raphanistrum* data set (Conner 2005) consisted of samples from 150 radishes from two families with five loci and five alleles per locus. There were 37 missing alleles among all the loci. The parent genotypes were available.

Shrimp. The tiger shrimp *Penaeus monodon* data set (Jerry et al. 2006) consisted of 59 individuals from 13 families with seven loci. There were 16 missing alleles among all the loci. The parent genotypes were available.

Fly. The *Scaptodrosophila hibisci* data set (Wilson et al. 2002) consisted of 190 same generation individuals (flies) from six families sampled at various number of loci with up to eight alleles per locus. Parent genotypes were known. All individuals shared two sampled loci that were chosen for our study. A substantial proportion of alleles were missing for some individuals.

4.1.2. Random Data. To create a set of simulation data, we developed a random population generator that works as follows. The generator first constructed a group of adults (parents) with the full genetic information. Based on this information, a single generation of sibling data were generated and the parentage information was retained so that the true sibling groups were known. The sibling population generator requires the following parameters: M is the number of adult males, F is the number of adult females, l is the number of sampled loci, a is the number of alleles per locus, j is the number of juveniles in the population per one adult female, and o is the maximum number of offsprings per parent couple. The procedure of our random generator can be described in detail as follows:

Step 1. First, we generated the parent population of M males and F females with parents with l loci, each having a distinct alleles per locus.

Step 2. After the parents were generated, we created a population of their offsprings by randomly selecting j pairs of parents. A male and a female were chosen independently and uniformly at random from the parent population.

Step 3. For each of the chosen parent pairs, we generated a specified number of offsprings, o, each randomly receiving one allele from its mother and one from its father at each locus.

This population generator is a rather simplistic approach; however, it is consistent with the genetics of known parents and provides a baseline for testing the accuracy of the algorithm. To produce a simulated data set used in this study, we varied the parameters of the population generator as follows:

- The number of adult females (F) and the number of adult males (M) are set to 10;
- The number of sampled loci (l) is set to 2, 4, 6, and 10;
- The number of alleles per locus (a) is set to 2, 5, 10, and 20;
- The number of families (j) is 1, 2, 5, and 10; and
- The maximum number of offsprings per couple (o) is set to 2, 5, and 10.

For each parameter setting, we obtained a set of offspring population with known parent pairs. In each population, there were $o \times j$ individuals with j known sibling groups. Random data are made available at http://kinalyzer.cs.uic.edu.

4.2. Evaluation and Assessment

We evaluated the effectiveness of our approaches by comparing the reconstructed sibling groups with the actual known sibling groups. The error measurement was obtained by calculating the minimum partition distance (Gusfield 2002). The error rate ($1 -$ accuracy) was defined as the ratio of the partition distance to the total number of individuals. In other words, the accuracy used in this study is the percentage of individuals correctly assigned to sibling groups.

The minimum partition distance is known to be equivalent to the maximum linear assignment problem (MLAP) that can be solved in polynomial time. This problem is also known as the maximum bipartite weighted matching problem. The MLAP for sibling reconstruction problem can be defined as follows: Given two collections of sibsets $\{A_1, \ldots, A_n\}$ and $\{B_1, \ldots, B_m\}$, let C be the $n \times m$ cost matrix where c_{ij} is the cost of the assigning sibset A_i to sibset B_j. Then the MLAP is to find an assignment of all sibsets in A to all sibsets in B at the maximum cost (individual matchings) such that each sibset in A is assigned to at most one sibset in B, and vice versa. The MLAP can be formulated as a MIP problem given by

$$\max \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij} \tag{16}$$

$$\text{s.t.} \sum_{j=1}^{m} x_{ij} \le 1 \quad \text{for } i = 1, \ldots, n; \tag{17}$$

