An Integrated Optimization Framework for Inferring Two Generation Kinships and Parental Genotypes from Microsatellite Samples

Daehan Won
Chun-An Chou[*]
W. Art Chaovalitwongse [†]

Departments of Industrial & Systems Engineering and Radiology, University of Washington
Seattle,WA 98195
wondae@uw.edu
joechou@uw.edu
artchao@uw.edu

Tanya Y. Berger-Wolf
Bhaskar Dasgupta
Ashfaq A. Khokhar
Marco Maggioni
Department of Computer Science, University of Illinois at Chicago
Chicago, IL 60607
tanyabw@cs.uic.edu
dasgupta@bert.cs.uic.edu
ashfaq@cs.uic.edu
mmaggi3@uic.edu

Saad Sheikh
Department of Computer & Information Science & Engineering, University of Florida
Gainesville, FL 32611
ufl@saadshekh.com

Mary V. Ashley
Jason Palagi
Department of Biology, University of Illinois at Chicago
Chicago, IL 60607
ashley@uic.edu
jpalag2@uic.edu

ABSTRACT

With the growing development and application of genetic data availability, it provides new possibilities in establishing the genealogical relationships of individual organisms such as sibling reconstruction, parentage inference, and inheritance investigation. We propose a new integrated optimization framework for parental reconstruction of a single-generation population using microsatellite data. Without prior information about the population, our optimization framework uses the combinatorial concepts of Mendel's laws of inheritance to reconstruct sibling groups and in turn identifies the associated parental genotypes. The effectiveness and robustness of our proposed approach were evaluated by both real biological and simulated data sets, covering different mating systems: monogamy, semi-monogamy, and polygamy. Additionally, we compared the results of the proposed approach with other state-of-the-art sibship reconstruction and parentage inference methods. The results demonstrate efficient and accurate inference for parental genotypes, and potentially suggest that our framework can provide an insightful roadmap for investigators to navigate fundamental ecological and evolutionary studies.

Categories and Subject Descriptors

G.1.6 [**Mathematics of Computing**]: Optimization–Integer programming
; G.2.1 [**Mathematics of Computing**]: Discrete mathematics–Combinatorics
; J.3 [**Computer Applications**]: Life and Medical Science–Biology and genetics

General Terms

Algorithms, Performance

Keywords

Computational biology, microsatellites, population biology, optimization algorithm, parentage inference

1. INTRODUCTION

Emerging technologies of molecular markers have enabled biologists to investigate the genealogical relationships among individuals and pedigree in wild populations. However, inferring such genetic information (e.g. kinship and pedigree) is still extremely hard to do from observations alone [2]. Microsatellite biomarkers provide new possibilities to develop more advanced computational methodologies of defining pedigree relationships that they are very vital in various

[*] Corresponding Author

[†] Corresponding Author

ACM-BCB '12, October 7-10, 2012, Orlando, FL, USA

research fields, for example, mating systems, social behavior or organizations, and isolation by distance and spatial genetic structure in wild populations [9, 12, 19, 22]. As part of pedigree relationships investigation, sibling group identification provides inference of many meaningful and useful biological parameters, including the number of reproducing adults, their fecundity, and the average size of litters [6]. For studies of evolutionary genetics, kinship inference can be used for assessing the inheritance of adaptive attributes and how they will behave in natural selection [6]. While real life applications of genetic markers are showing progress, in parallel, many statistical approaches which are based on Mendel's law of inheritance have been studied to analyze genetic marker data [2, 15, 22].

However, the above mentioned statistical methods have shown limited availabilities for practical problems [4]. Most methods infer a single generation relationship [15], either parentage or full sibships [2] with overlooking any other relationships and uncovered information [22]. Moreover, those methods require some prior knowledge of typical allele distribution and frequency, population size, and other information about the species [2]. To compensate these shortcomings, combinatorial approaches for sibship inference that do not require any prior genetic information have been proposed [6, 20]. Even though these approaches have been able to provide more practical and applicable solutions, the solutions of these approaches are not sufficient to define pedigree information because of the lack of parentage inference. Besides, these approaches solve for full or half sibship relationships independently, and they cannot construct such general pedigree structure at the same time.

In this paper, we present an integrated optimization framework for the sibling reconstruction and parental inference using microsatellite data acquired from a single generation. Similar to previous studies by our group [5, 6, 20], we assume that there is no prior knowledge about the population. Specifically, our framework can be divided into two sequential stages: half sibling reconstruction and full sibling reconstruction subsequently. Our approach employs mathematical programming methods that have been developed previously by our group [6, 20]. After the reconstruction steps, the parentage information is inferred by using the derived information from the proposed sibling reconstruction steps. Through our framework, we can acquire two genealogical information such as sibship reconstruction and parental inference simultaneously. Consequently, we can derive the meaningful information of genealogical relationship such as sibship and pedigree, using the solutions obtained from our framework.

