# Capacitated Clustering Problem in Computational Biology: Combinatorial and Statistical Approach for Sibling Reconstruction

Chun-An Chou

*Department of Industrial and Systems Engineering, Rutgers University, New Jersey, USA*

Wanpracha Art Chaovalitwongse

*Department of Industrial and Systems Engineering, Rutgers University, New Jersey, USA*
*Department of Industrial and Systems Engineering, University of Washington, Washington, USA*
*Department of Operations Research and Financial Engineering, Princeton University, New Jersey, USA*

Tanya Y. Berger-Wolf, Bhaskar DasGupta

*Department of Computer Science, University of Illinois, Chicago, Illinois, USA*

Mary V. Ashley

*Department of Biology, University of Illinois, Chicago, Illinois, USA*

## Abstract

The capacitated clustering problem (CCP) has been studied in a wide range of applications. In this study, we investigate a challenging CCP in computational biology, namely, sibling reconstruction problem (SRP). The goal of SRP is to establish the sibling relationship (i.e., groups of siblings) of a population from genetic data. The SRP has gained more and more interests from computational biologists over the past decade as it is an important and necessary keystone for studies in genetic and population biology. We propose a large-scale mixed integer formulation of the CCP for SRP, that is based on both combinatorial and statistical genetic concepts. The objective is not only to find the minimum number of sibling groups, but also to maximize the degree of similarity of individuals in the same sibling groups while each sibling group is subject to genetic constraints derived from the Mendel's laws. We develop a new randomized greedy optimization algorithm to effectively and efficiently solve this SRP. The algorithm consists of two key phases: construction and enhancement. In the construction phase, a greedy approach with randomized perturbation is applied to construct multiple sibling groups iteratively. In the enhancement phase, a two-stage local search with a memory function is used to improve the solution quality with respect to the similarity measure. We demonstrate the effectiveness of the proposed algorithm using real biological data sets and compare it with state-of-the-art approaches in the literature. We also test it on larger simulated data sets. The experimental results show that the proposed algorithm provide the best reconstruction solutions.

*Key words:* Clustering analysis, capacitated clustering problem, combinatorial optimization, sibling reconstruction, computational biology

## 1. Introduction

The capacitated clustering problem (CCP) has been one of the most challenging problems in clustering research. Several variants of CCP have been studied in the literature including a capacitated centred clustering problem (CCCP) as well as a capacitated $p$-median problem (CPMP). The CCP can be formally defined as follows. Given a set of data points with associated weights (or features), the CCP is to partition the data points into clusters such that the total weight of data points in each cluster does not exceed the capacity limit of the cluster. In general, the objective of CCP is to maximize the homogeneity (similarity) of the data points in each cluster or to maximize the separation (dissimilarity) among different clusters (Hansen and Jaumard, 1997). Although clustering techniques have been essential tools to solve many practical problems, previous studies on the CCP are mostly applied to facility location problems and they often focus on the development of solution algorithms. In this study, we consider an application of the CCP in computational biology to solve the sibling reconstruction problem (SRP). The SRP is one of the most important problems in genetic biology (Blouin, 2003). The knowledge of sibling relationships allows biologists to understand the fundamental biological phenomena including mating systems, ecological behaviors and evolutions, social organizations, etc. The SRP can be formally defined as follows. Given a population (or set) of individuals (or data points) with associated genetic features (or weights), the SRP is to partition individuals into sibling groups (or clusters) such that the combination of features in each sibling group does not violate the genetic constraints of sibling group (i.e., the Mendel's laws). The common objective of SRP is to maximize the similarity degrees of individuals in sibling groups and to minimize the number of sibling groups in the population. The genetic constraints of the sibling group make the SRP far more complicated than a standard CCP. The capacity constraint in the CCP only incorporates the total weight of data points in each cluster, which is usually one-dimensional. On the other hand, the genetic constraints of SRP are derived from the combinatorial concept of the Mendel's laws, which consider multiple pairwise genetic features of individuals. In addition, the actual number of sibling groups is not known a priori, and overlapping sibling groups are commonly seen in natural populations because wild animals are not always monogamous. In other words, an individual can be assigned to more than one sibling group simultaneously if the genetic constraints are satisfied.

To solve the SRP, most studies employ statistical likelihood techniques from genetic data (Painter, 1997; Smith et al., 2001; Thomas and Hill, 2002; Butler et al., 2004; Wang, 2004; Konovalov et al., 2004; Wang and Santure, 2009), while heuristic approaches are developed to integrate statistical likelihood such as graph-based approaches (Almudevar and Field, 1999; Beyer and May, 2003; Almudevar, 2007), simulated annealing (Almudevar, 2003), etc. More recently, there has been an increasing degree of interest to apply combinatorial concepts to the SRP. A new optimization model with complex combinatorial constraints derived from the Mendel's laws and its computational algorithms have been proposed (Berger-Wolf et al., 2005, 2007; Chaovalitwongse et al., 2007, 2010). The model is formulated in the format of a set covering problem (SCP), which is to find a minimum set of sibling groups subject to the combinatorial constraints. Note that, in those studies, only the combinatorial constraints and the concept of parsimony, which is to minimize the number of sibling groups, were considered in the model. More importantly, statistical similarity measure from genetic features of individuals can provide direct information to benefit the sibling relationships, while the combinatorial constraints give the robustness of reconstructing sibling groups.

In this study, we propose a new heuristic optimization algorithm, which has similar concept to a greedy randomized adaptive search procedure (GRASP) (Feo and Resende, 1995), that integrates the combinatorial constraints and the concept of parsimony with a statistical similarity measure. The proposed framework involves the following phases: the construction of clusters and the enhancement of quality of clusters. In the first phase, an efficient greedy approach, proposed by Chaovalitwongse et al. (2010), is employed repeatedly to construct a number of different possible partitions of (disjoint) sibling groups by introducing a randomized perturbation. Subsequently, among all possible partitions of sibling groups, a set covering problem (SCP) is solved to select the minimum set of sibling groups to cover the population. In the second phase, we propose a new two-stage local search with a memory function to improve the quality of sibling reconstruction based on the similarity of individuals in the sibling groups. Finally, a SCP is solved again to find the minimum number of sibling groups.

The remainder of this paper is organized as follows. In Section 2, we present the CCP and the sibling reconstruction problem with basic genetic terminologies and a mathematical representation of genetic data. The capacitated clustering formulation for the SRP is presented in detail in Section 3. The mathematical programming model with the combinatorial constraints of the Mendel's laws is formulated to integrate the statistical similarity measure of genetic data. In Section 4, we elaborate the proposed framework designed to solve the SRP. In Section 5, we demonstrate computational experiments using real biological data sets and simulated data sets, and show the effectiveness by comparing the proposed approach to the existing methods. Finally, this paper is concluded in Section 6.

## 2. Background

### *2.1. Capacitated Clustering Problem*

The mathematical model of the CCP was first proposed by Mulvey and Beck (1984) and its variants were used to study several practical problems in diverse applications. Here we consider one of the most common variants of CCP. Given a set of data points $i \in I$ with associated positive weights $\pi_i$ and resources $c_i$, and a set of edges $(i, i') \in E$ with associated positive weights (e.g., similarities) $w_{ii'}$, where $i \neq i'$. Assume that there is a set of clusters $j \in J$ used to cover (represent) all data points. Let $p$ be a predefined number of clusters. There is a resource limitation $W_j$ on each cluster $j$. The objective of CCP is to find a set of clusters with the maximum weight (or similarity) per cluster subject to a resource capacity.