$$\sum_{i=1}^{n} x_{ij} \le 1 \quad \text{for } j = 1, \ldots, m; \tag{18}$$

$$x_{ij} \in \{0, 1\}.$$

We note that the solution to MLAP, $|U| -$ (maximum assignment), can be represented as the minimum number of individuals to be deleted so that these two sibset collections are identical. This distance is the errors (misassignments) and is used to calculate the accuracy of our approaches. However, note that the relationships among parents are not necessarily monogamous; i.e., some sibsets in A (or B) may not be disjoint. As our MIP model is a covering model, the solution (set of full sibling groups) does not induce a partition on the individuals. Thus, to make this accuracy measure more appropriate, we propose the following modification of the MLAP: given two collections of non-disjoint sets $\{P_1, \ldots, P_n\}$ and $\{Q_1, \ldots, Q_m\}$ of elements in U and a solution (maximum assignment of $|\cup P_i \cap Q_j|$) to the MLAP over the matrix $c_{ij} = |P_i \cap Q_j|$, the minimum distance between two sibling sets is $|U| -$ (maximum assignment).

4.3. Empirical Results

We used the 2AOM, $2\widehat{AOM}$, and IMCS approaches described above to reconstruct the sibling groups from real and simulated data. We subsequently measured the accuracies of the reconstructed sibling sets in reference to the true sibling groups by solving the MLAP for every data set. In addition, we compared the performance characteristics (solution accuracies) of our approaches to the ones obtained by our previous approach (Berger-Wolf et al. 2007) and three other well-known sibship reconstruction methods in the literature (Almudevar and Field 1999, Beyer and May 2003, Konovalov et al. 2004). All tests of our new approaches were run on Intel Xeon Quad Core 3.0 GHz processor workstation with 8 GB RAM memory. Computational times reported in this section were obtained from the desktop's internal timing calculations, which included time used for preprocessing, perturbation, and postprocessing. All the mathematical modeling and algorithms were implemented in MATLAB and solved using a callable General Algebraic Modeling System (GAMS) library with CPLEX version 10.0 (default setting). The tests of our previous and other methods were run on a single processor with 4 GB RAM memory on the 64-node cluster running RedHat Linux 9.0. The difference in platforms and operating systems was dictated by the available software licenses and provided binary codes.

4.3.1. Results from Real Biological Data. We used the 2AOM, $2\widehat{AOM}$, and IMCS approaches described in §3 to reconstruct the sibling sets on all four real biological data sets. As mentioned, CPLEX was used to solve the optimization models in all approaches and the stopping criterion was set to be either less than 0.01% of solution gap or 20 hours (72,000 seconds) of running time. We note that the IMCS approach obtained the optimal solutions in all instances, whereas the 2AOM and $2\widehat{AOM}$ approaches obtained the optimal solution only in the radish data set. Specifically, in the salmon, shrimp, and fly data sets, when using 2AOM and $2\widehat{AOM}$ approaches, CPLEX failed to obtain the optimal solution under the 72-hour time limit, and the reported results were based on the best integer-feasible solutions. The quality of sibset solutions was assessed in terms of sibset accuracy as explained in §4.2. Table 2 gives the performance characteristics and solution quality of the 2AOM, $2\widehat{AOM}$, and IMCS approaches. The computational times reported in Table 2 are in seconds. We note that the objective functions of both approaches are to minimize the number of sibsets.

Although both the set covering (2AOM) and set partitioning ($2\widehat{AOM}$) models only obtained the optimal solution within the time limit for the radish data set, it provided quite good integer-feasible solutions of sibling reconstruction in all other data sets. In terms of the validity of the parsimony assumption, sibset solution gaps were computed with respect to the true numbers of sibling sets in the biological data sets. The set covering 2AOM approach yielded 33% (2/6), 50% (1/2), 8% (1/13), and 16% (1/6) relative gaps to the real number of sibsets, respectively. The set partitioning $2\widehat{AOM}$ approach yielded 133% (8/6), 50% (1/2), 8% (1/13), and 16% (1/6) relative gaps to the real number of sibsets, respectively. The optimal heuristic solutions of IMCS approach yielded relative gaps for the salmon, radish, and fly data sets of about 17% (1/6), 50% (1/2), and 33% (2/6), respectively. Nevertheless, we note that the objective function values were rather artificial because the real solution of

Table 2 Performance Characteristics of the Proposed 2AOM, 2-Allele Optimization Model with Set Partitioning Constraints ($2\widehat{AOM}$), and IMCS Approaches on Real Biological Data Sets