The organization of the paper is as follows. In Section 2, we describe the background of the two generation sibship problem and genetic data representation. In Section 3, we present the proposed framework. In Section 4, we demonstrate the computational experiments for both real and simulated data sets, and compare our approach with other existing methods. We conclude our work in Section 5.

2. BACKGROUND

Microsatellites are among the most commonly used genetic markers in sibship and population studies. Microsatellites are short tandem repeats of DNA sequences that vary in length. In the genome, microsatellites occur at specific locations (i.e., loci) on a chromosome. An allele is a distinct pattern of variable DNA sequences in microsatellites [6]. In this study, we present microsatellites data mathematically as a sequence of integer $\{0, 1, 2\}$ showing allele and locus information for each individual. To transform microsatellite data into such a mathematical representation, we define the following sets which will be used throughout the paper: I is a set of individuals, J is a set of reconstructed sibling groups, K is a set of alleles and L is a set of loci. A multi-dimensional data matrix is defined as $a^l_{ik} \in \{0, 1, 2\}$ of individual $i \in I$ at locus $l \in L$. This matrix represents the indication of heterozygous alleles ($a^l_{ik} = 1$) at a locus as well as homozygous alleles ($a^l_{ik} = 2$) and ($a^l_{ik} = 0$) represents allele k is not present.

In biology, Mendel's law of inheritance is the rule that describes the constraints of genetic 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 other loci [3, 18]. Based on the rule, the 2-allele constraints have been proposed as a sufficient condition for establishing sibling relationships [1, 5, 6]. The 2-allele constraints are defined: (1) the number of distinct alleles plus the number of homozygous alleles at each locus is less than four and (2) each allele cannot appear together with more than two other alleles at each locus. A full sibling group is defined as a group that all members have to satisfy the 2-allele constraints and have both parents in common. In addition, half sibling relationships have also been investigated. A half sibling group can be defined as a group that all members have one of parents in common [20]. Furthermore, full sibling groups belong to a half sibling relationship if one of shared parents appears to be the same across the full sibling groups. Both of half sibling groups and full sibling groups can be reconstructed by the mathematical optimization models developed in [6, 20].

3. METHOD

In this section, we present a new integrated optimization framework for the sibship reconstruction and parentage inference as described in the following steps.

First, we reconstruct half sibling groups such that each group member is a half sibling of each other. In a population, there exist a number of half sibling groups [20], and every individual is assigned to at least one half sibling group. Let H_j is the j-th half sibling group; $I = \bigcup_{\forall j \in J} H_j$, where J is the set of half sibling groups and I is the set of individuals in the population. After individuals in a half sibling group are assigned, the shared parent is identified and its associated genotypes are inferred. The detail is explained in Section 3.1.

Second, we reconstruct full sibling groups from the half-sibling solutions. In each half sibling group, we attempt to find subgroups whose members have both of parents in common, which in turn considered to be full sibling groups. Each individual can also be assigned to at least one full sibling group. We can represent a half sibling group $H_j = \bigcup_{\forall k \in K} F_k$, where F_k is k-th full sibling group belonging to half sibling group H_j. The procedure of the above two steps is repeated $|J|$ times for each half sibling group [6, 20].

Third, according to Mendel's law of inheritance, any individual who belongs to a full sibling group inherits its genotypes from parents. Since we determine individual assignments to single full sibling group and one of the parents from the previous two steps, we can infer the other one of par-

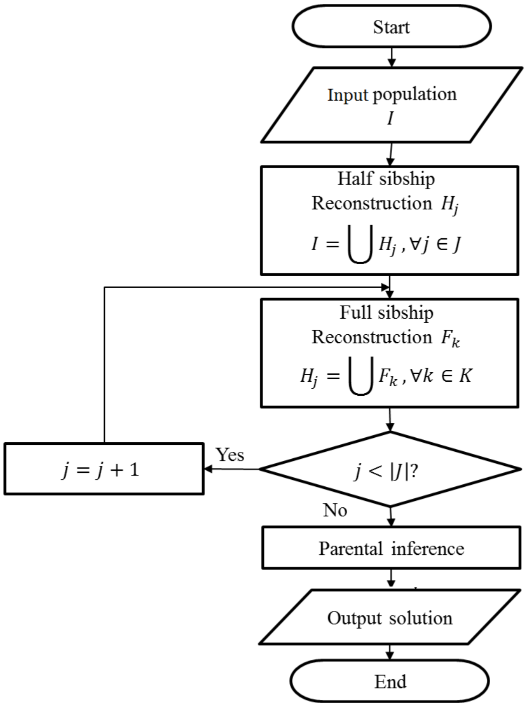

Figure 1: Flowchart of the proposed integrated framework for inferring two generation kinships and parental genotypes from microsatellite data.

ents using these predetermined information conversely (See Section 3.3).

