Column Generation Framework of Nonlinear Similarity Model for Reconstructing Sibling Groups

Chun-An Chou
Department of Systems Sciences & Industrial Engineering, SUNY Binghamton, Vestal, New York 13902
{cachou@binghamton.edu}

Zhe Liang
Department of Industrial Engineering & Management, Peking University, Beijing 100871, China
{liangzhe@coe.pku.edu.cn}

Wanpracha Art Chaovalitwongse
Departments of Industrial & Systems Engineering and Radiology, University of Washington, Seattle, Washington 98195 {artchao@uw.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}

Establishing family relationships, such as parentage and sibling relationships, can be extremely important in biological research, especially in wild species, as they are often key to understanding evolutionary, ecological, and behavioral processes. Because it is often not possible to determine familial relationships from field observations alone, the reconstruction of sibling relationships often depends on informative genetic markers coupled with accurate sibling reconstruction algorithms. Most studies in the literature reconstruct sibling relationships using methods that are based on either statistical analyses (i.e., likelihood estimation) or combinatorial concepts (i.e., Mendelian inheritance laws) of genetic data. In this paper we present a novel computational framework that integrates both combinatorial concepts and statistical analyses into one sibling reconstruction optimization model. To solve this integrated optimization model, we propose a column generation approach with a branch-and-price method. Under the assumption of parsimonious reconstruction, the master problem is to find the minimum set of sibling groups to cover the tested population. Pricing subproblems, which include both statistical similarity and combinatorial concepts of genetic data, are iteratively solved to generate high-quality sibling group candidates. Tested on real biological datasets, our approach is shown to efficiently provide reconstruction results that are more accurate than the ones provided by other state-of-the-art reconstruction algorithms in the literature.

Key words: column generation, branch-and-price, set covering problem, mixed-integer programming, computational biology, sibling reconstruction

1. Introduction

A *sibling group* is defined as a set of individuals that share the same parents. The *sibling reconstruction problem* (SRP) focuses on establishing sibling relationships (and groups) among individuals that are sampled and genotyped without parental information. In this paper, we are limiting our work to full sibling reconstruction, which is to group individuals that share both their mother and father. In studies of important biological phenomena, including mating systems, demographics, social structures and behavior, the SRP has become increasingly important. Among molecular genetic markers, *microsatellites* have been widely used for reconstructing pedigree relationships because they can exhibit high levels of intra-specific polymorphism, are codominantly inherited, and effectively detect variation among individuals and populations (Queller et al., 1993).

Over the past decade, using microsatellite data, several sibling reconstruction approaches have been developed. Those approaches are largely based on either statistical analyses or combinatorial concepts of microsatellite data. Statistical approaches in principle measure the likelihood of offspring partitions for the sample as a whole (aka group likelihood/relatedness score) that is in turn used to infer and construct probable sibling groups (Painter, 1997; Smith et al., 2001; Beyer and May, 2003; Thomas and Hill, 2002; Butler et al., 2004; Konovalov et al., 2004; Wang, 2004; Wang and Santure, 2009). They are generally accurate when applied to error-free datasets. However, almost every real-life dataset contains genotyping errors such as null alleles, allele dropouts, and false alleles. Also, statistical approaches are computationally expensive and memory-consuming as they have to estimate the likelihood or relatedness of all possible pairs or group partitions, making computational time and memory grow drastically with the population size. There have been a few studies that attempted to integrate optimization concepts into statistical approaches to improve the relatedness estimation, but the improvement was extremely limited (Almudevar and Field, 1999; Beyer and May, 2003; Almudevar, 2007, 2003). Combinatorial approaches have been recently developed by our group and applied to the SRP with some degree of success (Berger-Wolf et al., 2005, 2007; Chaovalitwongse et al., 2007). The basic idea of combinatorial approaches is to

impose the combinatorial concepts of Mendelian inheritance laws on reconstructing offspring partitions. Specifically, the SRP can be formulated as a mathematical programming problem with the combinatorial constraints derived from inheritance rules. These approaches first construct a large number of possible sibling groups, and then, based on the assumption of parsimony, select the minimum number of groups so that the entire population is covered. Chaovalitwongse et al. (2010) proposed a complete mixed integer programming (MIP) model of the SRP, and developed an iterative mathematical programming algorithm to reconstruct high-quality sibling groups one by one. Subsequently, Chou et al. (2012) developed a randomized greedy algorithm to efficiently find good reconstruction solutions.

Because of the hard constraints on the inheritance laws, the reconstruction accuracy of combinatorial approaches depends heavily on the integrity of the data. A few genotyping errors can make true (actual) sibling groups violate the Mendelian constraints. Consequently, combinatorial approaches fail to consider offspring partitions of the true sibling groups. Although our group developed an error-tolerant approach based on the consensus concept, it is extremely slow and requires massive computation memory – making it not useful in practice (Sheikh et al., 2008). In this study, we develop a new optimization framework that incorporates both combinatorial concepts and statistical analyses to efficiently provide a more robust and accurate reconstruction solution. In our framework, we use a column generation approach to model the SRP as a set covering problem (also called the master problem). The master problem is to select the minimum number of reconstructed sibling groups that can cover every individual in the population. To generate high-quality reconstructed sibling groups, we use a branch-and-price method and iteratively solve pricing subproblems, in which statistical similarity and combinatorial inheritance laws constraints are incorporated.

The outline of this paper is as follows. In Section 2, we introduce the biological background of the SRP and describe combinatorial concepts (constraints) derived from Mendel's laws and statistical similarity measures of genetic data. In Section 3, we present the basic framework of the column generation for the SRP. Computational results on real biological datasets and performance comparison with existing approaches are presented in Section 4. We draw final conclusions regarding our work in Section 5.

2. Biological Background of Sibling Reconstruciton

2.1 Basic Biological Definitions

A *chromosome* is an organized structure of DNA and protein found in cells. Microsatellite genotyping often targets multiple, specific locations of DNA sequence on the chromosome, called *loci* (singular *locus*). At each locus (i.e., DNA location) on a chromosome, a variant of repeated DNA sequences, called an *allele*, is present. In this study, we focus on diploid organisms, which have two copies of each chromosome – one set inherited from the mother and another set from the father. A genotype is thus the combination of alleles from the chromosome pairs present in an individual for a given set of loci. If two *identical* alleles are present at a given locus, an individual is considered to be *homozygous*. On the other hand, if two *different* alleles are present at a given locus, an individual is considered to be *heterozygous*. It is important to note that microsatellite loci are generally unlinked, so that transmission patterns at each locus are independent.

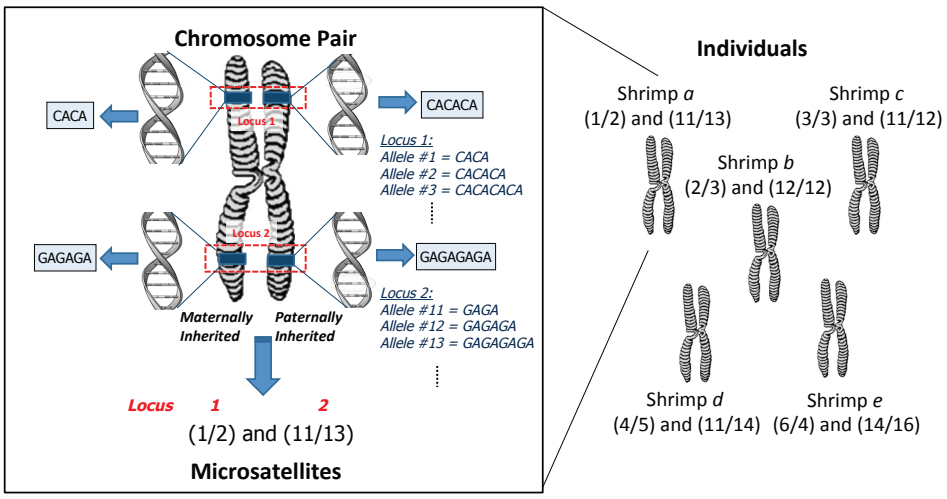

Figure 1: An example of microsatellite data from a sample of five shrimp genotyped at two loci from a chromosome pair. Individual genotypes are defined by a pair of codominant alleles at each locus.