		2AOM			$2\widehat{AOM}$			IMCS		
Species	Real no. of sibsets	No. of sibsets	Accuracy (%)	CPU time	No. of sibsets	Accuracy (%)	CPU time	No. of sibsets	Accuracy (%)	CPU time
Salmon	6	8	94.02	>72,000	14	76.07	>72,000	7	98.30	149.19
Radish	2	3	51.98	75.17	3	49.15	363.23	3	52.54	26.31
Shrimp	13	14	94.92	>72,000	13	100.00	>72,000	13	100.00	184.72
Fly	6	7	66.84	>72,000	7	55.26	>72,000	8	47.37	22.78

interest was the sibset assignments provided by these approaches. More importantly, the data sets were not perfect; i.e., there were missing alleles and genotyping errors. These errors would have made the true known sibset assignments violate the Mendelian constraints. Thus, the optimal solutions to both 2AOM and $2\widehat{AOM}$ models, if obtained, would have provided solution gaps in terms of the number of sibsets.

We observed that all methods provided relatively accurate reconstructed sibset results. In particular, the IMCS and $2\widehat{AOM}$ approaches provided a perfect sibset reconstruction for the shrimp data set with 100% accuracy. All approaches provided accurate reconstructed sibling relationships for the salmon data set, and the reason that we did not obtain 100% accuracy is because there were genotyping errors and an inaccurate known sibset assignment in the data set. For the fly data set, the 2AOM approach obtained a slightly more accurate sibset solution than that obtained by IMCS approach. On the other hand, the IMCS approach was more accurate in three other data sets. In all data sets except shrimp, the set covering 2AOM approach consistently provided better reconstruction results than the set partitioning $2\widehat{AOM}$ approach.

It is worth noting that none of the approaches performed well on the radish and fly data sets because there were a lot of missing data and genotyping errors. For the radish data set, we investigated the input genotypes and observed that the true sibsets (given solutions) violated the 2-allele property, which was biologically impossible. We did not cleanse or correct the data because this data set was used in the literature before and we wanted to compare our solution with the previous ones. It is important to note that although the IMCS approach was required to solve optimization problems iteratively, the computational times required by the IMCS approach were significantly less than those required by the 2AOM and $2\widehat{AOM}$ approaches.

Solution Perturbation. We investigated and compared the accuracy of reconstructed sibsets by perturbing the solutions in the first and second iterations of IMCS approach. Table 3 presents the performance characteristics of IMCS with Opt Cut and IMCS with Complementary Cut applied in the first iteration only. For the shrimp data set, all approaches performed very well and were able to perfectly reconstruct the true sibsets. In fact, all approaches reconstructed the same sibsets, but the only difference was the order of reconstructed sibsets because of alternate optimal solutions in the first iterations. For the salmon data set, the Complementary Cut also provided an alternate best sibset reconstruction while the Opt Cut provided a slightly less accurate solution. For other data sets, the perturbed solutions did not perform as well as the optimal solution to IMCS, although the accuracies are very close. Table 4 presents the performance characteristics of IMCS with Opt Cut and IMCS with Complementary Cut applied in the first and second iterations. For both shrimp and salmon data sets, the Complementary Cuts again provided an alternate best sibset reconstruction. Nevertheless, the Opt Cut was able to obtain slightly more accurate solutions in other data sets. Based on these results, we concluded that if the data had a good separation among sibsets like the shrimp data set, any of these techniques would have been able to accurately reconstruct the true sibsets. However, for imperfect data, these results suggested that the greedy approach that used only the optimal solution may be a better option in practice.