3.1 Half Sibling Model

We describe the reconstruction of half sibling relationships. We define binary variables used in the mathematical model as follows: x_{ij} indicates if individual i is assigned to be a member of half sibling group j, z_j indicates if group j is selected or not, w^l_{jk} indicates if allele k appears in half sibling group j at locus l, and a^l_{ik} is the number of times allele k appears in individual i at locus l. The optimization model for the half sibling reconstruction is formulated as follows.

$$
\begin{aligned}
\min \quad & \sum_{j \in J} z_j && (1)\\
\text{s.t.} \quad & x_{ij} \le z_j \qquad \forall i \in I, \forall j \in J && (2)\\
& \sum_{j \in J} x_{ij} \ge 1 \qquad \forall i \in I && (3)\\
& \sum_{k \in K} a^l_{ik} w^l_{jk} \ge x_{ij} \quad \forall i \in I, \forall j \in J, \forall l \in L && (4)\\
& \sum_{k \in K} w^l_{jk} \le 2 \qquad \forall j \in J, l \in L. && (5)
\end{aligned}
$$

Equation (1) is the objective function to minimize the number of half sibling groups. Equations (2) and (3) present the logical constraints to ensure that any individual i is to assigned to at least one of half sibling groups in J. Equations (4) and (5) ensure that half sibling groups contain no more than two alleles at each locus.

If binary decision variable w^l_{jk} is activated (e.g., $w^l_{jk} = 1$), then it represents that every individual assigned to half sibling group j commonly shares allele k at locus l. In other words, every individual in half sibling group j has an allele k at locus l in common. Hence, we can conclude that all of these common alleles represent the genotypes of the shared parents of half sibling groups.

3.2 Full Sibling Model

In this section, we present full sibling reconstruction. Full sibling reconstruction is formulated around the 2-allele constraints [6].

We define binary decision variables z_j and x_{ij} which indicate similar representation in that of the half sibling model, where J represents the set of full sibling group instead. Integer variable y^l_{jk} indicates if any members in full sibling group j has heterozygous ($y^l_{jk} = 1$) or homozygous ($y^l_{jk} = 2$) allele k at locus l and binary variable $v^l_{jkk'}$ indicates if allele k appears in the current full sibling group j at locus l. The optimization model for the full sibling reconstruction is formulated as follows.

$$
\begin{aligned}
\min \quad & \sum_{j \in J} z_j && (6)\\
\text{s.t.} \quad & x_{ij} \le z_j \quad \forall i \in I, \forall j \in J && (7)\\
& \sum_{i \in I} x_{ij} \ge 1 \quad \forall j \in J && (8)\\
& \sum_{i \in I} a^l_{ik} x_{ij} \le y^l_{jk} \quad \forall k \in K, \forall j \in J, \forall l \in L && (9)\\
& \sum_{k \in K} y^l_{jk} \le 4 \quad \forall j \in J, \forall l \in L && (10)\\
& \sum_{i \in I} a^l_{ik} a^l_{jk'} x_{ij} \le M v^l_{jkk'} && (11)\\
& \forall k \in K, \forall k' \in K \setminus k, \forall j \in J, \forall l \in L\\
& \sum_{k' \in K \setminus k} v^l_{jkk'} \le 2 \quad \forall j \in J, \forall k \in K, \forall l \in L. && (12)
\end{aligned}
$$

Equation (6) is the objective function that minimizes the number of full sibling groups. Equations (7) and (8) describe the logical constraints similar to Equations (2) and (3). Equation (9) ensures that allele indication variable y^l_{jk} is activated if individual i is assigned to full sibling group j. Equation (10) constrains that the numbers of heterozygous alleles and homozygous alleles are less than or equal to 4. Equation (11) is the restriction of the binary variable $v^l_{jkk'}$, which must be activated for any assignment of individual i to full sibling group j. M is a positive large number, where $M = |I|+1$. Equation (12) ensures that every allele in full sibling group j cannot exist with more than two other alleles at each locus.

