Error Tolerant Sibship Reconstruction in Wild Populations

Saad I. Sheikh[1], Tanya Y. Berger-Wolf[1], Mary V. Ashley[3], Isabel C. Caballero[3],
Wanpracha Chaovalitwongse[2], Bhaskar DasGupta[1]

[1] Dept. of Computer Science, University of Illinois at Chicago, {ssheikh,tanyabw,dasgupta}@cs.uic.edu
[2] Dept. of Industrial Engineering, Rutgers University, wchaoval@rci.rutgers.edu,
[3] Dept of Biological Sciences, University of Illinois at Chicago, ashley@eeb.uic.edu

Abstract. Kinship analysis using genetic data is important for many biological applications, including many in conservation biology. Wide availability of microsatellites has boosted studies in wild populations that rely on the knowledge of kinship, particularly sibship. While there exist many methods for reconstructing sibling relationships, almost none account for errors and mutations in microsatellite data, which are prevalent and affect quality of reconstruction. We present an error-tolerant method for reconstructing sibling relationships based on the ideas of consensus methods. We test our approach on both real and simulated data, with both pre-existing and introduced errors. Our method is highly accurate on almost all simulations, giving over 90% accuracy in most cases. Ours is the first method designed to tolerate errors while making no assumptions about the population or the sampling.

Keywords: Sibship Reconstruction, Kinship Analysis, Consensus, Combinatorial Optimization.

1 Introduction

Kinship analysis of wild populations is an important and necessary component of studying mating systems, dispersal patterns and kin selection. In wild populations, kinship relationships (lower order pedigree) are typically inferred from microsatellite markers, rather than SNPs which are more commonly used in model organisms (see [6] for a discussion). There are two main approaches to kinship inference from microsatellite data: using genetic distance estimates and statistical likelihood methods [1, 8, 11, 24, 25], and enumeration of feasible relationships based on Mendelian constraints [2, 5, 6, 10, 23]. However, with the exception of COLONY [25], none of the existing kinship reconstruction methods is designed to tolerate genotyping errors or mutation. Yet, both errors and mutation cannot be avoided in practice and identifying these errors without any prior kinship information is a challenging task. In [5, 6, 10, 23] we have presented a method for reconstructing sibling relationships from single generation microsatellite data that optimally identifies the most parsimonious set of sibling groups subject only to Mendelian inheritance constraints. We have shown that our method performs comparably or better than other sibling reconstruction approaches on both biological and simulated data. While our method was not designed for data with genotyping errors, it did perform relatively well on data that contained a limited number of errors. In this paper we present a new approach for reconstructing sibling relationships from microsatellite data designed explicitly to tolerate genotyping errors and mutation in data.

1.1 Microsatellite Markers

While there are several molecular markers used in population genetics, microsatellites (also known as SSRs, STRs, SSLPs, and VNTRs) are the most commonly used in population biology for non-model organisms. Microsatellites are repeats of short DNA sequences distributed throughout the genome. These are co-dominant, unlinked, multi-allelic markers that offer numerous advantages for population studies. Generally, phase or haplotype information is not available for microsatellite loci in non-model organisms.

1.2 Sibling Reconstruction Problem Statement

The main focus of our paper is to design a method that accurately reconstructs sibling groups from microsatellite data of a single generation in presence of genotyping errors and mutations. We have formally defined the problem of sibling reconstruction in [6] and we restate it here. Let $U = \{X_1, ...X_n\}$ be a population U of n diploid individuals of the same generation where each individual is represented by a genetic (microsatellite) sample at l loci. That is, $X_i = (< a_{i1}, b_{i1} >, ..., < a_{il}, b_{il} >)$ and a_{ij} and b_{ij} are the two alleles of the individual i at locus j represented as some identifying string. The goal is to reconstruct the full sibling groups which is a partition of individuals into $P_1, ...P_m$ where individuals in the same partition P_i have the same parents. We assume no knowledge of parental information.

1.3 2-Allele Algorithm

In [6] we presented a combinatorial 2-Allele Min set cover algorithm for the siblings reconstruction problem. We rely on Mendelian inheritance constraints that dictate that full siblings must share their parents' alleles at all loci. We formalize this rule as the 2-Allele Property in [5]: For a set of individuals there exists a swapping of individuals' alleles within a locus such that the total number of distinct alleles on each side at this locus is at most 2. Note, that the 2-allele property is a necessary constraint for a group of individuals to be siblings but not sufficient. Notice, also, that *any* two individuals necessarily satisfy the 2-allele property since by default the number of alleles on each side of any locus is at most two.