In Figure 1, we illustrate a schematic example of microsatellite data from a sample of five shrimp. Two microsatellite loci are genotyped from each shrimp. At locus 1 in shrimp a, DNA sequences (CA) are repeated twice (denoted by $(CA)_2$) in one chromosome and repeated three times (denoted by $(CA)_3$) in the other chromosome. $(CA)_2$ and $(CA)_3$ are encoded by alleles ♯1 and ♯2 at locus 1. Similarly, at locus 2 in shrimp a, DNA sequences $(GA)_2$, and $(GA)_3$ are encoded by alleles ♯11 and ♯13, for example. Note that allele encoding

can vary from one locus to another, i.e., allele ♯1 at loci 1 and 2 represent different DNA sequences. Shrimp a, d and e are *heterozygous* at both loci whereas shrimp b is *homozygous* at locus 2 and shrimp c is *homozygous* at locus 1.

It is important to note that it is not possible to distinguish which allele came from the mother or the father unless the parental genotypes are known and they share no alleles. Therefore, two pairs of alleles at a locus are unordered, e.g., (♯11/♯13) and (♯13/♯11), are considered to be the same genotype.

2.2 Mathematical Notations of Microsatellite Data

To represent microsatellite data in a mathematical form, we shall define the following notations, which will be used throughout the paper. Figure 2 illustrates a mathematical representation for a sample of five individuals genotyped at two microsatellite loci (shown in Figure 1).

Sets and Elements

I: the set of individuals in a population, indexed by i.

J: the set of all possible sibling groups, indexed by j.

I_j: the set of individuals in sibling group $j \in J$. Hence, $I_j \subseteq I$.

L: the set of loci, indexed by l.

K_l: the set of alleles at locus $l \in L$ in the entire population, indexed by k.

$\tilde{K}_{jl}$: the set of alleles at locus $l \in L$ of sibling group j.

$\hat{K}_{il}$: the set of alleles at locus $l \in L$ of individual i.

$\ddot{K}_{il}$: the homozygous allele at locus l of individual i.

From these notations, we can derived the following relationships. For any locus l, $|\hat{K}_{il}| \leq 2$, $\bigcup_{i \in I_j} \hat{K}_{il} = \tilde{K}_{jl}$ and $\bigcup_{i \in I} \hat{K}_{il} = K_l$. For any $l \in L$ and $i \in I$, $|\ddot{K}_{il}| \leq 1$ and $\ddot{K}_{il} \subseteq \hat{K}_{il}$, and the equal sign is valid if and only if individual i has a pair of identical alleles at locus l. For example, the microsatellite genotype of shrimp a can be represented by $K_{a1} = \{1, 2\}$ and $K_{a2} = \{11, 13\}$.

In Chaovalitwongse et al. (2010), our group introduced a matrix representation of microsatellite data. The input parameters from microsatellite data can be defined by the integer indicator α^l_{ik}, where $\alpha^l_{ik} = 1$ when a single allele k is present at locus l of individual i, $\alpha^l_{ik} = 2$ when homozygous alleles k are present at locus l of individual i, and 0 otherwise. For instance, in Figure 2, $\alpha^2_{d,14} = 1$ indicates that shrimp d has allele ♯14 at locus $l = 2$, while $\alpha^1_{c,3} = 2$ indicates that shrimp c has homozygous allele ♯3 at locus $l = 1$.

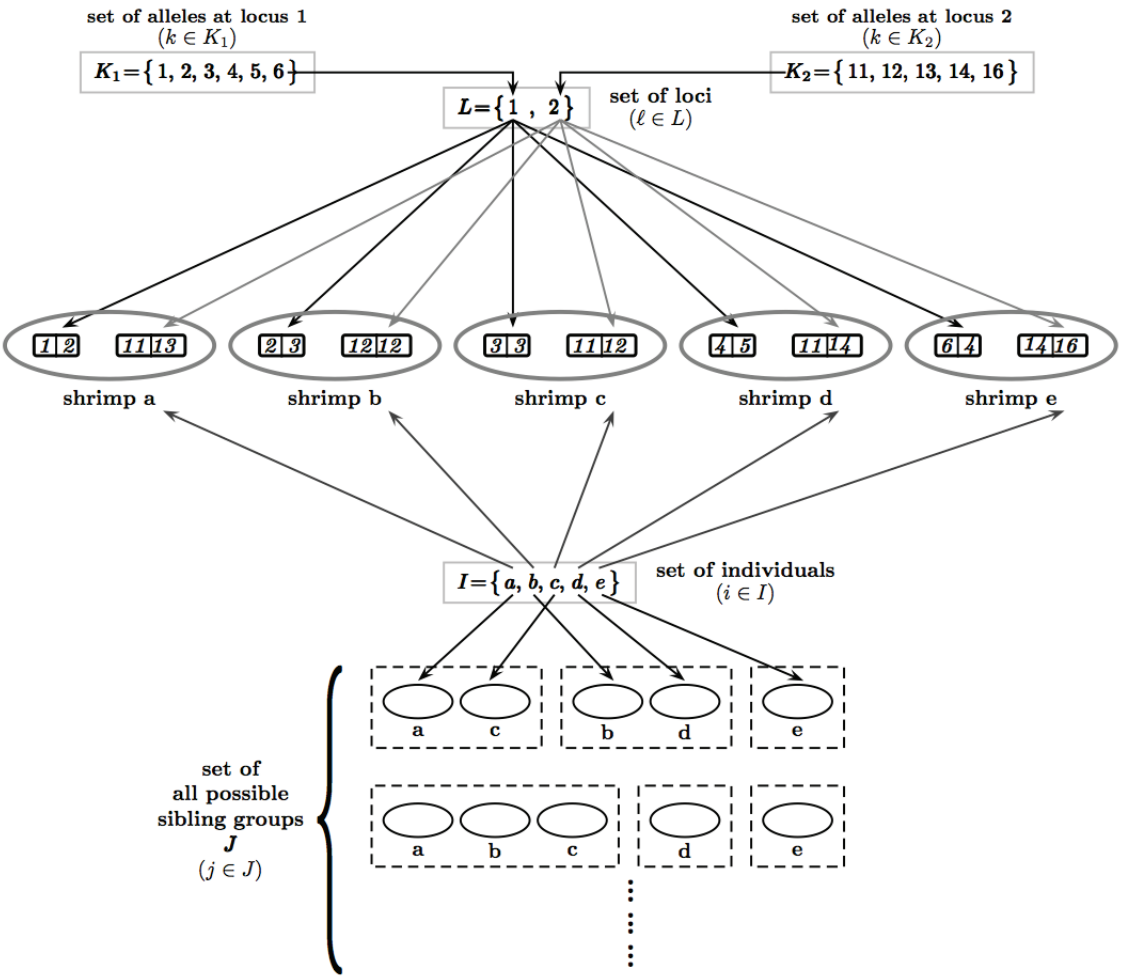

Figure 2: A visual illustration of the mathematical notations for a sample of five individuals genotyped at two microsatellite loci (shown in Figure 1).