Define $x_{ij}$ and $z_j$ as binary variables, where $x_{ij} = 1$ if data point $i$ is assigned to cluster $j$, and $x_{ij} = 0$ otherwise; $z_j = 1$ if cluster $j$ is selected, and $z_j = 0$ otherwise. The formulation of CCP is given in Equations (1)-(7). The objective in Equation (1) is to maximize the total weight of all selected clusters. The constraint set in Equation (2) calculates the total weight of data points assigned to cluster $j$. The constraint set in Equation (3) ensures that every data point is assigned to one cluster, while the constraint set in Equation (4) guarantees that a cluster must be selected if there is any data point assigned to it. The constraint set in Equation (5) ensures that only $p$ clusters are selected. The constraint set in Equation (6) is a knapsack constraint ensuring that the total resource of data points assigned to a cluster does not violate its capacity.

$$
\begin{aligned}
\text{(CCP)} \quad \max \quad & \sum_{j \in J} W_j z_j && (1)\\
\text{s.t.} \quad & W_j = \sum_{i \in I} \pi_i x_{ij} + \sum_{(i,\, i') \in E} w_{ii'} x_{ij} x_{i'j} \quad \forall\, j \in J && (2)\\
& \sum_{i \in I} x_{ij} \;=\; 1 \quad \forall\, j \in J && (3)\\
& x_{ij} \;\le\; z_j \quad \forall\, i \in I,\, j \in J && (4)\\
& \sum_{j \in J} z_j \;=\; p && (5)\\
& \sum_{i \in I} c_i x_{ij} \;\le\; C_j \quad \forall\, j \in J && (6)\\
& x_{ij},\, z_j \;\in\; \{0, 1\}. && (7)
\end{aligned}
$$

In the literature, exact solution methods have been proposed to solve different versions of CCP. Mehrotra and Trick (1998) used a column generation with a specialized branching technique and solved a maximum weighted cluster problem (MWCP) in the subproblem. Baldacci et al. (2002) presented a new exact algorithm by modeling the capacity location problem as a set partitioning problem with cluster-feasibility constraints. Lorena and Senne (2004) proposed an approach that integrates the column generation and Lagrangean/surrogate relaxation techniques to solve capacitated $p$-median problems. More recently, Ceselli et al. (2009) proposed a computational framework based on column generation and branch-and-price approaches to solve the capacitated network problems. Due to the computational complexity of real-life CCPs, a large number of heuristic approaches have been developed. Those include classical sub-gradient heuristics (Mulvey and Beck, 1984; Koskosidis and Powell, 1992), simulated annealing and tabu search (Franca et al., 1999; Osman and Creistofides, 2002), bionomic approach (Maniezzo et al., 1998), cluster search (Chaves and Lorena, 2010), GRASP-based algorithms (Samad and Osman, 2005; Deng and Bard, 2010), and other heuristics (Osman and Samad, 2002; Samad and Osman, 2002; Negreiros and Palhano, 2006; Avella et al., 2009).

### *2.2. Sibling Reconstruction Problem (SRP)*

In genetic and population biology, as more and more genetic markers become available for a wider range of species, biological researchers have attempted to better characterize evolutionary, ecological, population, and demographic parameters. With the knowledge of sibling relationships from genetic markers such as microsatellites, population biologists will be able to better understand the nature and organism behaviors such as the number of reproducing adults, their fecundity, and the average size of litters. For endangered species, this knowledge is important for conservation and management strategies. The sibling reconstruction problem can be defined as a problem of identifying sibling relationships from genotypic data where the organisms are sampled and genotyped without information about their parents.

There are several genetic markers used in population genetics, and *microsatellite* is one of the most widely used for the sibling reconstruction. Microsatellites are neutral and co-dominantly inherited, and allow the direct inference of genotypes at each locus (i.e., a site in the chromosome). Figure 1 displays an example. In this study, we restrict to *diploid individuals*, which are organisms that have a pair of alleles at each locus of a chromosome pair. Genotypes are comprised of *alleles*, which are distinct length variants of microsatellites. A *locus* is a site which the allele occupies on the chromosome. In an individual, homozygous (respectively, heterozygous) allele(s) represents a pair of identical (respectively, different) alleles at a particular locus.

Chaovalitwongse et al. (2010) proposed a multi-dimensional data structure to represent microsatellites mathematically. A multi-dimensional matrix is defined as $a_{ik}^{l} \in \{0, 1, 2\}$, where $i \in I$ is a set of individuals; $l \in L$ is a set of loci; and $k \in K$ is a set of alleles. This matrix reveals the indication of distinct alleles ($a_{ik}^{l} = 1$) at a locus as well as homozygous alleles ($a_{ik}^{l} = 2$). We show the representation in Figure 2. For example, $a_{45}^{1} = 1$ indicates that

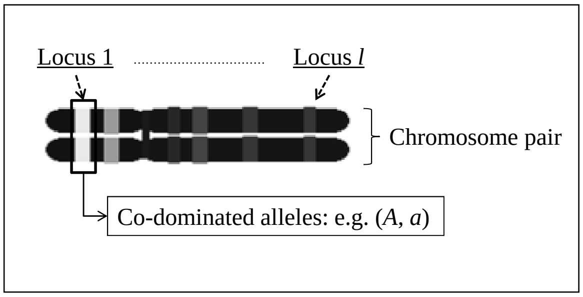

Figure 1: An illustration of a chromosome pair with several loci from a diploid individual. Genotype is co-dominated by a pair of alleles at each locus.

shrimp 4 has a distinct allele 5 while $a^2_{22} = 2$ indicates that shrimp 2 has homozygous alleles 12.

| Individual | Locus 1 | Locus 2 |
|---|---|---|
| Shrimp 1 | 1/2 | 11/13 |
| Shrimp 2 | 2/3 | 12/12 |
| Shrimp 3 | 3/3 | 11/12 |
| Shrimp 4 | 4/5 | 11/14 |
| Shrimp 5 | 6/7 | 14/16 |
| Shrimp 6 | 4/7 | 17/17 |

$\Longrightarrow$

|  | Locus 1 |  |  |  |  |  |  | Locus 2 |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Allele | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 11 | 12 | 13 | 14 | 16 | 17 |
|  | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
|  | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
|  | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
|  | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
|  | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
|  | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |

Figure 2: A multidimensional matrix interprets microsatellite markers from a population of six individuals with two loci. There are seven and six distinct alleles respectively.

When familial (parental) information is not known, genetic markers become a direct reference for identifying sibling groups. The sibling relationships are often inferred by using the statistical likelihood from the knowledge of genetic typical allele distribution and frequency. In the literature, statistical methods for SRP are categorized into pairwise and group approaches. The pairwise approaches infer the relationship of a pair of two individuals based of the genotypes. The group approaches take into account all individual simultaneously in the group partitions. To solve the SRP with the statistical likelihood concept, several methods have been proposed in the literature. Painter (1997) proposed a Bayesian approach to estimated sibling relationships in a generation and explore its feasibility. The technique proposed by Almudevar and Field (1999) is based on the exclusion principle to evaluate the set of feasible sibling groups. Smith et al. (2001) used Markov chain Monte Carlo (MCMC) algorithms to find a partition of sibling groups using pairwise likelihood ratios. Subsequently, Thomas and Hill (2002) applied the MCMC technique to reconstruct the nested structure of full-sibling within half-sibling relationships. Butler et al. (2004) compared existing methods (Painter, 1997; Almudevar and Field, 1999; Smith et al., 2001; Thomas and Hill, 2000, 2002) with a new Simpson algorithm, and compared performances in terms of accuracy, efficiency, and robustness. They have computational limitations on large scale data. Konovalov et al. (2004) developed an efficient computer program called KINGROUP using the likelihood formulas proposed by Queller and Goodnight (1989) and subsequently presented a modified version of Simpson algorithm (Konovalov et al., 2005).

Mendel's laws (also called Mendelian inheritance laws) (Mendel, 1901; Bowler, 1989)have also played an important role in using the combinatorial concept to solve the SRP. An offspring inherits one allele from each of its parents (either father or mother) at each locus and the inheritance pattern of one trait co-dominated by a pair of alleles is independent of another one. There are two main combinatorial principles. Thus, for any offspring, one of its allele pair only has at most two possibilities from parents and there are at most four alleles appearing possibly at each locus (Berger-Wolf et al., 2007; Chaovalitwongse et al., 2010). These combinatorial rules on genetic patterns provide the robustness of the sibling reconstruction. Recently, a few studies in the literature proposed the use of combinatorial concept to solve the SRP (Beyer and May, 2003; Almudevar, 2007). Almudevar (2003) developed a simulated annealing approach to solve the combinatorial version of SRP. Our group proposed a series of optimization algorithms based on the combinatorial constraints from the Mendel's laws (Berger-Wolf et al., 2005, 2007; Chaovalitwongse et al., 2007). Our approaches enumerated all possible sibling groups by following the Mendel's laws and solved a set covering problem to find a minimum set of representative sibling groups, which is based on the parsimony assumption when the actually number of sibling groups is not known a priori. Most recently, Chaovalitwongse et al. (2010) proposed an iterative heuristic approach, IMCS, to solve a new optimization model (2AOM) with the combinatorial constraints to find a partition of maximal sibling groups.