3.3 Parental Inference

The genotypes of the shared parents can be inferred by identifying a pair of alleles, for which every individual is inherited from at least one of the parents. The genotypes of the unknown parents are then determined by examining the combination of alleles from the known parents across loci. When the offspring population is known to contain half and full sibling relationships, we can construct the parental genotypes from the offspring genotypes [7, 8, 13].

From Mendel's law of inheritance [3, 18], if one of the parents is known, then the alleles of the other parent can be decided by subtracting the known parent's alleles from the offsprings. In other words, in order to infer the unknown parent, we need the genotypes of the known parent and their offsprings. Population with no known parents, the parentage reconstruction problem is not far more complicated than the populations with known parents [16].

According to our framework, a single half sibling group is in fact decomposed into several full sibling groups. From the definition of half sibling relationships, the decomposed group is referred as the group of full siblings that share common parents. Recall the half sibling model in Section 3.1, the binary decision variable w_{jk}^l is used to represent the genotypes of the shared parent of half sibling group j. Thus, one of the parents for each full sibling group can be consequently defined by w_{jk}^l. Using Mendel's law, the unknown parent for each full sibling group can be identified by the genotypes of individuals in the full sibling group and one of the parents from half sibling reconstruction stage.

Suppose that one shared parent of single full sibling group F_j is $p(F_j) = \{f^l, b^l\}$, where f^l represents front allele and b^l represents back allele at locus l, respectively. We can define that one shared parent $p(F_j)$ can be expressed by

$$p(F_j) = \{f^l, b^l\} = \{k \mid w_{jk}^l = 1\} \quad \forall j \in J, \forall l \in L.$$

Additionally, we present the i-th individual in full sibling group F_j as $x_i = \{f_i^l, b_i^l\} \quad \forall l \in L$. Similar to the above definition, f_i^l and b_i^l indicate the front and back allele of individual i at locus l. Unknown parent $\hat{p}(F_j)$ can be defined by

$$\begin{aligned} &\hat{p}(F_j) = \{\hat{f}^l, \hat{b}^l\} \\ &s.t. \quad \{\hat{f}^l, \hat{b}^l\} \cup \{f^l, b^l\} = \bigcup_{\forall i \in F_j} \{f_i^l, b_i^l\}, \quad \forall l \in L. \end{aligned}$$

4. EXPERIMENTAL RESULTS

In this section, we present the experimental results of the proposed approach for both real and simulated data sets. We implemented our approach in C# synchronized with the optimization solver ILOG CPLEX 12.2. All tests were run on Intel Xeon 2.33 GHz×8 processors workstation with 24 GB RAM memory.

4.1 Evaluation

To evaluate the effectiveness of the proposed framework, we employ group assignment accuracy and individual pairwise accuracy concurrently, which are addressed as follows.

4.1.1 Group Assignment Accuracy

The correctness of group assignment is obtained from a minimum partition distance, which is defined as the smallest number of individuals that need to be removed from the population to make two partitions equivalent [10, 20]. From our previous work [6], the minimum partition distance can be transformed into a maximum linear assignment problem (MLAP).

However, this evaluation may overestimate sibling assignment because these sibling groups are obtained by solving a set covering model. Since our mathematical models are based on a set-covering structure, using MLAP for calculating accuracy may not be a proper way to evaluate the effectiveness because it causes an overestimation. For instance, there are two original groups {1, 2, 3} and {4, 5, 6} (The number means an index of each individual). If reconstructed groups are {1, 2, 3} and {1, 2, 4, 5, 6}, then the calculated accuracy by MLAP would be 100%. In order to calculate accuracy precisely, we propose a modified model (MLAP-m) of MLAP. To prevent "overestimate" the reconstructed results, we propose a new way to quantify the accuracy. We combine "presence" with "absence" of individuals between actual and reconstructed groups. Note that "presence" is somewhat similar to, but not, the cost as input in the general mathematical model. We define a binary variable u_a^i indicating if individual i appears in actual group a correctly and a binary variable u_r^i indicating if individual i appears in reconstructed group correctly. For each pair of actual and reconstructed groups, we can identify "match" and "mismatch" as follows. $c_{ar} = \sum_{i \in I} c_{ar}^i$ for $u_a^i = u_r^i = 1$ and $d_{ar} = \sum_{i \in I} d_{ar}^i$ for $u_a^i = u_r^i = 0$.