The 2-Allele Min set cover algorithm works by first generating all maximal sibling groups that obey the 2-allele property. The algorithm then uses set cover [20] to find the minimum number of sibling groups necessary to explain the data.

1.4 Errors in Microsatellite Data

Errors and mutation cannot be verifiably avoided when genotyping wild populations. While there may be several sources and types of errors (see [14, 18]), we are concerned primarily with how they affect the sibling reconstruction problem. We now discuss errors typically present in microsatellite data.

Allelic Dropout occurs when one or both alleles are not amplified during polymerase chain reaction (PCR) and is one of the most common errors [14]. If one of the alleles is not amplified, the result can be taken as a homozygote. The case when both alleles are missing is easily identifiable and is handled by a simple extension of the 2-allele algorithm (see section 1.5).

Heterozygous Mistype occurs when two alleles are amplified by PCR but one or both of them, for a variety of reasons, are not recorded as present. In the context of sibling reconstruction any allele that was not present in either of the parents is a mistype.

Homozygous Mistype occurs when only one allele is amplified by PCR, and it is not any of the parental alleles.

Genetic Mutation is the actual variation in the alleles, also called *polymorphism*. This arises from mistakes made during DNA replication. A mutation may also be classified as Mistypes when reconstructing sibling relationships.

Allele Combination Error occurs when one or both alleles at a locus are present in the parents (or sibling group) but Mendelian inheritance rules are still violated.

Null Alleles is the lack of any amplification. When no allele is amplified it can be explicitly marked as a missing allele.

1.5 Accommodating Missing Alleles

To accommodate known missing alleles in the data we denote them by a special symbol, *e.g.* a wildcard (*). The 2-Allele Min set cover algorithm then proceeds to construct feasible sibling sets treating the wildcard as any possible allele.

1.6 Consensus Methods

We base our idea of error-tolerant sibling reconstruction on the consensus-based approach. The idea behind consensus methods is to combine different solutions to the same problem into one solution, *i.e.*, group decision making. The formal theory of voting and group decision making dates back to the eighteenth century [12, 13] and modernized by Kenneth J. Arrow in 1951 [3]. Recently mathematical and computational group choice and concensus techniques have been applied to biological problems, mostly in the context of phylogeny reconstruction [9]. Our solution is based on using such methods to tolerate genotyping errors. In Section 2.1 we define consensus in the context of siblings reconstruction problem, and discuss some approaches and their feasibility.

2 Consensus based approach for error-tolerant siblings reconstruction

We now describe our approach to reconstructing sibling relationships in presence of genotyping errors. Consider an individual X_i which has some genotyping error(s). Any error that is affecting siblings reconstruction must be preventing X_i's sibling relationship with at least one other individual X_j, who in reality is a sibling. It is unlikely that an error would cause two unrelated individuals to be paired up as siblings, unless all error-free loci do not contain enough information. It is possible that an individual has more than one error, yet it is unlikely that all the errors bias the solution in the same direction.

Thus, we can discard one locus at a time, assuming it to be erroneous, and obtain a sibling reconstruction solution based on the remaining loci. If all such solutions put the individuals X_i and X_j in the same sibling group (*i.e.*, there is a consensus among those solutions), we consider them to be siblings. The bulk of our error-tolerant approach is concerned with pairs of individuals that do not consistently end up in the same sibling group during this process, that is, there is no consensus about their sibling relationship.

We present a formal definition of consensus in the context of sibling reconstruction and describe our consensus-based algorithm for error- tolerant sibling reconstruction.

2.1 Consensus Methods for Siblings Reconstruction

Recall that for a population of individuals $U = \{X_1 \ldots X_n\}$ the goal of a siblings reconstruction problem is to find a partition of the population into sibling groups $S = \{P_1 \ldots P_m\}$ where all individuals are covered with no overlap:

$$\bigcup_{1 \leq j \leq m} P_j = U \qquad \text{and} \qquad \forall j, k \;\; P_j \cap P_k = \emptyset$$

A partition defines an equivalence relationship. Two individuals are equivalent if they are in the same partition of the solution S.

$$X_i \equiv_S X_j \iff \exists P_k \in S \;\text{ s.t. }\; X_i \in P_k \wedge X_j \in P_k$$

We are now ready to give the definition of a consensus method.