## 3. Capacitated Clustering Model for Sibling Reconstruction Problem

### *3.1. Capacitated Clustering Model*

We formulate the SRP as a CCP by using the statistical likelihood measure as the objective function subject to the Mendelian combinatorial constraints. We note that this is the first mathematical model that integrates both statistical and combinatorial concepts to reconstruct the sibling relationship. We shall mathematically define our integrated problem as follows.

Given a set of individuals $i \in I$ with associated weights $\pi_i$ and a set of edges $(i,\ i') \in E$ with associated similarity measures $w^l_{ii'}$ over all loci $l \in L$, where $i \neq i'$. Assume that there is a set of sibling groups $j \in J$ to represent the relationship of the given population. Because there is no prior parental information, the number of sibling groups is not known and will have to be determined by the model. Next we define the following decision variables.

- $z_j \in \{0,\ 1\}$: indicates if there is individual(s) assigned to be a member of sibling group $j$;
- $x_{ij} \in \{0,\ 1\}$: indicates if individual $i$ is assigned to be a member of sibling group $j$;
- $y^l_{jk} \in \{0,\ 1,\ 2\}$: indicates if any member in sibling group $j$ has distinct ($y^l_{jk} = 1$) or homozygous ($y^l_{jk} = 2$) allele(s) $k$ at locus $l$;
- $v^l_{jkk'} \in \{0\ 1\}$: indicates if allele $k$ appears with allele $k'$ in sibling group $j$ at locus $l$.

*3.1.1. Statistical Similarity Measure as Objective Function*

The overall objective here is to reconstruct a set of sibling groups such that the total similarity degree and weight of individuals assigned to the selected sibling groups is maximized. The objective function is given by

$$\max \quad \sum_{j \in J} W_j z_j, \tag{8}$$

where $W_j$ is the sum of weight and similarity score for a sibling group $j$, which can be calculated by

$$W_j = \sum_{i \in I} \pi_i x_{ij} + \sum_{(i,\ i') \in E} (\sum_{i \in L} w^l_{ii'}) x_{ij} x_{ij} \quad \forall\ j \in J. \tag{9}$$

The above equation takes into account not only the weights of individuals assigned to the sibling group $j$ but also the pairwise similarity measures over all loci. The weight of each individual can be estimated from the prior information; however, in our case all individuals are equally weighed because of the small sample size. To calculate the pairwise similarity score, we apply a simple pairwise approach to score the similarity based on genetic features at loci between a pair of individuals. The pairwise score can be calculated by

$$w_{ii'}^l := \begin{cases} 1 & \text{if } \sum\limits_{k \in K} |a_{ik}^l - a_{i'k}^l| = 0; \\ 0.5 & \text{if } \sum\limits_{k \in K} |a_{ik}^l - a_{i'k}^l| = 2; \\ 0 & \text{if } \sum\limits_{k \in K} |a_{ik}^l - a_{i'k}^l| = 4. \end{cases} \tag{10}$$

The sum of similarity score $\sum\limits_{l \in L} w_{ii'}$ over all loci represents the degree of similarity for a pair of individuals $i$ and $i'$. The higher the degree, the more similar two individuals.

*3.1.2. Capacity Constraints: Combinatorial Rules from Mendel's Laws*

The capacity constraints of SRP are more complex than those of simple CCP's because the capacity constraints are multi-dimensional. That is, each capacity constraint must be satisfied for individual independent locus of a sibling group.

In Berger-Wolf et al. (2007), the 4-allele and 2-allele properties were first proposed based on the Mendel's laws. Chaovalitwongse et al. (2010) augmented 2-allele property with a tighter constraint. For mathematical representation, we formulate combinatorial constraints from the modified 2-allele property by employing an indication matrix, $a_{ik}^l \in \{0, 1, 2\}$. From the first rule of the Mendel's laws, the combinatorial constraints are given in Equations (11)-(12). Equation (11) ensures that the integer variable $y_{jk}^l$ for distinct or homozygous indication must be activated for the existence of distinct or homozygous allele(s) at locus $l$ in sibling group $j$. Equation (12) ensures that the number of distinct allele and the number of homozygous alleles is less than or equal to four.

$$a_{ik}^l x_{ij} \leq y_{jk}^l \quad \forall\, j \in J,\ k \in K,\ l \in L, \tag{11}$$

$$\sum_{k \in K} y_{jk}^l \leq 4 \quad \forall\, j \in J,\ l \in L. \tag{12}$$

From the second rule of the Mendel's laws, the combinatorial constraints are given in Equations (13)-(14). Equation (13) restricts that the binary variable for allele pair indication $v_{jkk'}^l$ must be activated for any assignment of individual $i$ to sibling group $j$. Equation (14) ensures that every allele in the group does not appear with more than two other alleles (excluding itself). A big M number is defined by $M = |I| + 1$.

$$\sum_{i \in I} a^l_{ik} a^l_{ik'} x_{ij} \leq M v^l_{jkk'} \quad \forall\, j \in J,\ k \in K,\ k' \in K \backslash k,\ l \in L, \tag{13}$$

$$\sum_{k' \in K \backslash k} v^l_{jkk'} \leq 2 \quad \forall\, j \in J,\ k \in K,\ l \in L. \tag{14}$$

For the rest of the paper, a so-called "feasible sibling group (or cluster)" is a set of individuals that satisfies the capacity constraints in Equations (11)-(14) at every locus.

### *3.1.3. Covering constraints*

For certain species in natural populations that do not belong to the monogamous mating system, the overlapping situation where any individual can be assigned to more than one sibling group are commonly seen. We therefore consider the covering constraint set instead of the partitioning constraint set in Equations (4)-(5). The covering constraint sets are given by

$$\sum_{i \in I} x_{ij} \geq 1 \quad \forall\, j \in J, \tag{15}$$

$$x_{ij} \leq z_j \quad \forall\, i \in I,\ j \in J. \tag{16}$$

Equation (15) ensures that every individual is assigned to at least one sibling group. Equation (16) ensures that the binary sibling group variable must be activated for the assignment of any individual $i$ to sibling group $j$.

It is noted that because the actual number of sibling groups is not known in general, in this study, we therefore employ the parsimony assumption to find the minimum number of sibling groups instead of using the constraint set in Equation (5). For this purpose, sibling group selection can be formulated as a set covering problem (SCP) that incorporates the covering constraints.

## *3.2. Preliminaries of Solving CCP for SRP*

According to the formulation in the previous subsection, the CCP for SRP can be considered as a complete optimization model (CCP-SRP) shown in Equations (17)-(22). The objective of CCP-SRP in Equation (17) integrates the minimization of sibling groups and the maximization of similarity degrees of individuals in the same sibling groups, where a balancing parameter $\theta$ is introduced between the two terms. The constraint sets in Equations (18)-(22) follow the same definitions described in the previous section.

$$
\begin{aligned}
\text{(CCP-SRP)} \quad \max \quad & \sum_{j\in J}(\theta W_j - 1)z_j && (17)\\
\text{s.t.} \quad & W_j = \sum_{i\in I}\pi_i x_{ij} + \sum_{(i,\, i')\in E}(\sum_{i\in L} w^l_{ii'})x_{ij}x_{ij} \quad \forall\, j\in J && (18)\\
& \sum_{i\in I} x_{ij} \geq 1 \quad \forall\, j\in J && (19)\\
& x_{ij} \leq z_j \quad \forall\, i\in I,\ j\in J && (20)\\
& a^l_{ik}x_{ij} \leq y^l_{jk} \quad \forall\, j\in J,\ k\in K,\ l\in L && (21)\\
& \sum_{k\in K} y^l_{jk} \leq 4 \quad \forall\, j\in J,\ l\in L && (22)\\
& \sum_{i\in I} a^l_{ik}a^l_{ik'}x_{ij} \leq Mv^l_{jkk'} \quad \forall\, j\in J,\ k\in K,\ k'\in K\backslash k,\ l\in L && (23)\\
& \sum_{k'\in K\backslash k} v^l_{jkk'} \leq 2 \quad \forall\, j\in J,\ k\in K,\ l\in L. && (24)
\end{aligned}
$$

The CCP-SRP is a mixed-integer nonlinear programming (MINLP) problem, which is viewed as a generalization of 2AOM. To solve the CCP-SRP, there are issues encountered such as highly computational complexity and the calibration of the parameter $\theta$. Firstly, let us look back on the optimization model 2AOM in Chaovalitwongse et al. (2010), which is to find a minimum number of sibling groups subject to capacity constraints and without the integration of statistical similarity measure. The 2AOM has been proved to be an *NP-hard* problem with many discrete variables and many constraints. It is hard to solve directly to obtain an optimal solution. According to our computational experiments, we failed to find a feasible solution to 2AOM in CPLEX after 20 hours of run. Consequently, it is not easy to calibrate the balancing parameter at a precise level, which plays a role in solving the SCP-SRP, when the value of similarity varies with assignments of individuals into different sibling groups. These observations and experiences have motivated us to develop an efficient heuristic method to solve this problem. In the next section, we thus propose a new greedy optimization heuristic to solve the decomposed CCP-SRP model in two phases.