By taking summation over all individuals, we can obtain "presence" and "absence" information for all pairs. Here, c_{ar} and d_{ar} are sensitivity and specificity essentially. We then redefine the mathematical model with these two inputs as follows.

$$\begin{aligned} \max \quad & \alpha \sum_{a \in A} \sum_{r \in R} c_{ar} x_{ar} + \beta \sum_{a \in A} \sum_{r \in R} d_{ar} x_{ar} && (13) \\ \text{s.t.} \quad & \sum_{a \in A} x_{ar} \leq 1 \quad \forall\, r \in R && (14) \\ & \sum_{r \in R} x_{ar} \leq 1 \quad \forall\, a \in A && (15) \\ & x_{ar} \in \{0, 1\}. && (16) \end{aligned}$$

The objective in Equation (14) is to maximize the sum of sensitivity and specificity, while the constraints ensure that each reconstructed group is assigned to as most one actual group. α and β are the weights (between 0 and 1) of sensitivity and specificity in the objective function.

4.1.2 Individual Pairwise Accuracy

In this section, we introduce the other measurement for individual pairwise accuracy. The idea is to assess if a pair of individuals is in the same reconstructed groups or not compared to the actual groups. Let suppose sets A and R are binary, and their elements are denoted by a_{ij} and r_{ij}, respectively. These indicate that if individual i and j are in the same actual and reconstructed groups. $|A| = |R| = n$, where n is the total number of individuals. On the other hand, we can also describe "absence" information and introduce binary sets Ã and R̃ which satisfy following condition: $\tilde{a}_{ij} + a_{ij} = 1$ ($\tilde{a}_{ij} \in \tilde{A}$, $a_{ij} \in A$) and $\tilde{r}_{ij} + r_{ij} = 1$ ($\tilde{r}_{ij} \in \tilde{R}$, $r_{ij} \in R$).

Consequently, individual pairwise accuracy is calculated as follows. Sensitivity represents the "presence" information of individual pairs and specificity represents the "absence" information of individual pairs.

$$\begin{aligned} Sensitivity &= \frac{\sum_{\forall i,j \in I, i<j} a_{ij} r_{ij}}{\sum_{\forall i,j \in I, i<j} a_{ij}} \\ Specificity &= \frac{\sum_{\forall i,j \in I, i<j} \tilde{a}_{ij} \tilde{r}_{ij}}{\sum_{\forall i,j \in I, i<j} \tilde{a}_{ij}} \end{aligned}$$

4.1.3 Inferred Parental Information Accuracy

Additionally, we define another measurement to calculate an accuracy of inferred parental genotypes. We could acquire parental inference information through our framework. For calculating inferred parents accuracy, we com-

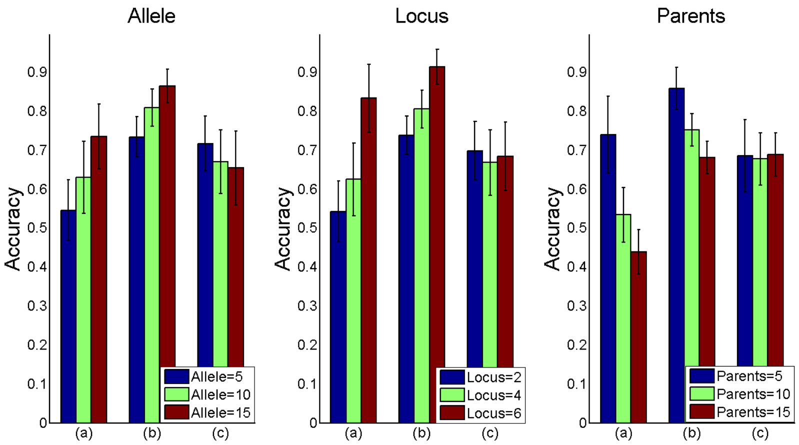

Figure 2: Group assignment accuracy(MLAP(a), MLAP-m(b)) and parentage inference accuracy (c) for simulated data sets

pare alleles between actual parents and inferred parents by a simple counting way: $\frac{2|L|-|U|}{2|L|}$, where $|L|$ and $|U|$ represent the number of sampled loci and the number of uncorrected match of alleles between actual and inferred parents, respectively.

4.2 Real Biological Data

We conducted experiments for both a real biological data set and simulated data sets. In this section, we present the experimental results using a real biological data set which is considered as benchmark data because its actual full sibling relationship was predetermined. The data set used in this paper was containing missing allele information because of technical errors in acquiring and scoring microsatellite data. In this study, any missing alleles or detected genotype error was replaced by a wild card "-1" to indicate the missing information. When our approach performs sibling reconstruction for each individual, a wild card could be transformed to any allele [6]. We used *Ant* data set and its specification is as follows.

Ant: The Leptothorax acervorum data set [11] is 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 [21]. There are 9% missing alleles in the data set.