Performance Comparison. Next, we compared the accuracies of sibset solutions obtained by 2AOM, $2\widehat{AOM}$, and IMCS approaches to four current state-of-the-art methods for sibling reconstruction in Table 5. These methods are based on very diverse approaches with different mechanisms and solution behaviors. The **BWG** algorithm, proposed previously by our group in Berger-Wolf et al. (2007), is based on 2-allele set construction version of the set covering method proposed in Chaovalitwongse et al. (2007). The **A&F** algorithm, proposed in Almudevar and Field (1999), is based on a completely combinatorial approach to exhaustively enumerate all possible sibling sets satisfying the 2-allele property (although the authors do not explicitly state the property) and obtain a maximal, not necessarily optimal, collection of sibling sets. The **B&M** algorithm, proposed in Beyer and May (2003), is based on a mixture of likelihood and combinatorial techniques used to construct a graph

Table 3 Performance Characteristics of the IMCS Approaches with Opt Cut and Complementary Cut Constraints Applied in the First Iteration

		IMCS with Opt Cut			IMCS with Complementary Cut		
Species	Real no. of sibsets	No. of sibsets	Accuracy (%)	CPU time	No. of sibsets	Accuracy (%)	CPU time
Salmon	6	7	98.01	133.59	7	98.30	124.78
Radish	2	3	52.35	23.34	3	51.41	19.94
Shrimp	13	13	100.00	159.30	13	100.00	154.31
Fly	6	8	44.74	23.16	8	36.84	18.67

Table 4 Performance Characteristics of the IMCS Approaches with Opt Cut and Complementary Cut Constraints Applied in the First and Second Iterations

Species	Real no. of sibsets	IMCS with Opt Cut			IMCS with Complementary Cut		
		No. of sibsets	Accuracy (%)	CPU time	No. of sibsets	Accuracy (%)	CPU time
Salmon	6	8	97.72	161.50	7	98.30	134.08
Radish	2	3	52.17	28.58	4	42.94	30.92
Shrimp	13	13	100.00	154.28	13	100.00	165.75
Fly	6	8	46.84	20.84	7	44.21	17.91

with individuals as nodes and the edges weighted by the pairwise likelihood (relatedness) ratio, and identify potential sibling sets by the connected components in the graph. The **KG** or **KinGroup** algorithm, proposed in Konovalov et al. (2004), is based on the likelihood estimates of partitions of individuals into sibling groups by comparing, for every individual, the likelihood of being part of any existing sibling group with the likelihood of starting its own group.

From the results in Table 5, we observed that the proposed 2AOM and IMCS approaches outperformed other methods on the shrimp data set. The main reason that our approaches performed very effectively was that this data set was almost complete in allele information and the average number of distinct alleles per locus was very high compared to other data sets, intuitively making the distinction among different sibsets easier. Nevertheless, our approaches were also competitive in the data sets with missing allele information. We observed that the radish data set presented a problem for all methods except **BWG**, since it had partial self-reproduction and offsprings of a selfed individual were hard to separate from their half-siblings produced by that and any other individual. Our approaches did not take this species-dependent constraint into account.

4.3.2. Results from Simulated Data. We also validated the proposed 2AOM and IMCS approaches on simulated data set and compared the results to the actual known sibling groups in the data to assess the accuracy of our constructed sibling sets. In addition, we compared the accuracy of our approaches to that of the **M4SCP** proposed in our previous paper (Chaovalitwongse et al. 2007). The reason that we did not compare the $2\widehat{AOM}$ approach was that it was almost always outperformed by the 2AOM approach.

Because there were several parameter combinations in this simulation, we limited the running time of CPLEX to two hours (7,200 seconds) for the 2AOM and IMCS approaches. The comparison of the three approaches on randomly simulated data is shown in Table 6. Because there were four-dimensional parameter settings (i.e., four parameters to a set), we reported the results by fixing one parameter at a time. The accuracies and computational time were calculated based on the average of all other varying parameters. From the results in Table 6, we observed that the proposed 2AOM and IMCS approaches outperformed the **M4SCP** on average with all the fixed parameters. Note that this was not always the case for all parameter settings. The reconstruction based on the IMCS approach was consistently better than that based on the 2AOM approach so was the computational time. We observed that the computational time of IMCS drastically increased when $l = 10$, $a = 10$, $j = 10$, and $o = 10$ because there was an instance when the IMCS approach failed to solve the simulated data with that setting to optimality. Therefore, the running

Table 5 Accuracies of the Sibling Sets Constructed by Our Approaches and Other Approaches from Real Biological Data Sets

Species	2AOM (%)	$2\widehat{AOM}$ (%)	IMCS (%)	**BWG** (%)	**A&F** (%)	**B&M** (%)	**KG** (%)
Salmon	94.02	76.07	98.30	98.30	N/A*	99.71	96.02
Radish	51.98	49.15	52.54	75.90	N/A*	53.30	29.95
Shrimp	94.92	100.00	100.00	77.97	67.80	77.97	77.97
Fly	66.84	55.26	47.37	100.00	31.05	27.89	54.73

*__A&F__ ran out of 4 GB memory as it enumerates all possible sibling sets.