## 4. Randomized Greedy Optimization Algorithm

In this study, we develop a new randomized greedy optimization algorithm (RGOA) to solve the CCP of SRP. The underlying concept behind the RGOA is motivated by the Greedy Randomized Adaptive Search Procedure (GRASP) (Feo and Resende, 1995). The RGOA is divided into two phases: construction and enhancement phases. Recall that the objective of CCP in Equation (8) and its total weight in Equation (9) contains two terms, the individual weight and the pairwise similarity, to be maximized. The individual weight of sibling group assignment is maximized in the construction phase while the pairwise similarity is maximized in the enhancement phase.

The flowchart of our RGOA is shown in Figure 3 and the associated pseudo-code is presented in Algorithm 1. In the construction phase, we modify an efficient approach, called IMCS, for the SRP (Chaovalitwongse et al., 2010) by introducing a randomized perturbation on the individual weight. The function of randomized perturbation is added into IMCS to construct diverse, yet high-quality feasible, partitions of (disjoint) sibling groups. A number of diverse partitions of sibling groups are accumulated over a number of iterations in the construction phase, where a parameter $max\_t$ is predetermined for limiting the maximum number of iterations. Subsequently, we perform cluster selection by solving a SCP to find the minimum set of sibling groups, which will be an initial solution for the next phase. In the enhancement phase, we propose a new local search with a memory function in two scales, cluster-based and individual-based neighborhoods, to improve the solution quality with respect to the pairwise similarity degree. In order to explore more high-quality solutions, we implement the RGOA procedure repeatedly to obtained a number of (high-quality) elite sets of sibling groups, where a parameter $max\_r$ is predetermined for limiting the maximum number of replications. Finally, among all (elite) solutions, the cluster selection is again performed by solving a SCP to obtain the final minimum set of sibling groups.

### *4.1. Construction Phase: Finding good and feasible sibling groups*

The goal of the construction phase is to construct high-quality partitions of feasible sibling groups, each maximizing the total weight of individuals assigned to it. In this paper, the greedy IMCS approach is employed and generalized by adding a new randomized weight perturbation to it. The idea behind the IMCS procedure is to iteratively construct a sibling group that

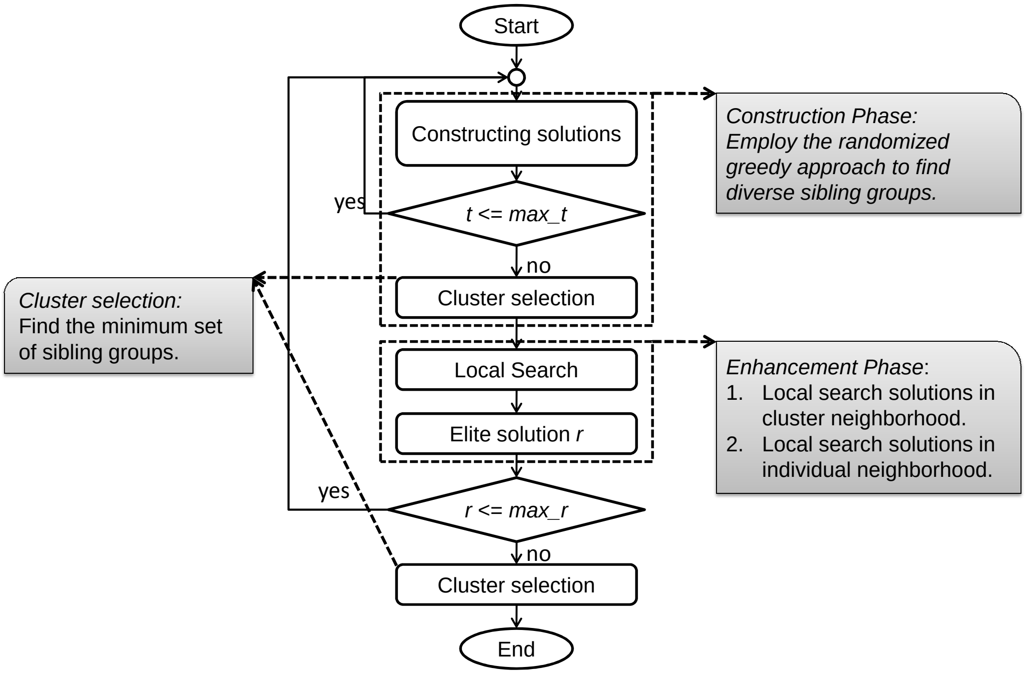
Start
Constructing solutions
*t <= max_t*
yes
no
Cluster selection
Local Search
Elite solution *r*
*r <= max_r*
yes
no
Cluster selection
End
*Construction Phase:*
*Employ the randomized greedy approach to find diverse sibling groups.*
*Cluster selection:*
Find the minimum set of sibling groups.
*Enhancement Phase*:
1. Local search solutions in cluster neighborhood.
2. Local search solutions in individual neighborhood.

Figure 3: Flow diagram of randomized greedy optimization algorithm. *Construction phase* is to construct a set of sibling groups with the randomized perturbations. *Enhancement phase* is to employ the two-stage local search to improve the solution quality. *Cluster selection* is to solve a set covering problem (SCP) to obtain the minimum set of sibling groups. A solution is defined a set of sibling groups (clusters).

covers the maximum number of individuals until no individuals are left while each group is subject to the Mendelian capacity constraints. Please refer to Chaovalitwongse et al. (2010) for more details. Because the IMCS uses a greedy-based optimization model that has an combinatorial objective function, it is very likely that there exist alternate or multiple optimal solutions. In other words, there may be several different groups with the same number of individuals that can be assigned to the group. In order to obtain diverse solutions in the construction phase, a randomized weight perturbation scheme is introduced. The weight of individual $i$ is defined by $\pi_i$ and added to the objective function of the IMCS. The concept behind the randomized perturbation is motivated by the noise method proposed in Charon and Hudry (1993). Note that, without the loss of generality, one can say that the IMCS in Chaovalitwongse et al. (2010) uses $\pi_i = 1,\ \forall i \in I$. In our case, the weight

**Algorithm 1** Randomized greedy optimization algorithm

```
Input: a set of individuals with genetic data
Output: a minimum set of sibling groups

procedure RANDOMIZED_GREEDY_OPTIMIZATION_ALGORITHM(input)
    repeat
        initialization: solution ← apply IMCS
        repeat
            solution ← solve IMCSP
            solution ← Update(solution)                    ▷ accumulate solution
        until $t > max\_t$
        solution ← ClusterSelection(solution)              ▷ solve a SCP
        solution ← LocalSearch_Cluster(solution)
        solution ← LocalSearch_Individual(solution)
        solution ← Update(solution)                        ▷ accumulate solution
    until $r > max\_r$
    solution ← ClusterSelection(solution)                  ▷ solve a SCP
return output
end procedure
```

is perturbed by adding a noise with a uniform distribution $[1-\epsilon,\ 1+\epsilon]$, where $\epsilon$ is a small positive number. The perturbed IMCS (IMCSP) can then be formulated as follows. Define the following decision variables:

- $x_i \in \{0,\ 1\}$: indicates if individual $i$ is assigned to be a member of the current sibling group;
- $y_k^l \in \{0,\ 1,\ 2\}$: indicates if any members in the current sibling group has distinct ($y_k^l = 1$) or homozygous ($y_k^l = 2$) allele(s) $k$ at locus $l$;
- $v_{kk'}^l \in \{0\ 1\}$: indicates if allele $k$ appears with allele $k'$ in the current sibling group at locus $l$.