The result for Ant data set is shown in Table 1. From the Table 1, our approach shows little short of perfect reconstruction. Especially, Higher values of group assignment and individual pairwise accuracies support that our framework provides robust reconstruction results with preventing overestimation.

Table 1: Reconstructed accuracies for the real data set.

	Group assignment Accur.		Individual pairwise Accur.	
	MLAP	MLAP-m	Sensitivity	Specificity
Ant	0.939	0.980	0.902	0.998

4.3 Simulated Data

To verify our framework's validity, we employed a random population generator to create various simulated data sets [6]. The random population generator requires the following parameters. M is the number of adults males, F is the number of adults females, l is the number of sampled loci, a is the number of alleles per locus, g is the number of groups in the population per one adult female, and o is the maximum number offsprings (individuals) per parent couple. The parameters are shown in Table 2. For each configuration, we generated 25 replications to verify our approach's stability. We conducted experiments for simulated data sets to demonstrate the effectiveness.

Table 2: Input parameters for simulated data sets.

		Input values		
Parameters	Symbol	Monogamy	Semi-Monogamy	Polygamy
Adult female	F	5,10,15	1	5,10,15
Adult male	M	5,10,15	5,10,15	5,10,15
Sampled loci	l	2,4,6	2,4,6	2,4,6
Allele per loci	a	5,10,15	5,10,15	5,10,15
Groups	g	5,10,15	5,10,15	5,10,15
Offsprings	o	5,10	5,10	5,10

Depending on the input data sets, our mathematical model may require huge running time to solve the problem. In order to conduct reasonable test, we set a stopping criterion which was set to be either 5 hours (18,000 seconds) running time or a predetermined solution quality. Test results contained two effectiveness measurements, "assignment accuracy" and "inferred parental information accuracy". The accuracies are reported as averaged accuracies of 25 replicates.

Figures 2 and 3 illustrate the performance trends for our approach when varying the number of alleles per loci, the number of sampled loci and the number of parents. In Figure 2, the accuracy increases as alleles and loci increases for our approach. In Figure 3, pairwise accuracy shows similar performance trend with group assignment accuracy. However, in Figures 2 and 3, we can also observe that the accuracy decreases as the number of parents increases. Increasing number of parents means the number of individuals in a population is increasing and it causes a high computational complexity to solve the problem. Since we restricted the running time as 5 hours for each stage, some data sets might contain relatively lower quality solutions (e.g. not optimal) than the data sets which were solved in time. In

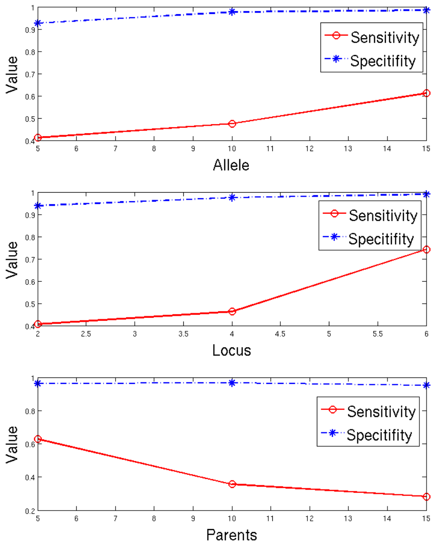

Figure 3: Individual pairwise accuracy with each parameter for simulated data sets.

Table 3: Performance comparison with existing approaches.

	GERUD2.0	COLONY	Our approach
MLAP	0.900	0.650	0.950
MLAP-m	0.825	0.825	0.925
Sensitivity	0.820	0.489	0.911
Specificity	0.760	0.999	0.800
Parentage accuracy	0.969	0.531	1.000
CPU time (second)	12	41	≤ 1

Figure 2, parental inference accuracy was shown to be unaffected by the changing of parameters. Although it has been shown to be slightly difference with alleles increasing, most results were not changed drastically along with parameters changing.

4.4 Comparison with Existing Approaches

In this section, we evaluated our approach compared to other existing approaches such as COLONY [17] and GERUD 2.0 [14]. Although there exist several applications for sibship reconstruction and parentage inference, COLONY and GERUD 2.0 seem to be appropriate applications for the performance comparison considering their availabilities and objectives. Even though all approaches were based on different algorithm, all of these algorithms could generate sibling reconstruction assignment result.