Table 6 Accuracies of the Sibling Sets Constructed by Our Approaches and the M4SCP Approach (Chaovalitwongse et al. 2007) from Simulated Data Set

Parameter settings	2AOM		IMCS		M4SCP	
	Accuracy (%)	CPU time	Accuracy (%)	CPU time	Accuracy (%)	CPU time
$l = 2$	59.25	2,273.04	57.61	2.28	54.18	0.26
$l = 4$	63.94	2,754.80	66.53	8.28	52.71	0.21
$l = 6$	64.28	3,005.49	71.44	28.96	54.78	0.19
$l = 10$	60.56	3,078.93	71.89	239.21	55.28	0.19
$a = 2$	26.67	0.56	26.67	0.21	36.98	0.16
$a = 5$	69.42	3,679.45	72.19	30.54	58.34	0.16
$a = 10$	71.81	3,699.62	81.83	225.17	60.71	0.39
$a = 20$	80.14	3,732.64	86.78	22.81	60.91	0.19
$j = 2$	76.67	1.50	78.13	0.72	62.88	0.02
$j = 5$	64.63	3,079.56	64.58	3.65	49.56	0.11
$j = 10$	44.73	5,253.14	57.90	204.68	34.00	0.75
$o = 2$	49.48	1,711.83	54.38	2.67	18.19	0.22
$o = 5$	69.46	3,250.27	69.83	14.41	36.66	0.27
$o = 10$	67.08	3,372.10	76.40	191.97	53.98	0.22

Note. The CPU time is reported in seconds.

time went up to 7,200 seconds. In all other cases, the IMCS approach always obtained optimal solutions in very reasonable time. In most cases, except when $a = 2$ and $j = 2$, the 2AOM approach failed to solve the sibling reconstruction problems to optimality.

Figure 5 illustrates the performance trend for all three approaches when varying the number of alleles per locus (a) and the number of sampled loci (l) and fixing the number of families (j) and the number of offsprings per family (o) to 10. Intuitively, the sibling reconstruction problem should be easier to solve when the number of alleles per locus increases because there is a greater variation in allele frequency distribution, which should help us to distinct one sibling group to another. In Figure 5, the accuracy increases as l increases for both 2AOM and IMCS approaches. However, we did not see the same behavior in the **M4SCP** approach, which was not robust to the goodness and completeness of the data. Similarly, the reconstructed sibling sets should be more accurate if there are more sampled loci (more combinatorial constraints). We observed a very nice accuracy trend in the IMCS approach. However, the accuracies of the **M4SCP** and 2AOM approaches did not improve with the allelic information from additional loci. We speculated that this happened with the 2AOM approach because it failed to efficiently and effectively solve the optimization problems as the problem size increased. Thus, the reconstructed sibling sets were from the best feasible integer (not optimal) solutions. It is easy to see that the proposed IMCS approach is a more robust, efficient, and accurate approach.

Figure 6 illustrates the performance trend for 2AOM, IMCS, **BWG**, **B&M**, and **KG** approaches when varying the number of alleles per locus (a) and the number of sampled loci (l). We observed an increasing accuracy as a increases with all approaches except the 2AOM approach. Although the 2AOM model was a complete mathematical formulation of the sibling reconstruction problem, it failed to obtain the optimal solutions in most cases because of the time limit. Compared with all other approaches, the IMCS approach was the best in terms of the trade-off between solution quality and computational time. Our previous **BWG** approach outperformed the IMCS approach when the number of alleles per locus was small; however, the computational time required by the **BWG** approach was much larger. In conclusion, the proposed 2AOM and IMCS approaches gave accurate and reliable sibset solutions when there was enough separation in the data (e.g., number of loci and number of alleles per locus). Note that although the 2AOM approach would require more computation time (e.g., days or weeks) to solve the MIP problem to optimality, it should deliver the best possible solution. The choice of use would solely be application dependent.