The optimization model of IMCSP is given by

$$\text{(IMCSP)} \quad \max \quad \sum_{i \in I} \pi_i x_i \tag{25}$$

$$\text{s.t.} \quad a_{ik}^l x_i \leq y_k^l \quad \forall\, i \in I,\ k \in K,\ l \in L \tag{26}$$

$$\sum_{k \in K} y_k^l \leq 4 \quad \forall\, l \in L \tag{27}$$

$$\sum_{i \in I} a_{ik}^l a_{ik'}^l x_i \leq M v_{kk'}^l \quad \forall\, k \in K,\ k' \in K \backslash k,\ l \in L \tag{28}$$

$$\sum_{k' \in K \setminus k} v_{kk'}^{l} \leq 2 \quad \forall\, k \in K,\ l \in L. \tag{29}$$

The objective in Equation (25) is to maximize the total weight of individuals selected to be in the sibling group. The constraint sets in Equations (26)-(27) are derived from the first rule of the Mendel's laws, which is to ensure that the sum of the total number of distinct alleles and the number of homozygous alleles is less than or equal to four. The constraint sets in Equations (28)-(29) are derived from the second rule of the Mendel's laws, which is to ensure that each and every allele does not appear with more than two other alleles, except itself, in each locus. The procedure of IMCSP approach is shown in Algorithm 2, which is to solve the IMCSP model iteratively.

**Algorithm 2** IMCSP

```
Input: a set of individuals with genetic data
Output: a partition of sibling groups

procedure IMCSP(input)
    initialization: generate a perturbation randomly
    repeat
        solution $\leftarrow$ solve IMCSP(input)
        solution $\leftarrow$ Update(solution)                    $\triangleright$ accumulate solution
        remove selected individuals from the input set
    until no individual is assigned
return output
end procedure
```

In addition to the randomized weight perturbation scheme, we introduce a cut constraint to explore and further diversify alternate optimal solutions of IMCSP. This situation discussed in Chaovalitwongse et al. (2010). The cut constraint is defined by

$$\sum_{i \in \bar{I}} x_i \leq |\bar{I}| - 1, \tag{30}$$

where $\bar{I} \subset I$ contains only the individuals assigned in the current group. The implementation of this cut constraint is described as follows. We first solve the original IMCSP model, add the cut constraint to the IMCSP to remove the current optimal solution from the feasible space, and then resolve the IMCSP model with the cut constraint to obtain an alternate optimal solution. By using this cut constraint, we propose two variants other than the original IMCSP:

1. IMCSP_1: add the cut constraint to the original IMCSP in the first and second iterations;
2. IMCSP_2: add the cut constraint to the original IMCSP repeatedly in the first iteration.

### *4.1.1. Cluster Selection: Minimum Set Covering Problem*

Cluster selection is the last step of the construction phase. The goal of cluster selection is to select the best subset of sibling groups from a pool of high-quality solution candidates generated by the iterative IMCSP. It can also be used to remove redundant or dominated groups from the solution pool. Cluster selection can thus be mathematically formulated as a SCP. Define a binary assignment matrix $d_{ij}$, which presents that individual $i \in I$ is assigned to sibling group $j \in S$, where $S$ is a pool of all sibling group candidates. The SCP is given by min $\sum_{j \in S} z_j$; s.t. $\sum_{i \in I} d_{ij} z_j \geq 1, \forall j \in S$. The objective of SCP is to find the minimum set of sibling groups. The constraint set ensures that each individual must be covered by at least one of sibling group candidates. Note that this SCP is relatively small, and it can be solved efficiently by any MIP solvers.

### *4.2. Enhancement Phase: Improving the solution quality*

The goal in the enhancement phase is to improve the solution quality with respect to the pairwise similarity degree of individuals assigned to the same sibling groups by performing local search. Generally, a local search starts with an initial solution, explores alternative solutions in the neighborhood, makes a move to a better solution, and terminates when no better solution is found. In our case, the initial solution is given as a set of sibling groups $j \in J$ selected in the construction phase. The associated feasible space is defined as all constructed sibling groups $j \in S$. The effectiveness of local search thus relies on its evaluation function, initial solution, neighborhood definition, and search strategies. The evaluation function, which we want to maximize, is herein defined by the pairwise similarity degree of individuals assigned to the same sibling groups, which is the second term in Equation (9),

$$\sum_{j \in J} \sum_{(i,\, i') \in E} (\sum_{l \in L} w^l_{ii'}) x_{ij} x_{ij}. \tag{31}$$

To improve the efficiency of search procedure, we employ a memory function, which is motivated by the tabu search (Glover, 1989, 1990). The memory function is used to collect the past movements, which are associated to solutions, and to guide the search path in an improving direction. In the search path, the most recently visited solution enters the memory, and the oldest one is removed from the memory. Each solution in the memory must be visited until it is removed from the list. This is mainly to prevent a local cyclic search where there are many similar solutions to explore. In addition, the memory length is one of keys to affect the search efficiency. Longer memory length may guide the search path in the wrong direction, while shorter memory length may not have any effect. However, there is not a standard setting for the memory length, which really depends on the problem complexity.

We herein propose a two-stage local search in cluster-based and individual-based neighborhoods. In the cluster-based search, a cluster switch is performed when a sibling group with a higher pairwise similarity is randomly selected from other solutions to replace a sibling group with a lower similarity in the current solution. To record the cluster movement, we define the memory structure as $(j_1, j_2, ..., j_n)$, where $j$ is the label of sibling group visited and $n$ is the memory length. Subsequently after the cluster-based search, local search in the individual-based neighborhood is performed. An individual shift is performed when an individual is randomly selected from one sibling group and shifted to another sibling group, also selected randomly. Similar to the cluster-based search, the memory structure is defined as $([j_1, i_1], [j_2, i_2], ..., [j_m, i_m])$, where $j$ and $i$ are the labels of sibling group and individual visited, and $m$ is the memory length. After some moves, the solution may no longer be feasible because the new individual added to the sibling group may violate the Mendelian capacity constraints. In such a case, this movement is forbidden and a new neighbor (solution) is reselected. Thus, it is necessary to check if the current movement is forbidden in every iteration. Note that, by definition of individual-based neighborhood, the feasible space is reduced from $S$ to $J$ and fixes on only sibling groups $j \in J$ determined from the first stage. The local search is performed iteratively. The stopping criteria are the maximum number of search iterations for both stages and the maximum number of no-improvement consecutive iterations. The local search terminates when whichever stopping criterion is reached first.

### *4.3. Final Cluster Selection*

The final step of RGOA is to perform the final cluster selection to find the minimum set of sibling groups from a number of elite sets. This step is similar to the last step of the construction phase.

## 5. Computational Experiments

### *5.1. Performances on Real Data sets*

### *5.1.1. Characteristics of Real Data sets*

In this study, we show the performance of our proposed algorithm on real biological data sets. These real data sets are considered benchmark data that have widely used in the literature because the true sibling relationships (ground truth) are known. The characteristics of the data sets are summarized in Table 1. In all data sets, except the salmon data set, there are some missing values in genotypic data. The percentage of missing genotypic data in each data set is reported in the last column of Table 1. Based on preliminary analysis of allele frequency, there are violations of the Mendel's laws in the salmon and turtle data sets. This might be due to genotyping errors. The background and more detailed information of these data sets are described below.

Table 1: Characteristics of the biological data sets

| | # of | # of | # of | # of | Missing |
|---|---|---|---|---|---|
| Species | individuals | sibling groups | loci | alleles per locus | 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 |

**Salmon:** The Atlantic salmon *Salmo salar* data set 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 data set is a subset of one of the samples of genotyped individuals used by Almudevar and Field (1999).

**Shrimp:** The tiger shrimp *Penaeus monodon* data set (Jerry et al., 2006) consists of 59 individuals from 13 families with 7 loci. There are 8 pairs of missing alleles.

**Fly:** The *Scaptodrosophila hibisci* data set (Wilson et al., 2002) consists of 190 individuals in the same generation from 6 families sampled at various numbers of loci with up to 7 alleles per locus. All individuals shared 2 sampled loci which were chosen for our study. A total of 37.89% of the alleles were missing in this data set.

**Ant:** The *Leptothorax acervorum* data set (Hammond et al., 2001) are haplodiploid species. The data set consists of 377 worker diploid ants. This data set is a subset of one of the samples used by Wang (2004). There are 9% missing alleles in the data set.

**Turtle:** Kemp's ridleys sea turtle data set, *Lepidochelys Kempi*, is polyandrous and sampled from 26 mothers and offspring groups at 3 loci (Kickler et al., 1999). There are 16.38% missing alleles in the data set.