2.3 Combinatorial Implications of Mendel's Laws

Mendel's laws (or Mendelian inheritance laws) (Bowler, 1989; Mendel, 1901) laid down simple rules of inheritance: *an offspring inherits one allele from each of its parents at each locus* and *the inheritance pattern of alleles at one locus is independent of the other loci.* Based on these rules, our group introduced the *4-allele condition* as a necessary (but not sufficient) condition to ensure the sibling construction to be genetically consistent (Berger-Wolf et al., 2005; Chaovalitwongse et al., 2007). Specifically, the 4-allele condition requires that for any sibling group, *the number of alleles at each locus is less than or equal to four* ($|\bigcup_{i \in I_j} \hat{K}_{il}| \leq 4$). Subsequently, our group also proposed the *2-allele condition*, which is tighter and more restricted than the 4-allele condition (Berger-Wolf et al., 2007). In Chaovalitwongse et al. (2010), we derived the mathematical constraints of the 2-allele condition for sibling group

as follows:

Definition 1 (*2-allele constraints*) *A sibling group j of individuals I_j satisfies the 2-allele condition if and only if they satisfy the following conditions:*
(i) at any locus $l \in L$, the sum of the numbers of different alleles and homozygous alleles is less than or equal to 4, i.e., $|\bigcup_{i \in I_j} \hat{K}_{il}| + |\bigcup_{i \in I_j} \ddot{K}_{il}| \leq 4$, and
(ii) at any locus $l \in L$, each allele k appears together with no more than two other alleles (excluding itself), i.e., $|\bigcup_{i:i \in I_j, k \in \hat{K}_{il}} \hat{K}_{il} \setminus \{k\}| \leq 2$.

From the example in Figure 2, shrimp a and b can be included in the same biologically consistent sibling group because they satisfy both constraints *(i)* and *(ii)*. Specifically, the group $j = \{a, b\}$ has alleles $\tilde{K}_{j1} = \{1, 2, 3\}$ at locus $l = 1$, and alleles $\tilde{K}_{jl} = \{11, 12, 13\}$ and homozygous alleles $\ddot{K}_{jl} = \{12\}$ at locus $l = 2$. We can see that the sum of the numbers of alleles and homozygous alleles is less than or equal to 4, and each allele appears with at most two others alleles at both loci. On the other hand, shrimps b, c, and d cannot be included in the same sibling group because the group $\{b, c, d\}$ fails to satisfy the constraint *(i)*. That is, the sum of the numbers of alleles $\{2, 3, 4, 5\}$ and homozygous alleles $\{3\}$ exceeds 4 at locus $l = 1$.

3. Mathematical Model and Solution Methods

3.1 Mathematical Formulation of Sibling Reconstruction

The SRP is mathematically formulated as follows. Given a set of individuals $i \in I$ characterized by a set of loci $l \in L$; each containing alleles $k \in K_l$, assume there is a completely enumerated set of all possible sibling groups J. We define the following variables:

Parameter and Variable

δ_{ij}: the binary indicator such that $\delta_{ij} = 1$ if individual i is included in sibling group j, and 0 otherwise.

z_j: the binary decision variable such that $z_j = 1$ if group j is included in the solution, and 0 otherwise.

The SRP herein aims to find a minimum set of sibling groups that cover the entire population. The set covering formulation (Min-SCP) of the SRP is as follows:

$$\text{(Min-SCP)} \quad \min \quad \sum_{j \in J} z_j \tag{1}$$

$$\text{s.t.} \quad \sum_{j \in J} \delta_{ij} z_j \geq 1 \quad \forall\, i \in I, \tag{2}$$

$$z_j \in \{0, 1\}, \quad \forall\ j \in J. \tag{3}$$

The objective function in Equation (1) minimizes the total number of the selected sibling groups. The constraints in Equation (2) ensure that every individual is covered by a selected group. The constraints in Equation (3) are the logical constraints for variables. As mentioned previously, the number of possible sibling groups grows exponentially with the population size. Therefore, it is impractical to enumerate all the groups. In the following sections, we develop a column generation approach to obtain optimal solutions to the LP relaxation of Min-SCP and a branch-and-price method to obtain optimal, integral solutions to Min-SCP.

3.2 Column Generation

Column generation has been widely and successfully applied to many large-scale combinatorial optimization problems when the number of variables (e.g., sibling groups) is too large to enumerate explicitly (Desrosiers et al., 1984). In this section, we propose a column generation approach developed for solving the Min-SCP. Note that sibling groups and columns are interchangeable in the following context. The procedure starts with an initial solution, i.e., a subset of sibling groups. We then solve the restricted master problem of the Min-SCP and obtain the dual costs of the constraints. To improve the restricted master problem, we solve the pricing subproblems based on the dual cost information to generate new sibling groups with positive reduced costs. Subsequently, these groups are added to the restricted master problem, and the updated restricted master problem is resolved again. The procedure iterates until no new sibling groups with positive reduced costs are found. This termination condition implies that the current LP solution of master problem is optimal. The components in the proposed column generation approach are described in the following subsections.

3.2.1 Pricing Subproblem

For the subproblem formulation, a set of variables are defined as follows:

Variables (continuous)

π_i: the non-negative dual variable associated to the constraints in Equation (2).

x_{ij}: the binary variable such that $x_{ij} = 1$ if individual i is assigned in sibling group j, and 0 otherwise.

y^l_{jk}: the integer variable such that $y^l_{jk} = 1$ if allele k is included in the allele set at locus l of group j, $y^l_{jk} = 2$ if allele k is included as a homozygous allele in the allele set at locus l of group j, and 0 otherwise.

$v^l_{jkk'}$: the binary variable such that $v^l_{jkk'} = 1$ if allele k appears with allele k' at locus l in group j, and 0 otherwise.

For a sibling group j, the reduced cost of z_j is written by

$$\bar{c}_j = \sum_{i \in I} \delta_{ij} \pi_i - 1, \quad \forall j \in J. \tag{4}$$

Therefore, the pricing subproblem at every column generation iteration attempts to find the sibling group j with a maximum value of $\bar{c}_j$.

In particular, when solving the subproblem, we determine the assignment of individuals into a new sibling group j, subject to the 2-allele constraints. This can be formulated as a weighted maximization problem (WMP) as follows:

$$\text{(WMP)} \quad \max \quad \sum_{i \in I} \pi_i x_{ij} \tag{5}$$

$$\text{s.t.} \quad \alpha^l_{ik} x_{ij} \le y^l_{jk} \quad \forall\, i \in I,\ k \in K_l,\ l \in L, \tag{6}$$

$$\sum_{k \in K_l} y^l_{jk} \le 4 \quad \forall\, l \in L, \tag{7}$$

$$\sum_{i \in I} \alpha^l_{ik} \alpha^l_{ik'} x_{ij} \le M v^l_{jkk'} \quad \forall\, k \in K_l,\ k' \in K_l \backslash k,\ l \in L, \tag{8}$$

$$\sum_{k' \in K_l \backslash k} v^l_{jkk'} \le 2 \quad \forall\, k \in K_l,\ l \in L, \tag{9}$$

$$x_{ij},\ v^l_{jkk'} \in \{0,1\};\ y^l_{jk} \in \{0,1,2\} \quad \forall\, i \in I,\ k \text{ and } k' \in K_l,\ l \in L. \tag{10}$$

It is important to note that the subscript j for all the variables can be dropped from the WMP because only one sibling group j is generated at a time when solving WMP. The dual variables π_i here are the weights of individuals to be selected. The objective in Equation (5) maximizes the weighted sum of individuals assigned to the current sibling group. The constraints in Equations (6)-(9) ensure that all individuals assigned to the current sibling group must satisfy the 2-allele constraints. Specifically, the constraints in Equations (6)-(7) ensure that the condition *(i)*, described in Section 2.3, is satisfied. The constraints in Equation (6) are the logical constraints of integer variable y^l_{jk} to keep track of the heterozygous or homozygous allele indication. The constraints in Equation (7) ensure that in group j, the sum of the numbers of different and homozygous alleles is less than or equal to 4. The constraints in Equations (8)-(9) ensure that the condition *(ii)*, described in Section 2.3, is satisfied. The constraints in Equation (8) are the logical constraints for $v^l_{jkk'}$ to indicate whether the allele pair k and k' appears together at locus l in group j. Here M is a large positive number defined by $M = |I| + 1$. The constraints in Equation (9) ensure that no allele at locus l in group j appears with more than two other alleles (excluding itself). The constraints in Equation (10) are the binary and integer constraints for variables.