Definition 1. *A* consensus *method for sibling groups is a computable function f that takes k solutions $\mathcal{S} = \{S_1, ..., S_k\}$ as input and computes* one *final solution.*

Definition 2. *A* strict consensus *[22] $\mathcal{C}$ is a partitioning of sibling groups where two individuals are together only if they were in the same partition for all solutions:*

$$\mathcal{C} = \{P_{\mathcal{C},1} \ldots P_{\mathcal{C},m}\} \qquad \textit{where} \quad X_j \equiv_{\mathcal{C}} X_k \iff \forall S_i \in \mathcal{S} \;\; X_j \equiv_{S_i} X_k$$

The strict consensus defines a true equivalence relation and, thus, is a transitive function:

$$(X_i \equiv_{\mathcal{C}} X_j \quad \wedge \quad X_j \equiv_{\mathcal{C}} X_k) \quad \Rightarrow \quad X_i \equiv_{\mathcal{C}} X_k$$

Any individual that is not consistently placed into a partition in all solutions will be added as a singleton. Such a consensus solution is reliable for the individuals that have been placed together in a group, but there may be many singleton groups.[4]

2.2 Distance-based consensus

The 2-Allele Algorithm finds the most parsimonious solution with the fewest number of sibling groups. While the algorithm performs well in absence of errors, it is not designed to handle errors. Moreover, the resulting sibling groups returned by the algorithm may overlap. The strict consensus, on the other hand, conservatively identifies reliable sibling relationships and puts the rest back into singleton groups. In order to combine the best aspects of both methods we present a distance based consensus method. We start with a strict consensus of the "leave-one-locus-out" solutions and search for the *nearest good parsimonious solution.* In order to search for such a solution we need quantitative measures to 1) assess quality of a solution, f_q, and 2) calculate the pairwise distance between solutions, f_d. Assume that we have the two functions f_q and f_d.

$$f_q : S \to \mathbf{R} \qquad \text{and} \qquad f_d : S \times S \to \mathbf{R}$$

Since we start with a strict consensus $\mathcal{C}$ the partitions in the solution cannot be refined any further. Therefore to improve the solution, we use the operations of merging two sets. The following monotonic property must be obeyed by any improved solution $\mathcal{C}'$:

$$\forall X_i, X_j \in U \quad X_i \equiv_{\mathcal{C}} X_j \implies X_i \equiv_{\mathcal{C}'} X_j. \tag{1}$$

Thus, given a solution $\mathcal{C}$, we look for an improved solution $\mathcal{C}'$ that minimizes $f_d(\mathcal{C}, \mathcal{C}')$ and maximizes $f_q(\mathcal{C}')$. To combine the two objectives we can formulate the following optimization problems:

1. Maximize f_q with an upper bound on f_d
2. Minimize f_d with a lower bound on f_q
3. Maximize/Minimize some (linear) combination of f_d and f_q

We prove all of these problems to be NP-Hard in general for arbitrary f_q and f_d.

[4] If we relax the above constraint to require not all but most of the solutions to agree on the equivalence relationship it gives us a "majority consensus". While it performs well in other applications, such as phylogeny reconstruction [7], it is too biased towards loci with errors in the context of sibling reconstruction.

Theorem 1. *Let $\mathcal{C}$ be a collection of sibling groups and $k \in \mathbf{R}$. Let $\mathcal{S}$ be the set of all solutions that are an improvement of $\mathcal{C}$ and are obtainable from $\mathcal{C}$ by merging sibling sets. The problem of finding an improved solution $\mathcal{C}' \in \mathcal{S}$ such that*

$$f_q(\mathcal{C}') = \max_{\substack{S \in \mathcal{S} \\ f_d(\mathcal{C},S) \leq k}} f_q(S)$$

is NP-hard.

Proof. We show that this problem is NP-hard by reducing from the 2-ALLELE MIN SET COVER problem, which we have proven to be MAX SNP-hard [4]. We start with a collection $\mathcal{C}$ of singleton only sets and aim to minimize the number of sibling groups.
Formally, for an input $U = \{X_1, ...X_n\}$ to the 2-ALLELE MIN SET COVER, the corresponding input to the distance-based consensus problem is $\mathcal{C} = \{\{X_1\}, ..., \{X_n\}\}$ and $k = 0$. We define the distance function f_d to be

$$f_d(\mathcal{C}, \mathcal{C}') = \begin{cases} 0 & \text{sibling groups in } \mathcal{C} \text{ can be merged to form } \mathcal{C}' \text{ without violating 2-allele property at any locus} \\ 1 & \text{otherwise} \end{cases}$$

Finally, we define the quality function $f_q(\mathcal{C}') = |U| - |\mathcal{C}'|$. This ensures the minimum number of sets to maximize the objective function since $|\mathcal{C}'| < |U|$.
The bound on f_d guarantees that any merged sibgroups obey the 2-allele property and the quality maximization objective ensures that the solution is a minimum set cover. □

Thus, finding improved solutions subject to the first objective is NP-hard. We now show that the second objective is NP-hard as well.