### *5.1.2. Computational Settings*

In this study, all computational experiments were programmed in MATLAB, and all MIP models were solved using a callable GAMS library with CPLEX version 10.0 (default setting). All experiments were run on an Intel Xeon Quad Core 3.0GHz processor workstation with 8 GB RAM memory. Execution time reported in this section were obtained from the desktop's internal timing calculations, which include time used for preprocessing and postprocessing.

The parameter settings of algorithm implementation are as follows. Each test data instance was implemented in a 20-hour computing time limit. The maximum number of RGOA replications was set to $max\_r = 100$. The maximum number of construction iterations was set to $max\_t = 50$ in the construction phase, where three variants of IMCSP, IMCSP_1, and IMCSP_2 were applied. In the enhancement phase, the major stopping criterion, the maximum number of search iterations, for two stages of local search were given by $50 \times |J|$ and $50 \times |I|$, respectively, and the auxiliary stopping criterion, the maximum number of no-improvement consecutive iterations, was set to 20, where $|J|$ is the cardinality of cluster set and $|I|$ is the cardinality of individual set.

### *5.1.3. Solution Assessment*

Although the overall similarity degree of reconstructed sibling groups is as an objective function to maximize in our approach, the ultimate objective of SRP is the accuracy of the reconstructed solutions. Specifically, the ground truth of sibling relationships of all test data sets is known. In sibling reconstruction research, reconstruction accuracy can be measured by calculating the percentage of individuals correctly assigned to resulting sibling groups in comparison to the actual sibling groups. The reconstruction accuracy can be calculated by quantifying an error measurement from the minimum partition distance (Gusfield, 2002). Specifically, the partition distance is equivalent to the minimum number of individuals that are removed from the reconstructed sibling groups so that they are identical to the actual sibling groups. The distance can be calculated by formulating and solving a maximum bipartite weighted matching problem. For more details, please refer to Chaovalitwongse et al. (2010). The ratio of the minimum distance to the total number of individuals provides a percentage of reconstruction errors. The reconstruction accuracy is equal to $(1 - error\ rate)$.

### *5.1.4. Reconstruction Results*

As mentioned in the previous section, there are three variants of our approach in the reconstruction phase: IMCSP, IMCSP_1, and IMCSP_2, and there are two stages in the enhancement phase: *cluster-based* and *individual-based*. The average and standard deviation of the reconstruction accuracies of all three variants after each phase of the framework are reported in Table 2. It can be seen that there are not significant differences among the three variants. Overall the accuracies gradually increase from the construction phase to the enhancement phase with the exceptions of the salmon and shrimp data sets. However, for the ant data set, the local search achieved a 100% reconstruction accuracy.

Table 3 presents the best final results of reconstruction accuracies and the numbers of sibling groups from the last step of elite cluster selection. It is observed that the proposed RGOA achieved 100% reconstruction accuracy on the shrimp and ant data sets. It is interesting to note that in other data sets that RGOA did not achieve 100% accuracy either there are missing allele information (fly and turtle) or violations in the Mendel's laws (salmon and turtle). For these reasons, RGOA did not provide accurate reconstruction results on those data sets. Nevertheless, even if the true optimal solutions were obtained, the reconstruction accuracies would be poor as well. The real reason is that the objective of our optimization framework and the Mendelian constraints assume that the data are not erroneous. In fact, most genetic data are erroneous. Thus a more robust optimization framework should be further investigated. From the table, it is also observed that IMCSP_1 and IMCSP_2 with the cut constraint are more time-consuming. From the last column in Table 3, for the same amount of time limit the numbers of replications of IMCSP_1 and IMCSP_2 are obviously smaller than IMCSP because each iteration of IMCSP takes much less time than that of IMCSP_1

Table 2: Reconstruction accuracies (%) in terms of *mean* ± *standard deviation* of the reconstruction results from different phases of RGOA tested on all data sets.

| | Constructive | Phase 1 | Phase 2 | Phase 2 |
|---|---|---|---|---|
| Species | strategy | | *cluster-based* | *individual-based* |
| Salmon | IMCSP | 98.29 ± 0 | 98.29 ± 0 | 98.29 ± 0 |
| | IMCSP_1 | 98.01 ± 0 | 98.01 ± 0 | 98.29 ± 0 |
| | IMCSP_2 | 98.29 ± 0 | 98.29 ± 0 | 98.29 ± 0 |
| Shrimp | IMCSP | 98.73 ± 2.54 | 98.73 ± 2.54 | 98.73 ± 2.54 |
| | IMCSP_1 | 98.73 ± 2.54 | 98.73 ± 2.54 | 98.73 ± 2.54 |
| | IMCSP_2 | 94.92 ± 0 | 94.92 ± 0 | 94.92 ± 0 |
| Fly | IMCSP | 52.82 ± 4.14 | 56.79 ± 4.78 | 59.59 ± 5.07 |
| | IMCSP_1 | 54.74 ± 5.86 | 56.56 ± 4.83 | 58.02 ± 5.25 |
| | IMCSP_2 | 53.16 ± 3.79 | 56.05 ± 3.90 | 58.36 ± 3.83 |
| Ant | IMCSP | 98.81 ± 0.94 | 99.60 ± 0.18 | 100 ± 0 |
| | IMCSP_1 | 98.81 ± 0.56 | 99.47 ± 0 | 100 ± 0 |
| | IMCSP_2 | 98.67 ± 0 | 99.47 ± 0 | 100 ± 0 |
| Turtle | IMCSP | 47.54 ± 1.87 | 48.57 ± 2.22 | 49.03 ± 2.05 |
| | IMCSP_1 | 46.50 ± 2.26 | 48.00 ± 2.07 | 48.43 ± 2.11 |
| | IMCSP_2 | 46.29 ± 6.47 | 46.29 ± 6.47 | 46.86 ± 5.66 |

and IMCSP_2. From our computational experience, we conclude that the IMCSP variant without the cut constraint should be used in order to save the computing time, yet maintain a good solution quality. On the other hand, the introduction of randomized perturbation can be helpful in terms of the diversification in the case where practitioners want to explore alternate solutions.

### *5.1.5. Comparison with Other Existing Methods*

To illustrate that our approach is among the best sibling reconstruction methods developed thus far, we compare the solution quality of RGOA and that of other state-of-the-art methods in the literature. The methods in the literature reported here include 2AOM, IMCS, A&F, B&M, KINGROUP, and COLONY. The IMCS approach solves a full optimization model 2AOM with 2-allele constraints to generate a partition of maximal sibling groups with 2-allele constraints while the statistical likelihood measure is not incorporated (Chaovalitwongse et al., 2010). The A&F algorithm is based on a completely combinatorial approach to exhaustively enumerate all possible sibling groups satisfying the 2-allele constraints and obtain a maximal, not necessarily optimal, collection of sibling groups (Almudevar and Field, 1999). The B&M algorithm is based on a mixture of likelihood and combinatorial

Table 3: Final results of the number of sibling groups, accuracy (%) and the number of replications. The computing time is limited within 20 hours (72,000 seconds). The perfect reconstruction are underlined.

| | | | Final Results | | | |
|---|---|---|---|---|---|---|
| Species | Constructing strategy | Actual # of sibling groups | # of sibling groups | Accuracy (%) | # of replications | Time (sec.) |
| Salmon | IMSCP | 7 | 7 | 98.29 | 6 | > 72,000 |
| | IMSCP_1 | 7 | 7 | 98.29 | 2 | > 72,000 |
| | IMSCP_2 | 7 | 7 | 98.29 | 1 | > 72,000 |
| Shrimp | IMSCP | 13 | 13 | 100.00 | 4 | > 72000 |
| | IMSCP_1 | 13 | 13 | 94.92 | 4 | > 72,000 |
| | IMSCP_2 | 13 | 13 | 94.92 | 1 | > 72,000 |
| Fly | IMSCP | 6 | 7 | 58.95 | 22 | > 72,000 |
| | IMSCP_1 | 6 | 7 | 65.79 | 22 | > 72,000 |
| | IMSCP_2 | 6 | 7 | 63.16 | 7 | > 72,000 |
| Ant | IMSCP | 10 | 10 | 100.00 | 2 | > 72,000 |
| | IMSCP_1 | 10 | 10 | 100.00 | 2 | > 72,000 |
| | IMSCP_2 | 10 | 10 | 100.00 | 1 | > 72,000 |
| Turtle | IMSCP | 26 | 18 | 56.57 | 10 | > 72,000 |
| | IMSCP_1 | 26 | 17 | 51.43 | 9 | > 72,000 |
| | IMSCP_2 | 26 | 18 | 42.86 | 2 | > 72,000 |

techniques used to construct a graph with individuals as the nodes and the edges weighted by the pairwise likelihood (relatedness) ratio. The algorithm identifies potential sibling groups by finding the connected components in the graph (Beyer and May, 2003). The KINGROUP algorithm 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 (Konovalov et al., 2004). The COLONY approach uses the maximum likelihood method to assign sibship and parentage jointly (Wang, 2004).