In column generation iteration, the subproblem is usually solved multiple times. The efficacy of the subproblem affects the overall performance of the column generation greatly. Therefore, it is critical to generate high quality variables in a timely fashion when implementing column generation. Next, we present several methods to improve the solution approach for the subproblem.

3.2.2 Integration of Similarity Measure in the Subproblem

Statistical approaches have been widely used to measure the similarity of microsatellite genotypes for a group of individuals (Beyer and May, 2003; Thomas and Hill, 2002; Butler et al., 2004; Konovalov et al., 2004). We here propose a simple pairwise statistical measure of the similarities for all individual pairs. For a pair of two individuals i and $i' \in I$, the similarity score at locus $l \in L$, denoted by $q^l_{ii'}$, can be calculated as follows:

$$q^l_{ii'} = \begin{cases} 1 & \text{if } \sum\limits_{k \in K_l} |\alpha^l_{ik} - \alpha^l_{i'k}| = 0; \\ 0.5 & \text{if } \sum\limits_{k \in K_l} |\alpha^l_{ik} - \alpha^l_{i'k}| = 2; \\ 0 & \text{if } \sum\limits_{k \in K_l} |\alpha^l_{ik} - \alpha^l_{i'k}| = 4. \end{cases} \tag{11}$$

If two individuals have both alleles in common at locus l, then $q^l_{ii'} = 1$. If there is only one allele in common, then $q^l_{ii'} = 0.5$. If there are no alleles in common, then $q^l_{ii'} = 0$. After the similarity score for each locus is obtained, the pairwise similarity score across all loci can be computed as the sum of locus similarities:

$$q_{ii'} = \sum_{l \in L} q^l_{ii'} \qquad \forall\ i, i' \in I. \tag{12}$$

Hence, $q_{ii'}$ can be viewed as the measure of how similar two individuals are. We then integrate the similarity measure with the dual variables as a weighted similarity score as follows:

$$\bar{q}_{ii'} = q_{ii'} \times \pi_i \times \pi_{i'} \qquad \forall\ i, i' \in I, \tag{13}$$

where π_i and $\pi_{i'}$ are the dual variables of individual i and i'. Consequently, the objective function of WMP is modified with the same constraints. A similarity maximization problem (SMP) is formulated as follows:

$$\begin{aligned} \text{(SMP)} \qquad & \max \sum_{i \in I} \sum_{i' \in I} \bar{q}_{ii'} x_i x_{i'} \\ & \text{s.t. } (6) - (9). \end{aligned} \tag{14}$$

Compared with WMP, SMP aims to generate a sibling group that includes as many valuable (i.e., $\pi_i > 0$) and similar (i.e., $\bar{q}_{ii'}$) individuals i as possible. For example, consider individuals a, d, e in Figure 1 and assume all the dual variables have the same positive value. The sibling group $\{a, d\}$ and the sibling group $\{a, e\}$ have the same objective values in WMP; however, the group $\{a, d\}$ is, in fact, preferable over the group $\{a, e\}$ in SMP.

In order to solve the quadratic SMP with a standard MIP solver, we employ a linearization technique proposed by Chaovalitwongse et al. (2004) to reformulate the quadratic program as an *equivalent* mixed integer linear program. The linearized model is reformulated as follows. We define the additional decision variables s_i as the total pairwise similarity score for individual i and r_i as a surplus variable. The linearized SMP is given by

$$\text{(SMP-L)} \quad \max \quad \sum_{i \in I} s_i \tag{15}$$

$$\text{s.t.} \quad \sum_{i' \in I \setminus i} \bar{q}_{ii'} x_{i'} - r_i - s_i = 0 \quad \forall\, i \in I, \tag{16}$$

$$s_i \leq \bar{M} x_i \qquad \forall\, i \in I, \tag{17}$$

$$r_i, s_i \geq 0 \qquad \forall\, i \in I, \tag{18}$$

$$(6) - (9).$$

In the above formulation, $\bar{M}$ is a large positive number, which can be set to $\bar{M} = \sum_{i \in I} \sum_{i' \in I} ||\bar{q}_{ii'}||$. From our preliminary study, we had shown it took only seconds to minutes to solve the SMP-L optimally using CPLEX. Interested readers are referred to the literature that discuss linearization approaches in detail (Chaovalitwongse et al., 2004; Adams and Forrester, 2007).

3.2.3 Hybrid Subproblem Solution Approach

In every column generation iteration, the reduced costs of new columns are checked to examine if the optimality is met. Although SMP can generate high-quality columns, the quadratic objective function in Equation (14) is not directly derived from the reduced cost in Equation (4) and not intuitively associated with the master problem. Therefore, SMP does not have the same optimality condition at termination as WMP. Hence, we propose a hybrid approach consisting of SMP and WMP. In every iteration, SMP is first solved to generate high-quality columns. If the reduced costs of all the sibling groups generated by SMP are not positive, we then solve WMP with the same dual variables one more time to ensure optimality of the solution.

3.2.4 Multi-Group Generation

In column generation it is not necessary to solve the subproblem optimally in every iteration, and, in fact, any columns with positive reduced costs can improve the current master problem (Vanderbeck, 1994; Barnhart et al., 1998). To accelerate the column generation procedure, one of the most widely used strategies is to generate multiple columns with positive reduced cost in an iteration, so that the total number of column generation iterations is reduced. In this paper, we propose two iterative approaches to generate multiple columns, which are described below.

Greedy set partitioning procedure (GSPP) The GSPP iteratively generates disjoint sibling groups by solving the WMP (or SMP). In each GSPP iteration, individuals that already assigned to the generated groups are removed. The procedure continues until all individuals are assigned.

Greedy set covering procedure (GSCP) The GSCP iteratively generates possibly non-disjoint sibling groups by solving WMP (or SMP). In each GSCP iteration, individuals with positive dual variables are in turn selected as a base individual (i.e., the individual that must be included in the sibling group), and a group is generated to cover the base individual in every iteration. The procedure continues until all individuals with non-zero dual variables are visited.

The difference between the two procedures is that in GSPP, the number of individuals assigned to a group decreases with the GSPP iterations, and the sibling group size generated in the latter iterations are usually small. In contrast, GSCP does not remove any individuals in any iteration, and the group size is not effected by the number of iterations. From our experiments, GSCP required more iterations and computational time than GSPP generally because the number of GSCP iterations is fixed to the number of individuals with positive dual variables, and the population size in most GSCP iterations is larger.

It is worth noting that we used GSPP and GSCP, respectively, to generate a set of sibling groups as the initial variables for the column generation.