Theorem 2. *Let $\mathcal{C}$ be a collection of sibling groups and $k \in \mathbf{R}$ be the lower bound on f_q. Let $\mathcal{S}$ be the set of all solutions that are an improvement of $\mathcal{C}$ and are obtainable from $\mathcal{C}$ by merging sibling sets. The problem of finding an improved solution $\mathcal{C}' \in \mathcal{S}$ such that*

$$f_d(\mathcal{C}, \mathcal{C}') = \min_{\substack{S \in \mathcal{S} \\ f_q(S) \geq k}} f_d(\mathcal{C}, S)$$

is NP-hard.

Proof. Similar to the proof of Theorem 1 above, we again reduce from the 2-ALLELE MIN SET COVER problem. Given an input $U = \{X_1, ...X_n\}$ to the 2-ALLELE MIN SET COVER, the corresponding input to the distance-based consensus problem is $\mathcal{C} = \{\{X_1\}, ..., \{X_n\}\}$ and $k = n$. We define the distance function f_d as follows:

$$f_d(\mathcal{C}, \mathcal{C}') = \begin{cases} \infty & \text{sibling groups in } \mathcal{C} \text{ can be merged to form } \mathcal{C}' \text{ without violating 2-allele property at any locus} \\ 1 & \text{otherwise} \end{cases}$$

We define the quality function f_q as the sum of the distance from strict consensus and the number of sets:

$$f_q(\mathcal{C}') = f_d(\mathcal{C}, \mathcal{C}') + |C'|.$$

A reduction from the 2-ALLELE MIN SET COVER problem follows. □

Lastly, for an arbitrary combination of f_q and f_d objective 3 is unattainable as well.

Theorem 3. *Let $\mathcal{C}$ be a collection of sibling groups. Let $\mathcal{S}$ be the set of all solutions that are an improvement of $\mathcal{C}$ and are obtainable from $\mathcal{C}$ by merging sibling sets and let $g(f_q, f_d)$ be a (linear) combination of the functions f_q and f_d. The problem of finding an improved solution $\mathcal{C}' \in \mathcal{S}$ such that*

$$g(f_d(\mathcal{C}, \mathcal{C}'), f_q(\mathcal{C}')) = \underset{S \in \mathcal{S}}{OPT}\{g(f_d(\mathcal{C}, S), f_q(S))\}$$

is NP-hard.

Proof. This theorem follows from the Theorem 1 (OPT is max) and Theorem 2 (OPT is min). Hence both the minimization and the maximization objectives are NP-Hard. □

We have shown that the three versions of the problem of finding the closest best solution to a given solution of the sibling reconstruction problem are NP-hard. In the next section we present a heuristic approach that efficiently finds good solutions.

3 Greedy Distance-Based Consensus

We now present a greedy algorithm that given a collection of sibling reconstructions attempts to find a good solution with few sibling groups while allowing for a small number of errors in the data. The Greedy Consensus Algorithm uses costs associated with errors in data to define a merging cost and to find and merge the pair of sibling groups with the minimum (merging) cost. The sibling groups to be merged are selected by exhaustively examining all pairs of groups and identifying the merge that results in *lowest total merging cost* for the *merged* group. Our quality function based on the parsimony assumption: we try to find the minimum number of sibling groups and errors that explain the data. Therefore, to get the minimum number of sibling groups our quality function is defined as $f_q = |U| - |\mathcal{C}|$.

3.1 Distance Function

We define two functions necessary to calculate the distance f_d : the cost and the benefit of assigning an individual to a sibling group. The cost of an assignment is used when an individual cannot be assigned to a group without violating the 2-allele property. The total cost of tolerating errors is computed using user-defined costs for each type of possible error in data. These costs are provided by the user depending on the expected error rates and number of loci. By default, these may be uniform. The *benefit* of an assignment is determined by the shared alleles and allele pairs of the new individual, which can be added *without violating 2-allele property.*

More formally, we assume that we are given as an input the relative costs of the four distinct error types and the upper bounds on the number of errors per individual, per sibling group, and per individual in a sibling group. [5] The cost and the benefit of assigning an individual X to a sibling group P_i is defined as:

$$f_{assign}(P_i, X) = \begin{cases} benefit & \text{If the 2-allele property is not violated in adding } X \text{ to } P_i \\ cost & \text{otherwise} \end{cases}$$

[5] Note that COLONY [25], the only other sibling reconstruction method that explicitly tolerates errors in data, requires considerably more detailed information about the types, costs, and frequencies of errors.