Table 4: Comparison results in accuracy (%) with other state-of-the-art approaches on five different species. The best results are underlined.

| Species | RGOA [a] | IMCS | 2AOM | A&F | B&M | KG | COLONY |
|---|---|---|---|---|---|---|---|
| Salmon | 98.29 | 98.29 | 94.02 | –[b] | 98.29 | 94.60 | 56.70 |
| Shrimp | 100.00 | 100.00 | 96.61 | 67.80 | 100.00 | 77.97 | 100.00 |
| Fly | 63.16 | 47.37 | 66.84 | 31.05 | 19.62 | 54.73 | –[c] |
| Ant | 100.00 | 93.10 | – [d] | –[b] | 97.61 | 97.10 | 100.00 |
| Turtle | 56.57 | 40.00 | –[d] | –[b] | 38.18 | 39.40 | 40.00 |

[a] We report the best accuracy among all experiments.

[b] A&F ran out of 4GB memory as it enumerates all possible sibling groups.

[c] There are no results available.

[d] No feasible solutions are obtained within 20 hours time limit.

Reconstruction accuracies of the above-mentioned reconstruction methods and RGOA on all biological data sets are shown in Table 4. Note that the best reconstruction results of RGOA among different parameter settings are reported. The most accurate reconstruction results are underlined. In all cases, RGOA obtained the best reconstruction results and outperformed all other methods. It is worth noting that although the RGOA's construction phase is based on the IMCS approach, randomized perturbation and local search can greatly improve the reconstruction accuracies. Specifically for the fly and turtle data sets, in which there are a lot of missing values, RGOA was able to increase the accuracies by about 15%. Both B&M and KINGROUP appear to be inaccurate on the data sets with a lot of missing values. We were not able to obtain the reconstruction results from the A&F algorithm on the salmon, ant, and turtle data sets because it ran out of memory when enumerating all possible combinations.

In Figure 4, we show the reconstruction accuracies of RGOA with the constructing strategy IMCSP on two real data sets (ant and turtle) over the time shift, which are compared to 2AOM and IMCS. Accuracies of RGOA are averaged by the number of replications at the time of 4, 8, 12, and 16 hours, and accuracies at 20 hours are obtained by final cluster selection. RGOA approach can achieve as good as, even better than accuracies IMCS approach although it takes longer computing time to obtained solutions. Moreover, it guarantees to have more diverse solutions so that we obtain better reconstruction accuracies on these two data sets. On the other hand, compared to 2AOM, we can always obtain good feasible solutions in a relatively short time ($<$ 20 hours) for large and complex data sets.

### *5.2. Performances on Simulated Data sets*

To show the ability of the proposed RGOA approach for larger complex data sets, we apply a random population generator (Chaovalitwongse et al., 2010) to generate larger simulated data sets when the real data sets at hand are relatively small-size, even the largest available in the literature. Essentially, the mechanism of the random population generator is to first construct a group of parents with the full genetic information such that a single generation of true sibling groups is known a priori. The generation process is as follows with parameters required: $M/F$ is the number of male/female adults, $l$ is the number of sampled loci, $a$ is the number alleles per locus, $j$ is the number of juveniles in the population per one adult female, and $o$ is the number of maximum number of offsprings per parent couple.

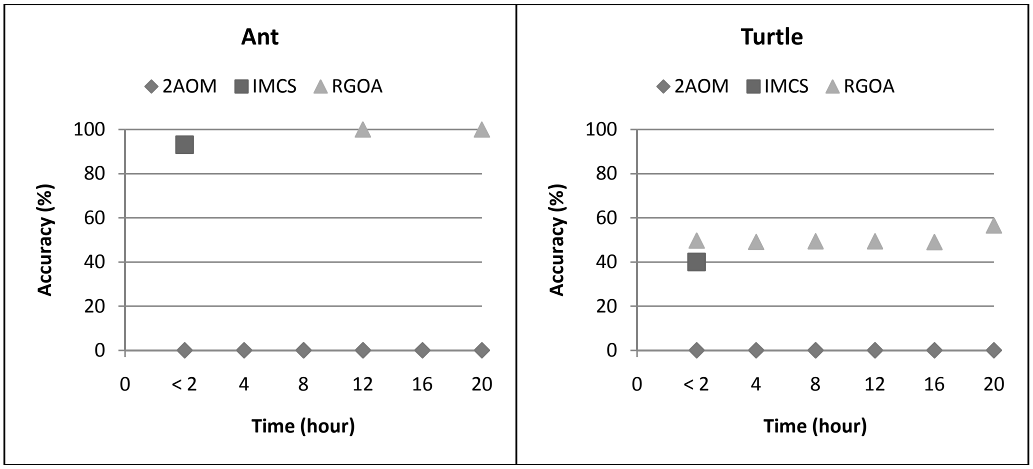

Figure 4: Averaged accuracies of RGOA on real data sets (ant and turtle) are obtained over time shift, compared to 2AOM and IMCS approaches in Chaovalitwongse et al. (2010). Accuracy = 0 represents that no feasible solution is available by 2AOM at the time. For IMCS, all solutions are obtained within two hours.

*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.

The parameter settings for larger simulated data sets are given: $M$ and $F = 30$, $j = 10$, $o = 40$ and 50, $l = 2, 3$, and 4, and $a = 10$. Additional computational settings are considered as follows. As mentioned previously, we suggest to adopt the constructive strategy IMCSP in the construction phase to save the computing time. For diversification reason, we expect to have more replications of ROGA) within a fixed computing time by shortening the construction phase. We add a stopping criterion of the maximum number of no-improvement consecutive iterations based on the similarity score in the construction phase and slightly reduce the maximum number of construction iterations to 20. Thus, the construction procedure terminates when whichever stopping criterion is reached first. The results are reported in Table 5, in turn, the number of sibling groups, accuracy (%) and the number of replications within 20 hours time limit, and compared to the known sibling relationships. We can still obtain good results in terms of the number of sibling groups and reconstruction accuracy. More accurate reconstruction is obtained when there are more genetic information (i.e., more loci). However, more loci make the problem more complex to solve, which can be seen that the number of replications of RGOA decreases with the complexity of problems because it is more time-consuming to solve for a single solution in the construction phase. Moreover, we compare the performance of the RGOA to IMCS and 2AOM approaches in Chaovalitwongse et al. (2010). The accuracies are reported in Table 6. With proposed randomized perturbation and local search, we obtain better reconstruction accuracies than IMCS. 2AOM can not be solved to obtain the solutions within 20 hours time limit. It is shown that our proposed approach is capable of solving larger complex problems effectively.

Table 5: Results of RGOA approach tested on larger simulated data sets. Final results are reported, in turn, the number of sibling groups, accuracy (%) and the number of replications within 20 hours (72,000 seconds) time limit, and compared to the known sibling relationships. The perfect reconstruction are underlined.

| | | Final Results | | | |
|---|---|---|---|---|---|
| Simulated data set | Actual # of sibling groups | # of sibling groups | Accuracy (%) | # of replications | Time (sec.) |
| Rand-j10-o40-l2-a10 | 10 | 10 | 91.00 | 24 | > 72,000 |
| Rand-j10-o50-l2-a10 | 10 | 10 | 91.60 | 13 | > 72,000 |
| Rand-j10-o40-l3-a10 | 10 | $\underline{10}$ | $\underline{100.00}$ | 7 | > 72,000 |
| Rand-j10-o50-l3-a10 | 10 | 10 | 99.80 | 7 | > 72,000 |
| Rand-j10-o40-l4-a10 | 10 | $\underline{10}$ | $\underline{100.00}$ | 3 | > 72,000 |
| Rand-j10-o50-l4-a10 | 10 | $\underline{10}$ | $\underline{100.00}$ | 3 | > 72,000 |

Table 6: Accuracy results of RGOA approach compared to IMCS and 2AOM approaches (Chaovalitwongse et al., 2010) from the simulated data sets.

| Simulated data set | RGOA | IMCS | 2AOM[a] |
|---|---|---|---|
| Rand-j10-o40-l2-a10 | 91.00 | 89.00 | - |
| Rand-j10-o50-l2-a10 | 91.60 | 79.40 | - |
| Rand-j10-o40-l3-a10 | 100.00 | 98.25 | - |
| Rand-j10-o50-l3-a10 | 99.80 | 96.80 | - |
| Rand-j10-o40-l4-a10 | 100.00 | 99.25 | - |
| Rand-j10-o50-l4-a10 | 100.00 | 100.00 | - |

[a] No feasible solution is obtained within 20 hours time limit.