3.3 Branch-and-Price

A branch-and-price approach is further introduced to obtain optimal IP solution to the Min-SCP. The branch-and-price is a branch-and-bound algorithm in which we employ a

branch-on-follow-on rule in the search procedure and carry out a column generation at each node (Ryan and Foster, 1981; Barnhart et al., 1998). Specifically, in the left branch, we ensure that both individuals i and i' appear together within any selected sibling groups, whereas on the right branch, individuals i or i' do not appear together within any selected sibling groups. Therefore, when searching in the branch-and-bound tree, we need to select individuals i and i' at each node, and to enforce the branching rules in the master problem and subproblem. Next, we present a node selection rule and branching constraints at each node.

3.3.1 Node Selection

Given a fractional solution, we select a pair of individuals i and i' with the highest chance of being together. A function $p(i, i')$ is introduced to quantify how likely individuals i and i' belong to the same sibling group as follows:

$$p(i,i') = \frac{\sum_{j:i \text{ and } i' \in I_j} z_j}{\sum_{j:i \text{ or } i' \in I_j} z_j} \qquad \forall i,\ i' \in I. \tag{19}$$

The pair with the highest p value will be selected as the branching node:

$$\{i, i'\} = \arg \max_{(i,i') \in I} p(i, i'). \tag{20}$$

For instance, consider three generated sibling groups: $j_1 = \{a, b, c, d\}$, $j_2 = \{a, b, c\}$ and $j_3 = \{c, d\}$. Their LP solutions are $z_1 = 0.5$, $z_2 = 0.5$, and $z_3 = 0.5$. From Equations (19), we know that $p(a, b) = (0.5 + 0.5)\ /\ (0.5 + 0.5) = 1$, whereas for other pairs $p(a, c) = p(b, c) = p(c, d) = 2/3$ and $p(a, d) = p(b, d) = 1/3$. Therefore, we branch on the pair (a, b) as it provides the highest p value. If a pair has already been selected at previous nodes, we select the pair with the next highest value of p.

3.3.2 Branching Constraints

After deciding the branching node with the two individuals i and i', we enforce the branch-on follow-on rule by adding a set of constraints to the master problem and the subproblem. The additional constraints for the left branch are:

$$\text{for Min-SCP} \qquad z_j = 0 \qquad \forall j : i \in I_j,\ i' \in I_j,\ \{i, i'\} \not\subseteq I_j, \tag{21}$$

$$\text{for WMP or SMP} \qquad x_i - x'_i = 0. \tag{22}$$

The constraints in Equation (21)-(22) ensure that both individuals i and i' appear together in the solutions of the master problem and the subproblem, respectively. The additional constraints for the right branch are:

$$\text{for Min-SCP} \qquad z_j = 0 \qquad \forall j : \{i, i'\} \subseteq I_j, \tag{23}$$

$$\text{for WMP or SMP} \qquad x_i + x'_i \leq 1. \tag{24}$$

The constraints in Equation (23)-(24) ensure that both individuals i and i' do not appear together in a sibling group.

Note that in the branch-and-bound tree, we perform a depth-first search; the left branch is always searched first so that a feasible solution can be found in the shortest time.

4. Computational Experiments

In this section, we discuss the implementation of our approaches on real biological datasets, and compare the solution quantity and the computational time with other state-of-the-art combinatorial and statistical approaches.

4.1 Biological Datasets

We tested the performance of the proposed approaches on five real biological datasets. These benchmark data have been previously used in the literature because the true sibling groups ("ground truth") are known. The characteristics of the datasets are summarized in Table 1. There are some missing values and the percentages of missing alleles are reported in the last column. Also, it is interesting to note that there are violations of Mendel's laws in the salmon and turtle datasets, which might be due to genotyping errors. The background of the datasets is briefly described below.

Salmon: The Atlantic salmon *Salmo salar* dataset comes from the genetic improvement program of the Atlantic Salmon Federation (Herbinger et al., 1999). We use a truncated sample of microsatellite genotypes of 250 individuals from 5 families with 4 loci per individual. This dataset is a subset of one of the samples of genotyped individuals used in Almudevar and Field (1999).

Shrimp: The tiger shrimp *Penaeus monodon* dataset (Jerry et al., 2006) consists of 59 individuals from 13 families with 7 loci. This dataset has 2.66% missing alleles.

Table 1: Characteristics of the biological datasets

Dataset	No. of individuals	No. of groups	No. of loci	No. of types of alleles per locus	Missing alleles (%)
Salmon	351	6	4	(9, 11, 9, 7)	0.00
Shrimp	59	13	7	(20, 18, 12, 7, 23, 9, 16)	2.66
Fly	190	6	2	(7, 7)	37.89
Ant	377	10	6	(22, 16, 15, 3, 5, 8)	9.00
Turtle	175	26	3	(5, 13, 10)	16.38
Turtle-m[a]	55	9	3	(5, 9, 8)	12.12

[a] Turtle-m is a subset of turtle dataset by removing most indefinite sibling groups.

Fly: The *Scaptodrosophila hibisci* dataset (Wilson et al., 2002) consists of 190 individuals in the same generation from 6 families sampled at various number of loci with up to 8 alleles per locus. All individuals shared 2 sampled loci which were chosen for our study. This dataset has 37.89% missing alleles.

Ant: The *Leptothorax acervorum* dataset (Hammond et al., 2001) are from a haplodiploid species. This is a subset of a sample used in Wang (2004), which consists of 377 worker diploid ants. This dataset has 9% missing alleles.

Turtle: Kemp's ridleys sea turtle dataset, *Lepidochelys kempi*, is polyandrous species and was sampled from 26 mothers and offspring groups at 3 loci (Kickler et al., 1999). There are 16.38% missing alleles. The other dataset we used (**Turtle-m**) is a subset obtained by eliminating most violated and indefinite sibling groups. There are 12.12% missing alleles.

4.2 Performance Assessment

In the sibling reconstruction studies, the true sibling groups are used for assessing the performance of the tested algorithms. The reconstruction accuracy measures the percentage of individuals correctly assigned to sibling groups compared with the true sibling groups (Gusfield, 2002). The problem can be modeled as a maximum linear assignment problem (MLAP) as follows. Denote a set of true sibling groups (ground truth) by J_A, indexed by j_a. Denote a set of our reconstructed sibling groups by J_B, indexed by j_b. For every sibling group pair j_a and j_b, $c_{j_a j_b}$ denotes the number of individuals appearing in both groups, i.e., $c_{j_a j_b} = |I_{j_a} \cap I_{j_b}|$. We define a binary decision variable: $x_{j_a j_b} = 1$ if group j_a is assigned to

group j_b, and 0 otherwise. The mathematical programming formulation is given as follows:

$$\text{(MLAP)} \quad \max \quad \sum_{j_a \in J_A} \sum_{j_b \in J_B} c_{j_a j_b} x_{j_a j_b} \tag{25}$$

$$s.t. \quad \sum_{j_a \in J_A} x_{j_a j_b} \leq 1 \quad \forall\, j_b \in J_B, \tag{26}$$

$$\sum_{j_b \in J_B} x_{j_a j_b} \leq 1 \quad \forall\, j_a \in J_A, \tag{27}$$

$$x_{j_a j_b} \in \{0,1\} \quad \forall\, j_a \in J_A,\ j_b \in J_B. \tag{28}$$

The objective in Equation (25) maximizes the total number of individuals in the group assignment from set J_B to set J_A. The constraints in Equations (26) and (27) ensure that each group in J_B (J_A, respectively) is assigned to at most one group in J_A (J_B, respectively). Here the reconstruction accuracy is reported as the percentage of individuals whose sibling groups are correctly reconstructed, that is, $\frac{\sum_{j_a \in J_A} \sum_{j_b \in J_B} c_{j_a j_b} x_{j_a j_b}}{|I|} \times 100\%$.