5. Conclusion and Discussion

This paper presents a novel optimization model and solution approach for the problem of sibling reconstruction from single generation microsatellite genetic data. The sibling reconstruction problem is an extremely difficult problem that has been shown to be NP-complete and cannot be approximated to the ratio of n^{ϵ}, where n is the number of individual and $0 < \epsilon < 1$. A new optimization model for this problem 2AOM, was herein developed and shown to be a generalization of the well-known NP-hard set covering problem. A heuristic approach, IMCS, was developed to efficiently solve the 2AOM model based on a well-known approximation algorithm of the set covering problem to iteratively solve the decomposed problems of 2AOM. The IMCS approach is able to accurately reconstruct sibling groups without the knowledge of underlying population allele frequencies, which is required by other likelihood-based sibling reconstruction approaches. This has made our work very

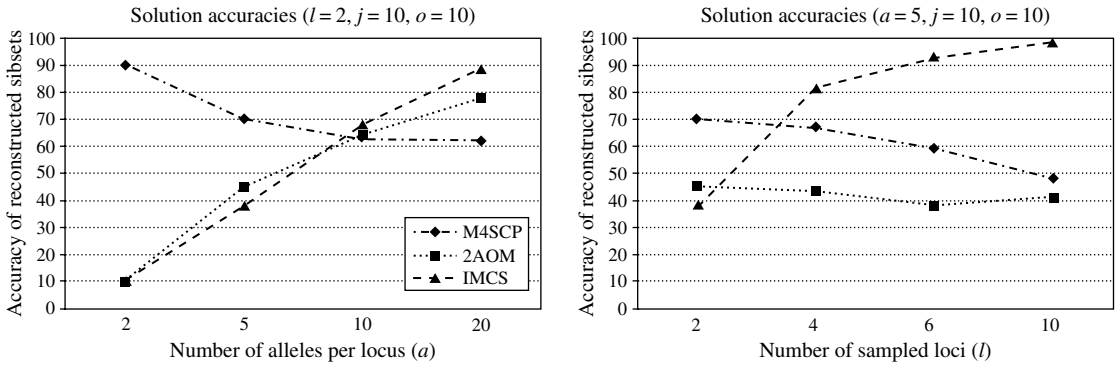


Figure 5 Accuracies of the Sibling Sets Constructed by the 2AOM, IMCS, and M4SCP Approaches on Randomly Generated Data

Notes. The y-axis shows the accuracy of reconstruction as a function of the number of alleles per locus (left) and the number of sampled loci (right). The title shows the value of the fixed parameters.