Suppose $\mathcal{C} = \{P_1, ..., P_m\}$ is a collection of sibling groups and $\mathcal{C}'$ is a collection of groups obtained from $\mathcal{C}$ by merging groups P_i and P_j. Then we define the distance between $\mathcal{C}$ and $\mathcal{C}'$ as follows:

$$f_d(\mathcal{C}, \mathcal{C}') = \min \left\{ \sum_{X \in P_i} f_{assign}(P_j, X), \sum_{X \in P_j} f_{assign}(P_i, X) \right\}$$

3.2 Greedy Algorithm

Given an upper bound on the number of errors and the relative error costs, Greedy Consensus algorithm searches for the solution with the fewest number of sibling groups and errors necessary to explain the data. We denote by $U_{|i}$ the set of individuals U with the ith locus omitted. The Greedy Consensus algorithm has three phases:

1. Calculate the 2-allele min set cover solutions for $U_{|1} \ldots U_{|l}$
2. Calculate the strict consensus $\mathcal{C}$ of the above solutions
3. Merge sibling groups greedily as allowed by the parameters

Phase 1 runs the 2-allele min set cover algorithm to obtain solutions for dropping one locus. Any technique for siblings reconstruction may be used here. We use the 2-allele min set cover algorithm as the basis since it performs as well or better than other available methods and makes fewest assumptions [6]. Phase 2 works by examining all the solutions from phase 1 and placing two individuals in the same sibling group only if all the solutions agree. Unpaired individuals are placed in singleton groups. Finally, phase 3 works iteratively by merging the *closest* pair of sibling groups. This is done by calculating the f_d distance for all pairs of sibling groups at every iteration. The pair that gives the smallest distance is merged. This continues until the minimum distance is greater than either the *maximum editing cost per sibling group* or the average edit cost exceeds *maximum average editing cost per sibling group*. Both of these costs are input parameters.

To analyze the computational time complexity of the Greedy Consensus algorithm we consider each phase separately. Computing the 2-allele min set cover for each subset of the input is the most expensive part of the algorithm. The 2-allele min set cover problem is MAX SNP-hard [4], which means that it cannot be approximated within some constant factor in polynomial time, unless $P = NP$. We use the commercial mixed integer program solver CPLEX[6] to solve the problem to optimality. Greedy Consensus algorithm executes l runs of 2-allele min set cover to compute the l solutions for consensus method.

The consensus part (steps 2-3) of Greedy Consensus algorithm is polynomial: The total time for $O(n)$ iterations is $O(n^3 l)$. Note that our approach is not exclusive to the 2-allele min set cover but may be used with a faster algorithm for a base solution.

4 Experimental Methodology

We tested our approaches on random datasets generated by coupling the random simulations used in [5, 6, 10] and adding random errors to the dataset. We compared results of our Greedy Consensus algorithm with the original 2-allele min

[6] CPLEX is a registered trademark of ILOG

set cover [6] as well as those of the Family Finder software [8] and a limited comparison to COLONY software [25].

We also tested our approach on biological datasets with known sibling groups: Tiger Shrimp *Penaeus monodon* ([19]), Ants *Leptothorax acervorum* [16], and Atlantic Salmon *Salmo salar* [17]. Only the Shrimp dataset had original errors in it. We introduced errors into other datasets to test our approach.

4.1 Random Simulations

We validate our approach using random simulations. We first create random diploid parents and then generate complete genetic data for offspring varying the number of males, females, alleles, loci, number of offsprings and juveniles. We then introduce errors into the data and use various methods to reconstruct the sibling groups. We compare our results to the actual known sibling groups in the data to assess accuracy. We measure the error rates using the Gusfield's partition distance [15]. The base population is generated using uniform distribution as described in [6].

We used the following ranges of parameter settings for the fixed error rate of 10% of individuals:

- The number of adult females F and the number of adult males M were equal and set to $5, 20$.
- The number of loci sampled $l = 4, 6, 8, 10$
- The number of alleles per locus (for the uniform allele frequency distribution) $a = 10, 15$.
- The the number of true sibling groups $j = 5, 10, 20$.
- The maximum number of offspring per couple (for the uniform family size distribution) $o = 5, 10$.