4.3 Implementation Settings

In our study, all programs were coded in MATLAB with synchronization of CPLEX version 10.0 in GAMS on the platform of an Intel Xeon Quad Core 3.0 GHz processor workstation with 8 GB RAM memory. The computational times reported were obtained from the desktop's internal timing calculations, which include time used for preprocessing and post-processing. The LP relaxation solution to the master problem in each column generation iteration was obtained using CPLEX barrier LP solver in order to reduce the heading-in and the tailing-off effects. There are six different configurations for the subproblem solution approach as shown in Table 2.

Table 2: Configurations for the subproblem solution approach.

	Approach in subproblem	
Subproblem	GSPP	GSCP
WMP	GSPP-WMP	GSCP-WMP
SMP	GSPP-SMP	GSCP-SMP
HYBRID	GSPP-HYBRID	GSCP-HYBRID

In our experiments, we first employed the proposed column generation approach to obtain the optimal LP solution of Min-SCP, and then we solved an integer programming formulation of Min-SCP containing all the generated columns. If the IP solution was shown to be optimal

(with respect to the optimal LP solution), the final reconstruction results were reported. Otherwise, we further performed a branch-and-price to obtain the optimal integral solution.

In column generation, to prevent the degeneracy, we perturb the coefficients of the objective function in the master problem of Min-SCP. We rewrite the objective function in Equation (1) as $\sum_{j\in J}\sigma_j z_j$, where σ_j is a random variable uniformly distributed between $[1-\epsilon, 1+\epsilon]$ and ϵ is a small positive number. The perturbation is only executed when there is no improvement on the objective function of the master problem after a consecutive number of iterations. The stopping criteria are set as follows: (1) an LP solution is optimal if the optimality condition is met and (2) there is no improvement in the objective value after 50 iterations, including 10 perturbation iterations.

In the branch-and-price framework, we set the computational time $T(L)$ limited at nodes, where L is the level of a node in the search tree. $T(0) = 5$ hours at the root node, and $T(L) = \max\{0.25, T(0)/2^{(L-1)}\}$ at level L. The total computational time is limited to 20 hours.

4.4 Reconstruction Results

Table 3 presents the results of the proposed approaches tested for all instances. The LP and IP solutions and reconstruction accuracies are reported. It is obvious from the table that there were a relatively small number of columns generated for the Min-SCP. Note that those columns that were repeatedly generated were not taken into account. When solving the LP relaxation of the Min-SCP, the column generation algorithms were terminated for almost all instances by the degeneracy condition, i.e., no improvement in the objective value after 50 iterations. There are 31 out of 36 instances in which the optimal IP solutions were found after the column generation was finished within 5 hours. For the remaining instances, the optimality condition was not proved and the branch-and-price approach was applied. Finally, our approaches were shown to result in very high reconstruction accuracies in most instances. Particularly, the best accuracies for all instances were obtained using GSCP-SMP. This shows that the effectiveness of incorporating the similarity measure was very prominent, and the overall reconstruction accuracies were increased, especially for the turtle dataset. In fact, 100% accuracies were obtained for the shrimp and ant datasets. It is noted that for the fly and turtle datasets, unstable results of IP solutions and accuracies obtained may result from the large percentages of missing values in these data.

It is interesting to note that although the LP and IP objective values provided by the hybrid approach are as good as the objective values provided by SMP approach, the accuracy of the hybrid solutions was worse than the SMP solutions in a few experiments. The reason is that the hybrid method could generate some columns with low similarities using WMP that does not include the similarity measure in the model. It is possible to select such groups in the final IP solution when multiple optimal solutions exist. This is because the Min-SCP model only accounts for the number of the selected groups, but not the similarity scores of the selected groups. Since sibling groups with low similarities are likely to reduce the accuracy of a final solution, the hybrid approach could generate less accurate solutions than the SMP approach.

Table 3: The results of our proposed approaches on real biological datasets. The results of the solution information of the LP relaxation are in the columns under 'Column Generation'. The IP solutions obtained from the best node in the brand-and-price are under 'Branch-and-Price'. The accuracy measures of the IP solution is given in the last column ('Assessment Accuracy').

		Column Generation				Branch-and-Price			Solution Assessment	
Dataset	Subproblem method	No. of new columns	LP val	IP val	Time (Sec.)	No. of nodes	IP val	Time (Sec.)	True No. of groups	Accuracy (%)
Salmon	GSCP-WMP	140	7.00	7	807				6	98.29
	GSPP-WMP	79	7.00	7	366				6	98.29
	GSCP-SMP	3	7.00	7	2,091				6	98.29
	GSPP-SMP	19	7.00	7	11,860				6	98.01
	GSCP-HYBRID	13	7.00	7	3,477				6	98.01
	GSPP-HYBRID	26	7.00	7	11,565				6	98.29
Shrimp	GSCP-WMP	392	13.00	13	3,539				13	100.00
	GSPP-WMP	163	13.00	13	3,147				13	100.00
	GSCP-SMP	2	13.00	13	1,262				13	100.00
	GSPP-SMP	1	13.00	13	4,420				13	100.00
	GSCP-HYBRID	15	13.00	13	1,549				13	100.00
	GSPP-HYBRID	12	13.00	13	6,275				13	100.00
Fly	GSCP-WMP	2,638	5.77	7	10,739	19	6	30,964	6	62.63
	GSPP-WMP	64	7.50	8	94				6	82.11
	GSCP-SMP	93	7.00	7	649				6	84.74
	GSPP-SMP	66	6.40	7	890				6	69.47
	GSCP-HYBRID	83	7.00	7	650				6	75.26
	GSPP-HYBRID	63	6.40	7	1,125				6	66.84
Ant	GSCP-WMP	543	10.00	10	3,890				10	99.73
	GSPP-WMP	109	11.00	11	1,496				10	93.10
	GSCP-SMP	12	10.00	10	2,998				10	100.00
	GSPP-SMP	58	10.00	10	15,558				10	97.61
	GSCP-HYBRID	23	10.00	10	4,849				10	100.00
	GSPP-HYBRID	56	10.00	10	15,999				10	98.41
Turtle	GSCP-WMP	2,716	15.61	17	18,079	43	17	>72,000	26	53.71
	GSPP-WMP	1,150	15.55	18	18,026	43	18	>72,000	26	48.57
	GSCP-SMP	245	30.00	30	5,212				26	70.29
	GSPP-SMP	185	16.19	18	18,728	32	18	>72,000	26	55.43
	GSCP-HYBRID	236	30.00	30	5,475				26	70.29
	GSPP-HYBRID	185	16.19	18	18,739	32	18	>72,000	26	55.43
Turtle-m	GSCP-WMP	615	6.75	7	602				9	67.27
	GSPP-WMP	123	7.14	8	293				9	74.55
	GSCP-SMP	36	10.00	10	689				9	83.64
	GSPP-SMP	43	6.80	7	357				9	69.09
	GSCP-HYBRID	30	10.00	10	369				9	81.82
	GSPP-HYBRID	57	6.80	7	781				9	67.27

In Figure 3, we illustrate the behaviors of IP/LP solutions and accuracies. The ant dataset is used as an example with implementing GSCP-WMP and GSCP-SMP. Compared to GSCP-WMP, GSCP-SMP integrating the similarity measure, resulted in more stable and better solutions in a small number of iterations.