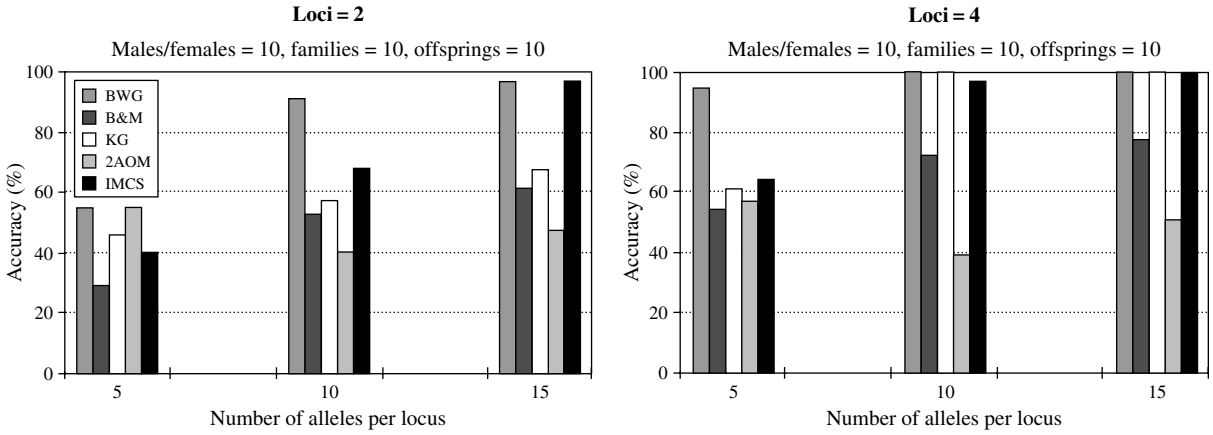


Figure 6 **Accuracies of the Sibling Sets Constructed by 2AOM, ICMS, BWG, B&M, and KG Approaches from Simulated Data Sets with the Parameter Settings** $M = F = 10$, $j = 10$, $o = 10$, $l = 2, 4$, **and** $a = 5, 10, 15$

practical because it may be difficult to obtain accurate estimates of underlying population allele frequencies independently of the sample of potential siblings.

We implemented and tested our approaches on both real biological and simulated data, and then compared the solution quality of our approaches with other state-of-the-art sibling reconstruction methods in the literature. For biological data, our approaches performed as well or better than other methods. Most importantly, our approaches were able to perfectly reconstruct the true sibling sets in the shrimp data set—a result not obtained by our previously published methods. The results suggested that our combinatorial-based approaches gave accurate and reliable sibset solutions for clean and well-separated data sets. On the other hand, our approaches did not perform well for the radish and fly data sets because of missing alleles and biologically inconsistent sibset solutions. These are errors typically present in microsatellite data. For example, allelic dropout occurs when one or both alleles are not amplified during polymerase chain reaction (PCR). Heterozygous mistyping occurs when two alleles are amplified by PCR, but one or both of them, for a variety of reasons, are not recorded as present. Homozygous mistyping occurs when only one allele is amplified by PCR, and it is not any of the parental alleles. Allele combination error occurs when one or both alleles at a locus are present in the parents (or sibling group) but Mendelian inheritance rules are still violated. To reasonably assess our approaches on error-free data, the experiments on simulated data allowed us to estimate the accuracy of our approaches. In all cases except $a = 2$, the proposed IMCS approach successfully reconstructed the sibling sets with an accuracy greater than 50%.

In parallel with this work, we have addressed the possibilities of errors in data or missing allele information by using the concept of consensus methods (Sheikh et al. 2008). We have developed an error-tolerant method for reconstructing sibling relationships to tolerate genotyping errors and mutations in data. The key idea of this method is to remove microsatellite data from one locus at a time, assuming it to be erroneous, and obtain a sibling reconstruction solution based on the remaining loci. We consider an individual pair to be siblings if there is a consensus among (almost) all the reconstructed solutions. Its preliminary results are presented in Sheikh et al. (2008). In the future, we plan to validate our approaches on other biological data sets and more realistic simulated populations (e.g., non-uniform allele distributions). In addition, we will also modify our approaches for populations that contain partial self-reproduction and half-siblings by incorporating species-dependent constraints (field knowledge) into our models.

Acknowledgments

This research is supported by the following grants: National Science Foundation (NSF) IIS-0611998 and NSF CCF-0546574 (to the first author), NSF IIS-0612044 (to the third, fourth, and sixth authors), DBI-0543365 and IIS-0346973 (to the fourth author), Fullbright Scholarship (to the fifth author), and DIMACS special focus on Computational and Mathematical Epidemiology (to the fourth author). The authors are grateful to the people who have shared their data: Jeff Connor, Atlantic Salmon Federation, Dean Jerry, and Stuart Barker. The authors would also like to thank Anthony Almudevar, Bernie May, and Dmitri Konovalov for sharing their software.

References

Almudevar, A. 2003. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoret. Population Biol.* **63**(2) 63–75.

Almudevar, A., C. Field. 1999. Estimation of single-generation sibling relationships based on DNA markers. *J. Agricultural, Biol., Environment. Statist.* **4**(2) 136–165.

Ashley, M. V., T. Y. Berger-Wolf, P. Berman, W. Chaovalitwongse, B. DasGupta, M.-Y. Kao. 2009. On approximating four covering and packing problems. *J. Comput. System Sci.* **75**(5) 287–302.