4.2 Random Errors

Errors were introduced uniformly at random with a probability of 0.1 of an individual having an error. Once an individual to have an error is chosen, we choose the locus to introduce an error uniformly at random. Then the type of error to introduce is chosen by generating a random number between 0 and 1, and choosing the corresponding error from Table 1(a). While the probability of 0.1 for having an individual with an error may seem large, it is meant to exhibit how robust our method is to genotyping errors. It also is affected by the number of loci since we introduce only one erroneous locus for an individual. We further test our approach by varying the error rate for selected parameters.

Table 1. Random Errors and Associated Costs

(a) Error ranges for the different error types

Type of Error	Random Number Range
Allelic Dropout	$[0, 0.5]$
Heterozygous Mistype	$(0.5, 0.7]$
Homozygous Mistype	$(0.7, 0.95]$
Genetic Mutation	otherwise

(b) Costs and relative thresholds used for Greedy Algorithm Simulations

Cost	value
Allelic Dropout	0.34
Heterozygous Mistype	0.7
Homozygous Mistype	1
Allele Combination Error	0.4
Maximum Editing per individual	2.0
Maximum Editing per group	∞
Maximum Avg Edit Ind in Group	0.45

The error rates in Table 1(a) has been derived from biological data as well as [14]. The values used in our experiments are shown in Table 1(b).

4.3 Evaluation

We measure the accuracy of the solution by comparing the known sibling sets with those generated by our algorithm, and calculating the minimum partition distance [15]. The solution error is the percentage of individuals that would need to be removed to make the reconstructed sibling set equal to the true sibling sets. Note that the 2-ALLELE MIN SET COVER does not return a partitioning of the individuals, whereas the GREEDY CONSENSUS ALGORITHM partitions them.

The experiments were run on Intel^{TM} Quad Core Xeon Processor (2.66GHz) with 24 GB RAM memory.

4.4 Sibling Group Reconstruction Methods

We compare the performance of the GREEDY CONCENSUS ALGORITHM to the 2-ALLELE MIN SET COVER algorithm and two other sibship reconstruction methods. While there are other sibling reconstruction methods available, in our evaluations, partially presented in [6], Family Finder and COLONY, together with the 2-ALLELE MIN SET COVER, were the best.

Family Finder. The approach proposed in [8] is a mixture of likelihood and combinatorial techniques. The algorithm constructs a graph with individuals as nodes and the edges weighted by the pairwise likelihood ratio that the individuals are siblings versus being unrelated. Very light edges are ignored. Sibling groups are then dense areas of the graph.

Colony. Wang [25] has proposed the only other known error tolerant approach. The method uses a simulated annealing algorithm that works by starting with known sibling groups. Similar to the consensus approach, individuals whose sibling groups are not known are placed into singleton sibling groups. Iteratively alternate solutions are created by randomly changing group memberships of individuals. It uses a "cool-down" approach to reduce exploration after a large number of iterations.

5 Results

We have compared the accuracy of reconstruction of sibling groups by the new error-tolerant approach to the best existing sibling reconstruction methods. We use simulated data with a wide range of parameters. On simulated data our GREEDY CONSENSUS algorithm performs better than all other methods on almost all parameters. When the number of loci is small the 2-ALLELE MIN SET COVER performs better in some cases, but overall the consensus method performs best on simulated data. Both Family Finder and COLONY are very inaccurate when the number of loci is small, thus making them expensive for wild populations. For all simulations with 6 or more loci, our approach was 95% or more accurate, even if the number of erroneous individuals went up to 20%. Family Finder and COLONY showed considerable improvement with increase in the number of loci and alleles per locus.

We present the results on simulated data in Figure 1. We show the accuracy as a function of the number of sampled loci, number of alleles per locus, number of families, and the size of a family.