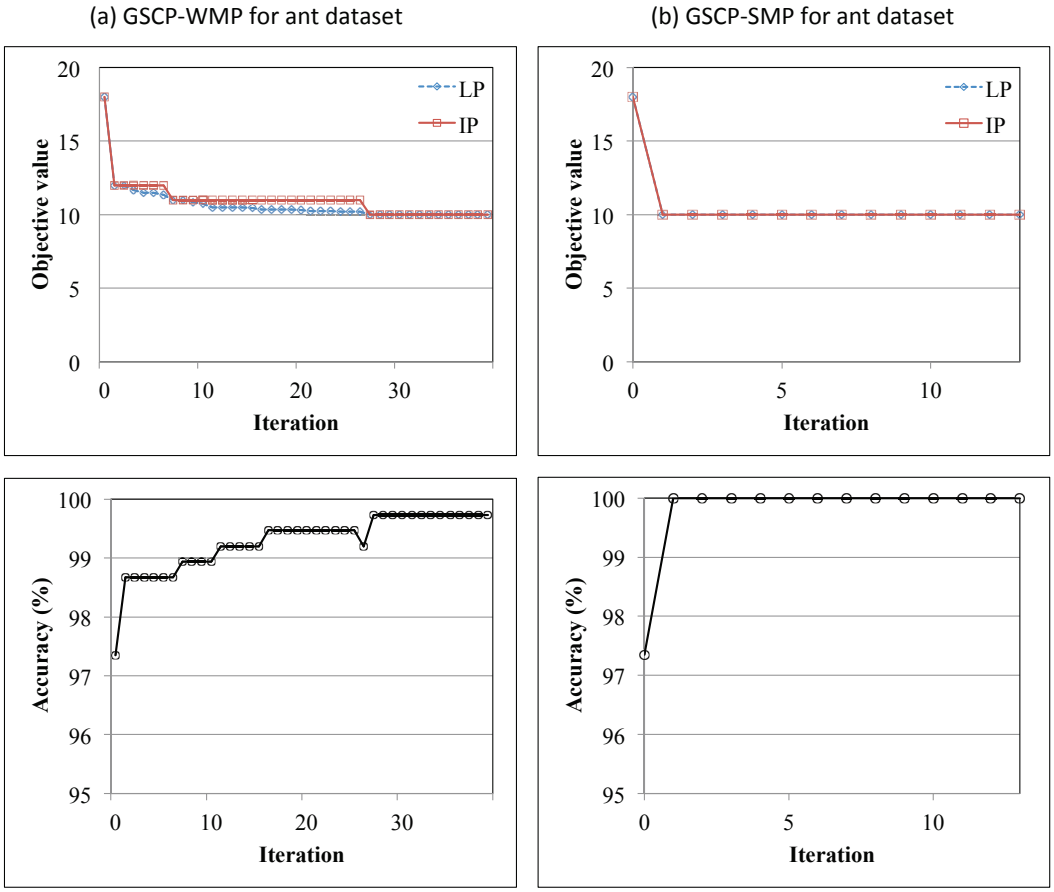

Figure 3: Objective function values and accuracies throughout the column generation iterations from (a) GSCP-WMP and (b) GSCP-SMP for the ant dataset.

The reconstruction of siblings is negatively affected by missing allele data as mentioned before. In this study, the missing alleles were treated as a wild card, which can represent any alleles when compared to the other alleles in a group. Therefore, the reconstruction of siblings in a group would not be overestimated and the worst-case reconstruction was then promised. Surprisingly, for the fly dataset, a reconstruction with high accuracy was obtained although there were many missing alleles. In addition, note that there were violations of the 2-allele constraints in the salmon dataset. After correcting these violations, we obtained the reconstruction result with 100% accuracy.

4.5 Comparison with Existing Approaches

In this section, we compare the reconstruction results obtained by the proposed approaches with other existing methods for the SRP. In Table 4, we report the comparison of the accuracies of GSCP-SMP (chosen from the best results among our experiments) to existing methods. The 2-allele optimization model (2AOM) is a mixed integer programming (MIP) model of the SRP based on the 2-allele constraints, without considering the similarity measure (Chaovalitwongse et al., 2010). The 2AOM construction results were obtained by using CPLEX to directly solve the MIP of 2AOM. The iterative maximum covering set (IMCS) is an iterative greedy algorithm that our group proposed to efficiently find good solutions to the 2AOM (Chaovalitwongse et al., 2010). The BWG algorithm is a brute-force approach that enumerates all possible maximal sibling groups that satisfy the 2-allele condition, and subsequently to solve a set covering problem to find the minimum set of sibling groups (Berger-Wolf et al., 2007). The A&F algorithm is a combinatorial approach that exhaustively enumerates all possible sibling groups satisfying the 2-allele condition (although the authors did not explicitly state the condition) and to obtain a maximal, not necessarily optimal, collection of sibling groups (Almudevar and Field, 1999). The B&M approach is based on a mixture of likelihood and combinatorial techniques, in which the SRP is presented as a graph problem; individuals as nodes and edges weighted by the pairwise likelihood (relatedness) ratio. The KINGROUP (KG) algorithm is an approach that estimates the likelihood 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 (Konovalov et al., 2004). The COLONY approach uses the maximum likelihood method to assign sibling groups and parentage jointly (Wang, 2004). As shown in Table 4, our approach achieved better and more robust reconstruction solutions than all other existing methods, especially for the datasets with many missing alleles (e.g., fly and turtle). Although the BWG approach can provide reconstruction results as good as GSCP-SMP for some datasets, it required more than 48 hours of computational time.

Furthermore, in Table 5, we present the comparison of the computational times of GSCP-SMP to the combinatorial approaches (based on the 2-allele condition) such as 2AOM, IMCS, BWG, and A&F. It is prominent in the table that our approach was far more efficient because it only generated highly accurate sibling groups. Our GSCP-SMP approach was able to find the optimal IP solutions in much shorter time whereas solving the 2AOM model

Table 4: Recovery values of true full sibling groups (accuracy) when comparing our method with other existing approaches in five different species. The best accuracies are marked in boldface.

	Combinatorial Approach					Statistical Approach		
Dataset	GSCP-SMP[a]	2AOM	IMCS	BWG	A&F	B&M	KG	COLONY
Salmon	**98.29**	94.02	**98.29**	**98.29**	–[d]	**98.29**	94.60	56.70
Shrimp	**100.00**	96.61	**100.00**	**100.00**	67.80	**100.00**	77.97	**100.00**
Fly	**84.74**	66.84	47.37	–[c]	31.05	19.62	54.73	–[c]
Ant	**100.00**	–[b]	93.10	**100.00**	–[d]	97.61	97.10	**100.00**
Turtle	**70.29**	–[b]	40.00	48.00	–[d]	38.18	39.40	40.00
Turtle-m	**83.64**	47.27	61.82	–[c]	–[d]	–[c]	–[c]	–[c]

[a] The best configuration was chosen from Table 3.

[b] There are no results acquired within the time limit of 72,000 seconds.

[c] There are no results available.

[d] A&F ran out of 8GB memory as it enumerates all possible sibling groups.

Table 5: Comparison of computational time (in second) of the proposed approach to the previous combinatorial approaches based on the time limit of 20 hours.

Dataset	GSCP-SMP[a]	2AOM[b]	IMCS	BWG[c]	A&F[d]
Salmon	8,382	>72,000	130	>72,000	-
Shrimp	4,367	>72,000	150	>72,000	2
Fly	2,648	>72,000	20	>72,000	>72,000
Ant	14,667	>72,000	505	>72,000	-
Turtle	16,105	>72,000	78	>72,000	-
Turtle-m	2,111	>72,000	12	>72,000	-

[a] The best configuration was chosen from Table 3.

[b] The gaps between IP and LP solutions are 63%, 67%, 55%, -, -, and 60%, respectively, when time limit is reached.

[c] BWG ran over the time limit of 72,000 seconds for all datasets.