Berger-Wolf, T. Y., B. DasGupta, W. Chaovalitwongse, M. V. Ashley. 2005. Combinatorial reconstruction of sibling relationships. *Proc. 6th Internat. Sympos. Computational Biol. Genome Informatics (CBGI 05), Salt Lake City, UT*, 1252–1255.

Berger-Wolf, T. Y., S. Sheikh, B. DasGupta, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, S. L. Putrevu. 2007. Reconstructing sibling relationships in wild populations. *Bioinformatics* **23**(13) i49–i56.

Beyer, J., B. May. 2003. A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology* **12**(8) 2243–2250.

Blouin, M. S. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecology Evolution* **18**(10) 503–511.

Bowler, P. J. 1989. *The Mendelian Revolution: The Emergence of Hereditarian Concepts in Modern Science and Society*. The Johns Hopkins University Press, Baltimore.

Butler, K., C. Field, C. M. Herbinger, B. R. Smith. 2004. Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Molecular Ecology* **13**(6) 1589–1600.

Chaovalitwongse, W., T. Y. Berger-Wolf, B. DasGupta, M. V. Ashley. 2007. Set covering approach for reconstruction of sibling relationships. *Optim. Methods Software* **22**(1) 11–24.

Conner, J. K. 2005. Personal communication (December 8).

Eskin, E., E. Haleprin, R. M. Karp. 2003. Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinformatics and Comput. Biol.* **1**(1) 1–20.

Gusfield, D. 2002. Partition-distance: A problem and class of perfect graphs arising in clustering. *Inform. Processing Lett.* **82**(3) 159–164.

Herbinger, C., P. T. O'Reilly, R. W. Doyle, J. M. Wright, F. O'Flynn. 1999. Early growth performance of Atlantic salmon full-sib families reared in single family tanks or in mixed family tanks. *Aquaculture* **173**(1–4) 105–116.

Jerry, D. R., B. S. Evans, M. Kenway, K. Wilson. 2006. Development of a microsatellite DNA parentage marker suite for black tiger shrimp *Penaeus monodon*. *Aquaculture* **255**(1–4) 542–547.

Khuller, S., A. Moss, J. Naor. 1999. The budgeted maximum coverage problem. *Inform. Processing Lett.* **70**(1) 39–45.

Konovalov, D. A., C. Manning, M. T. Henshaw. 2004. KINGROUP: A program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes* **4**(4) 779–782.

Li, J., T. Jiang. 2003. Efficient inference of haplotypes from genotype on a pedigree. *J. Bioinformatics Comput. Biol.* **1**(1) 41–69.

Mendel, G. 1866. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforscheden Vereins in Brünn, Bd. IV fur das Jahr 1865*, 3–47. [Translated as Experiments in plant hybridisation (*J. Roy Horticultural Soc.* **26** 1–32, 1901)].

Painter, I. 1997. Sibship reconstruction without parental information. *J. Agricultural, Biol., Environment. Statist.* **2** 212–229.

Queller, D. C., J. E. Strassman, C. R. Hughes. 1993. Microsatellites and kinship. *Trends Ecology Evolution* **8** 285–288.

Sheikh, S. I., T. Y. Berger-Wolf, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, B. DasGupta. 2008. Error-tolerant sibship reconstruction in wild populations. *Proc. 7th Ann. Internat. Conf. Computational Systems Bioinformatics, Stanford, CA*, 273–284.

Smith, B. R., C. M. Herbinger, H. R. Merry. 2001. Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics* **158**(3) 1329–1338.

Thomas, S. C., W. G. Hill. 2002. Sibship reconstruction in hierarchical population structures using Markov Chain Monte carlo techniques. *Genetic Res.* **79**(3) 227–234.

Vazirani, V. V. 2001. *Approximation Algorithms*. Springer-Verlag, New York.

Wang, J. 2004. Sibship reconstruction from genetic data with typing errors. *Genetics* **166**(4) 1968–1979.

Wilson, A. C. C., P. Sunnucks, J. S. F. Barker. 2002. Isolation and characterization of 20 polymorphic microsatellite loci for *Scaptodrosophila hibisci*. *Molecular Ecology Notes* **2**(3) 242–244.