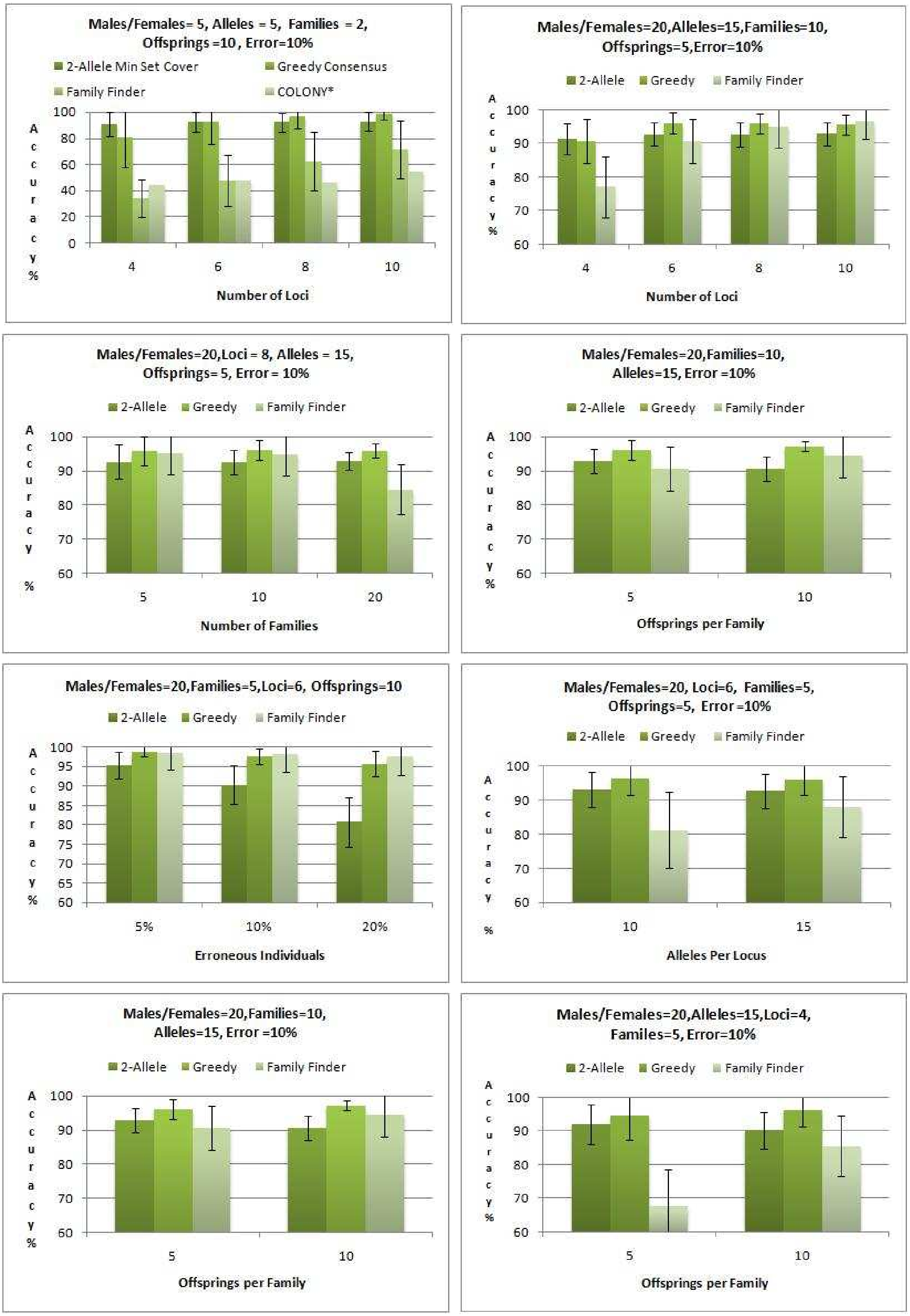

Fig. 1. Results on simulated datasets. Only 50 iterations were used for the COLONY algorithm due to its computational inefficiency and time constraints.

On biological data all methods performed comparably well with slight variation around the 90% accuracy. The concensus approach achieved over 90% accuracy for all the biological datasets, which was slightly better than the 2-ALLELE MIN SET COVER.

6 Conclusions

We have proposed an error-tolerant approach for reconstructing sibling relationships from microsatellite data. Our method is based on the idea of taking a consensus of partial solutions obtained by omitting one locus at a time and then locally improving the resulting combined solution. We have proven the intractability of any general formulation of distance based consensus methods. We have proposed a new combinatorial algorithm for the problem of reconstructing sibling relationships from single generation microsatellite genetic data in presence of genotyping errors. We have implemented and tested our approach on both simulated and real data. We have provided a framework for distance based consensus methods which may be used with any combination of distance and quality functions, possibly yielding to better results.

Consensus methods give a partition, unlike a set cover, and the proposed GREEDY CONSENSUS ALGORITHM has over 95% accuracy for most datasets and performs comparably or better than other approaches in most cases. Moreover, unlike any other approach, our method can *identify* errors in a given dataset.

While Family Finder and COLONY perform comparably well in some scenarios, our method requires considerably less input, makes fewer assumptions, and is consistently over 90% accurate. Moreover, our method is considerably faster than COLONY which performs an almost exhaustive search for a global minimum of the likelihood function. COLONY also requires one of the parents to be monogamous which is an unrealistic assumption for many species. Family Finder does not perform well for large families, especially if the allele frequency is low.