We set performance measurements which were group assignment accuracy (MLAP, MLAP-m), pairwise accuracy (sensitivity, specificity), and parental inference accuracy. Since GERUD 2.0 had several restrictions to process, we used a simple data set which could be properly solved in all of methods. We set input data as following: The number of females: 1, the number of males: 2, the number of alleles per loci: 15, the number of loci: 4, the number of groups: 2, and the number of offsprings per each group: 10. Basically, GERUD 2.0 supported only semi-monogamy mating system so we used semi-monogamy simulated data sets made by random population generator. Test results are shown at Table 3.

From Table 3, we observed that our approach showed relatively higher performance than others. Especially our approach provided 100% match for parental inference accuracy and showed fastest running time. The main reason was that our model was focused on showing effectiveness and optimality about those measurement performance unlikely other algorithms which were focused on feasibility. Actually, GERUD 2.0 was based on a brute - force algorithm. It may be easier to implement but less effective. When we compared with COLONY, our approach showed comparable performance result.

Since the GERUD 2.0 supported only monogamous mating system, we conducted additional tests with COLONY under three different mating systems for specified comparison. Basically, COLONY has an input parameter setting module as prior information such as mating type, species, analysis method, likelihood precision, run specification and sibship prior [17]. In our test, we set input parameters for COLONY as default except mating types selection. In COLONY, we selected different mating systems such as monogamy, semi-monogamy and polygamy, and conducted comparison tests using simulated data sets which had three different mating systems.

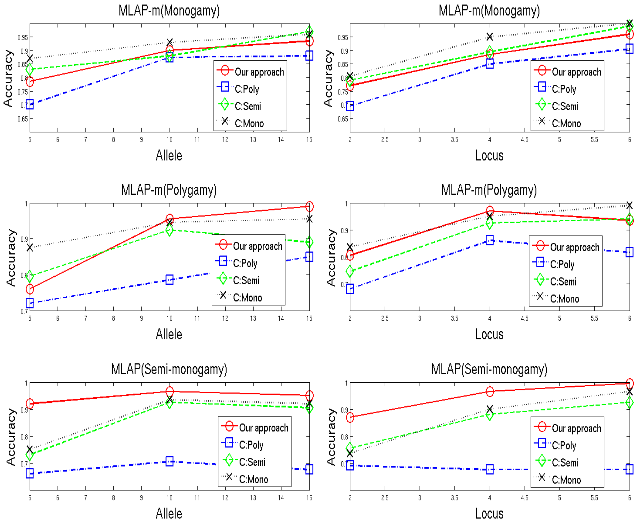

Figure 4: Group assignment accuracy (MLAP-m) for our approach and COLONY with three different mating systems.

In Figures 4, 5, 6 and 7, "C:Poly", "C:Semi" and "C:Mono" indicate that input parameter setting for COLONY is "Polygamy", "Semi-monogamy",and "Monogamy", respectively. We can observe the performance trend for our approach and COLONY when varying the number of alleles per loci and the number of sampled loci. In semi-monogamy and polygamy mating systems, our approach has shown relatively accurate solution than COLONY even COLONY was performed under correct setting. In monogamy mating system, COLONY has shown slightly better solution than ours. In Figure 7, for comparing processing time, our approach was faster than COLONY.

5. CONCLUSIONS

In this paper, we presented an integrated optimization framework for inferring sibship and parentage from single

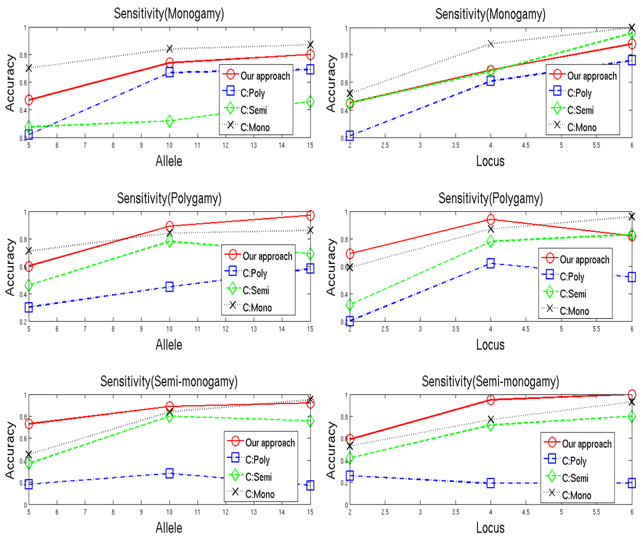

Figure 5: Individual pairwise accuracy (Sensitivity) for our approach and COLONY with three different mating systems.