## 6. Conclusion and Discussion

In this study, we modeled the SRP as a special case of the CCP and developed a new efficient solution approach, called RGOA, algorithm to solve the SRP. The algorithm employs both combinatorics and statistical likelihood concepts to analyze the microsatellite genetic data. To our knowledge, the proposed algorithm is among the first approach to put both statistical and combinatorial concepts together into a single optimization model. The statistical likelihood measure is used as an objective function while the combinatorial concept of the Mendel's laws is used as a capacity constraint. To practically solve this hard, large-scale combinatorial optimization, RGOA was designed based on the core concept of GRASP. Nevertheless, the difference between the proposed RGOA and the GRASP lies on the incorporation of randomness and greediness in the construction phase. In addition, the solution structures of RGOA and GRASP are also different. GRASP uses an explicit evaluation function to assess the quality of solutions directly. For our problem in particular, unlike GRASP, the solution quality is defined by the reconstruction accuracy, which cannot be measured until the end of the final step of RGOA because the actual sibling groups should not be known in practice. The solution quality in RGOA can be implicitly evaluated by similarity scores. We thus employed the cluster selection to find the (possibly best) combination of sibling groups from all possible solutions at the end of construction phase.

RGOA was tested on real biological data sets and larger simulated data sets, and compared with other state-of-the-art SRP methods. The experimental results suggest that RGOA is the most accurate method among all tested methods. In addition, for the data sets with a lot of missing data, RGOA appears to be more robust than other methods. Although RGOA is heuristic-based and cannot guarantee the optimality, the best solution should be the most accurate solution, not the one that maximizes the statistical likelihood measure. More importantly, RGOA is able to provide multiple high-quality solutions so that practitioners can investigate a set of good solutions as opposed to the optimal solution. Note that RGOA is quite demanding in terms of computational time. The stopping criteria are rather too specific on the computational time. In the future, a systematic framework based on the column generation may be considered. Moreover, in practice, the full sibling reconstruction is limited to monogamous species. New framework for the study of half sibling reconstruction (Sheikh et al., 2010) should be studied. Other biological objectives such as minimizing the parent pairs and inferring higher ordered sibling relationships are worthy to explore in the future.

## Acknowledgments

This research is supported by the following grants: NSF IIS-0611998 (Chaovalitwongse), NSF CCF-0546574 (Chaovalitwongse), NSF IIS-0612044 (Berger-Wolf, Ashley, DasGupta). We first thank Saad Sheikh and Isabel C. Caballero for knowledge and corrections. 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.

## References

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 (2), 213–229.

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

Avella, P., Boccia, M., Sforza, A., Vasil'ev, I., 2009. An effective heuristic for large-scale capacitated facility location problems. Journal of Heuristics 15, 597–615.

Baldacci, R., Hadjiconstantinou, E., Maniezzo, V., Mingozzi, A., 2002. A new method for solving capacitated location problems based on a set partitioning approach. Computers & Operations Research 29, 365–386.

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

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

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

Blouin, M., 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. TRENDS in Ecology and 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.

Butler, K., Field, C., Herbinger, C., Smith, B., 2004. Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from dna marker data. Molecular Ecology 13, 1589–1600.

Ceselli, A., Liberatore, F., Righini, G., 2009. A computational evaluation of a general branch-and-price framework for capacitated network location problems. Annals of Operations Research 167, 209–251.

Chaovalitwongse, W., Berger-Wolf, T. Y., DasGupta, B., Ashley, M. V., 2007. A robust combinatorial approach for sibling relationships reconstruction. Optimization Methods and Software 22 (1), 11–24.

Chaovalitwongse, W., Chou, C.-A., Berger-Wolf, T. Y., DasGupta, B., Sheikh, S., Putrevu, S. L., Ashley, M. V., Caballero, I. C., 2010. New optimization model and algorithm for sibling reconstruction from genetic markers. INFORMS Journal on Computing 22 (2), 180–194.

Charon, I., Hudry, O., 1993. The noise method: a new method for combinatorial optimization. Operations Research Letters 14, 133–137.

Chaves, A. A., Lorena, L. A., 2010. Clustering search algorithm for the capacitated centred clustering problem. Computers & Operations Research 37, 552–558.

Deng, Y., Bard, J., 2010. A reactive grasp with path relinking for capacitated clustering. Journal of Heuristics, 1–34.

Feo, T., Resende, M., 1995. Greedy randomized adaptive search procedures. Journal of Global Optimization 6, 109–133.

Franca, P. M., Sosa, N. M., Pureza, V., 1999. Adaptive tabu search approach for solving the capacitated clustering problem. International Transactions of Operations Research 6, 665–678.

Glover, F., 1989. Tabu search - part I. ORSA, Journal on Computing 1, 190–206.

Glover, F., 1990. Tabu search - part II. ORSA, Journal on Computing 2, 4–32.

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

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

Hansen, P., Jaumard, B., 1997. Cluster analysis and mathematical programming. Mathematical Programming 79, 191–215.

Herbinger, C., O'Reilly, P., Doyle, R., Wright, J., O'Flynn, F., 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., Evans, B., Kenway, M., Wilson, K., 2006. Development of a microsatellite DNA parentage marker suite for black tiger shrimp penaeus monodon. Aquaculture 255, 542–547.

Kickler, K., T., M., Holder, Davis, S., M*á*rquez-M, R., Owens, D. W., 1999. Detection of multiple paternity in the kemp's ridley sea turtle with limited sampling. Molecular Ecology 8 (5), 819–830.

Konovalov, D. A., Bajema, N., Litow, B., 2005. Modified simpson $o(n^3)$ algorithm for the full sibship reconstruction problem. Bioinformatics 21 (20), 3912–3917.

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

Koskosidis, Y., Powell, W., 1992. Clustering algorithms for consolidation of customer orders into vehicle shipments. Transportation Research 26B, 365–379.

Lorena, L. A., Senne, E. L., 2004. A column generation approach to capacitated p-median problems. Computers & Operations Research 31, 863–876.

Maniezzo, V., Mingozzi, A., Baldacci, R., 1998. A bionomic approach to the capacitated p-median problem. Journal of Heuristics 4, 263–280.

Mehrotra, A., Trick, M. A., 1998. Cliques and clustering: A combinatorial approach. Operations Research Letters 22, 1–12.

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

Mulvey, J., Beck, M., 1984. Solving capacitated clustering problems. European Journal of Operations Research 18, 339–348.

Negreiros, M., Palhano, A., 2006. The capacitated centred clustering problem. Computers & Operations Research 33, 1639–1663.

Osman, I. H., Creistofides, N., 2002. Capacitated clustering problems by hybrid simulated annealing and tabu search. International Transactions of Operations Research 1, 317–336.

Osman, I. H., Samad, A., 2002. Guided construction search for the capacitated *p*-median problem. Working Paper, School of Business, American Univeristy of Beirut, Lebanon.

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

Queller, D. C., Goodnight, K. F., 1989. Estimating relatedness using genetic markers. Evalution 43 (2), 258–275.

Samad, A., Osman, I. H., 2002. Density based problem space search for the capacitated clustering problems. Annals of Operations Research 131, 21–43.

Samad, A., Osman, I. H., 2005. Greedy random adaptive memory programming search for the capacitated clustering problems. European Journal of Operations Research 162, 30–44.

Sheikh, S., Berger-Wolf, T. Y., Khokar, A., Chou, C.-A., Chaovalitwongse, W., Ashley, M. V., Caballero, I. C., DasGupta, B., 2010. Combinatorial reconstruction of half-sibling groups: Models and algorithms. Journal of Bioinformatics and Computational Biology 8 (2), 1–20.

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

Thomas, S. C., Hill, W. G., 2000. Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics 15, 1961–1972.

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

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

Wang, J., Santure, A. W., 2009. Parentage and sibship inference from multilocus genotype data under polygamy. to appear in Genetics.

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