6.1 Future Work

Our approach can be combined with a variety of methods for both generating the input solutions, and to develop a consensus among them. In future we intend to explore other than greedy optimization objectives to try to avoid local minima in the distance function.

Our technique can be extended to solve other problems in kinship analysis. Since our approach is not restricted to the methods used for generating input solutions, it can be used as a general consensus between different methods of sibling reconstruction. For example, a tree-based consensus method can be used to merge pedigrees.

Acknowledgments

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

References

1. A. Almudevar. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, 63, 2003.
2. A. Almudevar and C. Field. Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:136–165, 1999.
3. K. J. Arrow. *Social Choice and Individual Values.* John Wiley, New York, second edition, 1963.
4. M. Ashley, T. Y. Berger-Wolf, P. Berman, W. Chaovalitwongse, B. DasGupta, and M.-Y. Kao. On approximating four covering/packing problems with applications to bioinformatics. Technical report, DIMACS, 2007.
5. T. Y. Berger-Wolf, B. DasGupta, W. Chaovalitwongse, and M. V. Ashley. Combinatorial reconstruction of sibling relationships. In *Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics (CBGI 05)*, pages 1252–1255, Utah, July 2005.
6. T. Y. Berger-Wolf, S. I. Sheikh, B. DasGupta, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, S. P. Lahari. Reconstructing sibling relationships in wild populations. *Bioinformatics*, 23(13), 07.
7. T. Y. Berger-Wolf, T. L. Williams, B. E. Moret, and T. J. Warnow. An experimental evaluation of phylogenetic consensus methods. Technical Report TR-CS-2003-19, Department of Computer Science, University of New Mexico, 2003.
8. Jen Beyer and B. May. A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, 12:2243–2250, 2003.
9. O. R. P. Bininda-Emonds. MRP supertree construction in the consensus setting. In M. Janowitz, F.J. Lapointe, F. McMorris, B. Mirkin, and F. Roberts, editors, *Bioconsensus*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. DIMACS-AMS, 2001.
10. W. Chaovalitwongse, T. Y. Berger-Wolf, B. Dasgupta, and M. V. Ashley. Set covering approach for reconstruction of sibling relationships. *Optim, Methods and Soft.*, 22(1):11 – 24, Feb 2007.
11. S. C.Thomas and W. G.Hill. Sibship reconstruction in hierarchical population structures using markov chain monte carlo techniques. *Genet. Res., Camb.*, 79:227–234, 2002.
12. J.C. de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sci.*, 1784.
13. Marie Jean Antoine Nicolas de Caritat marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions, 1785.
14. P. Gagneux, C. Boesch, and D. S. Woodruff. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear dna amplified from shed hair. *Mol.Eco.*, 6(9):861-8,Sep 1997.
15. D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, May 2002.
16. R.L. Hammond, A.F.G. Bourke, and M.W. Bruford. Mating frequency and mating system of the polygynous ant, leptothorax acervorum. *Molecular Ecology*, 10(11):2719–2728, 1999.
17. C. M. Herbinger, P. T. O'Reilly, R. W. Doyle, J. M. Wright, and F. O'Flynn. Early growth performance of atlantic salmon full-sib families reared in single family tanks or in mixed family tanks. *Aquaculture*, 173(1–4):105–116, 1999.
18. J. I. Hoffman and W. Amos. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, 14(2):599–612, 2005.
19. D. R. Jerry, B. S. Evans, M. Kenway, and K. Wilson. Development of a microsatellite dna parentage marker suite for black tiger shrimp penaeus monodon. *Aquaculture*, 542-547, 2006.
20. R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, 85–103. Plenum Press, 1972.
21. D. A. Konovalov, C. Manning, and M. T. Henshaw. KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Mol. Ecol. Notes*, 2004.
22. F. R. McMorris, D. B. Meronik, and D. A. Neumann. A view of some consensus methods for trees. In J. Felsenstein, editor, *Numerical Taxonomy*, pages 122–125. Springer-Verlag, 1983.
23. S. I. Sheikh, T. Y. Berger-Wolf, W. Chaovalitwongse, and M. V. Ashley. Reconstructing sibling relationships from microsatellite data. In *Proc. of the Europ. Conf. on Comp.Bio.(ECCB)*, Jan 07.
24. B. R. Smith, C. M. Herbinger, and H R. Merry. Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, 158(3):1329–1338, July 2001.
25. J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, 166:1968–1979, April 2004.