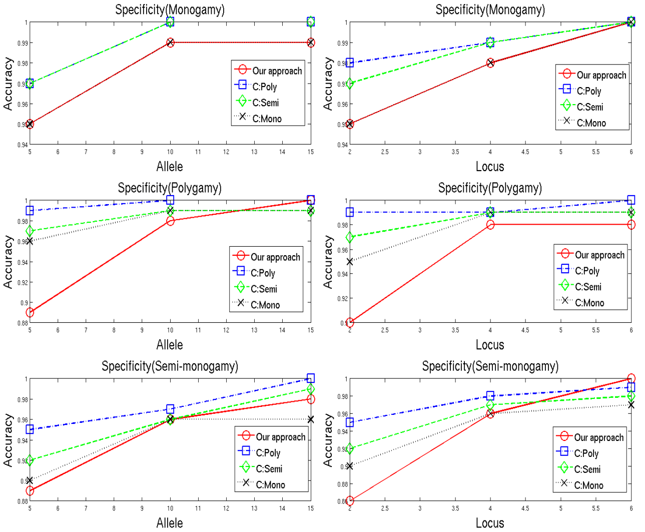

Figure 6: Individual pairwise accuracy (Specificity) for our approach and COLONY with three different mating systems.

generation microsatellite samples. Our new framework was developed and shown to be generalization of the sibship reconstruction and parentage inference. We implemented and tested our framework on both real biological and simulated data and compared the performance of our approach with other state-of-the art method. By analyzing the results, we confirmed that our approach provided accurate and robust solutions to sibship reconstruction and parentage inference. Our approach outperformed other existing methods.

6. ACKNOWLEDGMENTS

This work was supported by NSF under grants IIS-1231132.

7. REFERENCES

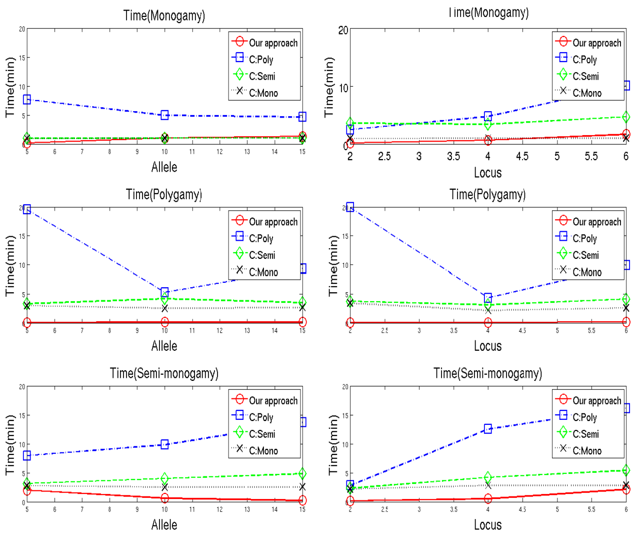

Figure 7: Processing time (minute) for our approach and COLONY with three different mating systems.

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

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

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

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

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

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

[7] J. A. Dewoody, Y. D. Dewoody, A. C. Fiumera, and J. C. Avise. On the number of reproductives contributing to a half-sib progeny array. *Genetics Research*, 75:95–105, 2000.

[8] J. A. Dewoody, D. Walker, and J. C. Avise. Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals. *Genetics*, 154:1907–1912, 2000.

[9] M. A. D. Goodisman and R. H. Crozier. Population and colony genetic structure of the primitive termite

mastotermes darwiniensis. *Evolution*, 56:70–83, 2002.

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

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

[12] D. Heg and R. van Treuren. Femal-female cooperation in polygynous oystercatchers. *Nature*, 391:687–691, 1998.

[13] A. Jones and J. C. Avise. Microsatellite analysis of maternity and the mating system in the gulf pipefish syngnathus scovelli, a species with male pregnancy and sex-role reversal. *Molecular Ecology*, 6:203–213, 1997.

[14] A. G. Jones. Gerud 2.0: a computer program for the reconstruction of parental genotypes from half-sib progeny arrays with known or unknown parents. *Molecular Ecology Notes*, 5:708–711, 2005.

[15] A. G. Jones and W. R. Ardren. Methods of parentage analysis in natural populations. *Molecular Ecology*, 12:2511–2523, 2003.

[16] A. G. Jones, C. M. Small, K. A. Paczolt, and N. L. Ratterman. A practical guide to methods of parentage analysis. *to appear in Genetics*, 10(1):6–30, 2010.

[17] O. Jones and J. Wang. Colony: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10:551–555, 2009.

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

[19] P. Morin, J. Moore, R. Chakraborty, L.Jin, and J.Goodal. Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, 265:1193–1201, 1994.

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

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

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