[d] A&F ran over the time limit of 72,000 seconds for the fly dataset and ran out of 8GB memory for most datasets.

directly by CPLEX resulted in large solution gaps between the IP and LP solutions after the 20-hour time limit. The IMCS approach was very fast but the quality of reconstruction solutions were not as accurate and robust as the ones by our approach. The BWG approach provided reconstruction results as accurate as those by our GSCP-SMP approach because it is based on a brute-force enumeration of maximal sibling groups. However, it required more than 48 hours of computational time. The A&F approach, which is based on sibling group enumeration, ran out of memory in most cases, except the easy shrimp dataset.

It is also important to note that the main goal of this study to present an approach that is both efficient and accurate. On one hand, we proposed an approach to efficiently solve an optimization model of SRP. Our approach is far more efficient than all but one

(i.e., IMCS) combinatorial approaches. On the other hand, we proposed an approach to accurately provide reconstruction results that are more accurate than those of statistical approaches. In the literature, it is not common to compare the computational efficiency between combinatorial approaches and statistical approaches. This is because some statistical approaches use simplistic statistical models, making them very fast to solve but inaccurate. In our computational experiments, it was consistently observed that both the B&M and KG approaches provided reconstruction results very fast but the results were inaccurate. COLONY, on the other hand, used a very sophisticated statistical model to provide relatively accurate solutions but required more than 3 days of CPU time (far exceeding that required by our approach). In addition, it is difficult to fairly compare exact computational times of statistical approaches because some run solely on Windows and some run solely on Linux.

5. Conclusion

When solving the SRP using microsatellite genotype data, existing combinatorial and statistical methods in the literature face challenges of computational complexity and reconstruction accuracy. In this study, an exact, integrated framework was successfully developed to overcome this challenge. Specifically, this study presented an optimization model for the SRP that integrates the combinatorial constraints of Mendel's laws with the statistical likelihood of the genetic data. We proposed a new column generation approach with a branch-and-price method to solve the SRP, in which the master problem finds the minimum set of sibling groups and three variants of the subproblem solution approach were proposed to efficiently generate high-quality sibling groups with high similarity scores. The proposed approaches were tested on real biological datasets, and their performances were compared with other recent SRP approaches in the literature. The results demonstrated that our approaches outperformed the alternative approaches and were be more robust.

In addition to full-sibling reconstruction, there are various related questions in pedigree reconstruction that also encounter the common challenges of accuracy and efficiency, such as half-sibling reconstruction (where only one parent is shared) (Sheikh et al., 2010) and multi-generational relationship reconstructions (Won et al., 2012). We plan to extend our approach to meet the challenges of related pedigree problems in our future work.

Acknowledgments

This research was supported by the following grants: NSF IIS-0611998 (Chaovalitwongse), NSF CCF-0546574 (Chaovalitwongse), NSF IIS-1219638 (Chaovalitwongse) NSF IIS-0612044 (Berger-Wolf, Ashley, DasGupta), NSF IIS-1064681 (Berger-Wolf, Ashley, DasGupta, Khokar) and Fullbright Scholarship (Sheikh). We are grateful to the people who have shared their data with us: Jeff Connor, Atlantic Salmon Federation, Dean Jerry, and Stuart Barker. We would also like to thank Anthony Almudevar, Bernie May, and Dmitri Konovalov for sharing their software. Finally, we especially thank the associate editor and the reviewers for very detailed and insightful suggestions on the present work.

References

Adams, W. P., R. J. Forrester. 2007. Linear forms of nonlinear expressions: New insights on old ideas. *Operations Research Letters* **35** 510–518.

Almudevar, A. 2003. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology* **63** 63–75.

Almudevar, A. 2007. A graphical approach to relatedness inference. *Theoretical Population Biology* **71** 213–229.

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

Barnhart, C., E. L. Johnson, G. L. Nemhauser. 1998. Branch-and-price: Column generation for solving huge integer programs. *Operations Research* **46** 316–329.

Berger-Wolf, T. Y., B. DasGupta, W. Chaovalitwongse, M. V. Ashley. 2005. Combinatorial reconstruction of sibling relationships. *Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics (CBGI 05)*. 1252–1255.

Berger-Wolf, T. Y., S. I. Sheikh, B. DasGupta, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, S. Lahari Putrevu. 2007. Reconstructing sibling relationships in wild populations. *Bioinformatics* **23** 49–56.

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

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

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** 1589–1600.

Chaovalitwongse, W., T. Y. Berger-Wolf, B. DasGupta, M. V. Ashley. 2007. A robust combinatorial approach for sibling relationships reconstruction. *Optimization Methods and Software* **22** 11–24.

Chaovalitwongse, W., C.-A. Chou, T. Y. Berger-Wolf, B. DasGupta, S. I. Sheikh, S. Lahari Putrevu, M. V. Ashley, I. C. Caballero. 2010. New optimization model and algorithm for sibling reconstruction from genetic markers. *INFORMS Journal on Computing* **22** 180–194.

Chaovalitwongse, W., P. M. Pardalos, O. A. Prokopyev. 2004. A new linearization technique for multi-quadratic 0-1 programming problems. *Operations Research Letters* **32** 517–522.

Chou, C.-A., W. Art Chaovalitwongse, T. Y. Berger-Wolf, B. DasGupta, Mary V. Ashley. 2012. Capacitated clustering problem in computational biology: Combinatorial and statistical approach for sibling reconstruction. *Computers & Operations Research* **39** 609–619.

Desrosiers, J., F. soumis, M. Desrochers. 1984. Routing with time windows by column generation. *Networks* **14** 545–565.

Gusfield, D. 2002. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters* **82** 159–164.

Hammond, R. L., A. F. G. Bourke, M. W. Broford. 2001. Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum*. *Molecular Ecology* **10** 2719–2728.

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** 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** 542–547.

Kickler, K., M. T. Holder, S. K. Davis, R. Márquez-M, D. W. Owens. 1999. Detection of multiple paternity in the Kemp's ridley sea turtle with limited sampling. *Molecular Ecology* **8** 819–830.

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** 779–782.

Mendel, G. 1901. Experiments on plant hybridization (versuche ber pflanzen-hybriden). *Journal of the Royal Horticultural Society* **26** 1–32.

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

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

Ryan, D., B. Foster. 1981. An integer programming approach to scheduling. A. Wren, ed., *Computer Schedule of Public Transport Urban Passenger Vehicle and Crew Scheduling*. Elsevier Science B. V., 269–280.

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. *7th Annual International Conference on Computational Systems biology* .

Sheikh, S. I., T. Y. Berger-Wolf, A. Khokar, C.-A. Chou, W. Chaovalitwongse, M. V. Ashley, I. C. Caballero, B. DasGupta. 2010. Combinatorial reconstruction of half-sibling groups: Models and algorithms. *Journal of Bioinformatics and Computational Biology* **8** 1–20.

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** 1329–1338.

Thomas, S. C., W. G. Hill. 2002. Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetic Research* **79** 227–234.

Vanderbeck, F.çois. 1994. *Decomposition and Column Generation for Intger Programs*. Ph.D Thesis, Universite Catholique de Louvain, Belgium.

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

Wang, J., A. W. Santure. 2009. Parentage and sibship inference from multi-locus genotype data under polygamy. *Genetics* **181** 1579–1594.

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

Won, D., C.-A. Chou, W. A. Chaovalitwongse, T. Y. Berger-Wolf, B. DasGupta, A. A. Khokhar, M. Maggioni, M. V. Ashley, J. Palagi, S. I. Sheikh. 2012. An integrated optimization framework for inferring two-generation kinships and parental genotypes from microsatellite samples. *Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. 392–399.