In: xxxx
Editor: xxxx, pp. 1-28

ISBN xxxx

Chapter xx

Full Sibling Reconstruction in Wild Populations From Microsatellite Genetic Markers

***Mary V. Ashley**[*] **Tanya Y. Berger-Wolf**[†] **Isabel C. Caballero**[†],*
***Wanpracha Chaovalitwongse**[‡] **Bhaskar DasGupta**[†] **Saad I. Sheikh**[†]*

> *I do not believe that the accident of birth makes people sisters and brothers. It makes them siblings. Gives them mutuality of parentage.*
>
> – Maya Angelou

Abstract

New technologies for collecting genotypic data from natural populations open the possibilities of investigating many fundamental biological phenomena, including behavior, mating systems, heritabilities of adaptive traits, kin selection, and dispersal patterns. The power and potential of genotypic information often rests in the ability to reconstruct genealogical relationships among individuals. These relationships include parentage, full and half-sibships, and higher order aspects of pedigrees. Some areas of genealogical inference, such as parentage, have been studied extensively. Although methods for pedigree inference and kinship analysis exist, most make assumptions that do not hold for wild populations of animals and plants.

In this chapter, we focus on the full sibling relationship and first review existing methods for full sibship reconstructions from microsatellite genetic markers. We then describe our new combinatorial methods for sibling reconstruction based on simple

[*] Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60607. Email: ashley@uic.edu

[†] Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607. email: {tanyabw,dasgupta,ssheikh}@cs.uic.edu

[‡] Department of Industrial Engineering, Rutgers University, Piscataway, NJ 08854. email: wchaoval@rci.rutgers.edu

Mendelian laws and its extension even in the presence of errors in the data. We also describe a generic consensus method for combining sibling reconstruction results from other methods. We present experimental comparison of the best existing approaches on both biological and simulated data. We discuss relative merits and drawbacks of existing methods and suggest a practical approach for reconstructing sibling relationships in wild populations.

1. Introduction

Kinship analysis of wild populations is often an important and necessary component of understanding an organism's biology and ecology. Population biologists studying plants and animals in the field want to know how individuals survive, acquire mates, reproduce, and disperse to new populations. Often these parameters are difficult or impossible to infer from observational studies alone, and the establishment of kinship patterns (parentage or sibling relationships, for example) can be extremely useful. The powerful toolbox provided by advances in molecular biology and genome analysis has offered population biologists a growing list of possibilities for inferring kinship. Paternity analysis in wild populations became common upon the arrival of the first DNA-based markers in the mid-1980s, when multi-locus DNA fingerprinting methods became available. Probably the most notable discoveries came from studies of avian mating systems. Multi-locus DNA fingerprinting revealed that many bird species that were behaviorally monogamous were in fact often reproductively promiscuous. Females of such species would furtively engage in extra-pair copulations, apparently unbeknownst to their cuckolded male social partners. In fact, the frequency of extra-pair fertilizations (up to 50% in some species) led avian behavioral ecologist to distinguish between *social mating systems* and *genetic mating systems* (reviewed in [55]). The invention of the polymerase chain reaction (PCR) [38] quickly led to the replacement of multi-locus fingerprinting with single-locus PCR-based techniques by the mid 1990s [3, 39]. Microsatellites (also known as SSRs and STRs) were the first and still are the most widespread molecular marker for inferring kinship in wild populations, although their development in each new species studied is often a time-consuming and expensive obstacle. Microsatellite genotypes, which could be obtained from tiny amounts of blood, tissue, or even feces, have been used to infer parentage, particularly paternity, in a large number of wild species. Notable examples include the study of pollination patterns in forest trees [13, 14, 47], identifying fathers of the famed chimpanzees of Gombe [12], and evaluating the success of alternative mating strategies used by male big horn sheep [24]. A breakthrough in paternity assignment came with the release of the software program CERVUS [30] that provided a user-friendly Windows-based program that employed a statistical likelihood method to assign paternity to a candidate father with an estimated level of statistical confidence.

There are many cases where field studies can sample cohorts of offspring yet sampling putative parents is problematic. In these cases, sibling relationships (sibship) reconstruction, rather than parentage assignment, is required. For genetic markers showing Mendelian inheritance, such as microsatellites, parentage assignment (maternity or paternity) is computationally much simpler than sibship reconstruction. In diploid organisms, a parent and each offspring must share an allele at every genetic locus (barring rare mutations). On the other hand, full siblings will share, on average, half their alleles, but at any one locus, they may share 0,1, or 2 alleles. Sibling reconstruction methods have lagged behind those developed for paternity assignment, but several methods of sibling reconstruction are now available. In this review, we will examine the constraints that Mendelian inheritance dictates for sibling reconstructing, review the use of microsatellite genotyping in wild populations, and evaluate alternative genetic markers. We will then review the various methods for full sibling reconstruction that are currently available and present experimental validation of various methods using both real biological data and simulated data.

1.1. Microsatellites

While there are several molecular markers used in population genetics, microsatellites are the most commonly used in kinship studies in wild populations. First discovered in the late 1980s when genomic sequencing studies began [48, 54], microsatellites are short (one to six base pairs) simple sequence repeats, such as $(CA/GT)_n$ or $(AGC/TCG)_n$ that are scattered around eukaryotic genomes. A genomic library for a study species is screened for such repeats and primers for PCR amplification are constructed from the regions flanking the short repeats. Alternatively, microsatellite primers developed for one species may be used for closely related species. For example, microsatellites developed for humans amplify homologous loci in chimpanzees [12]. Figure 1 shows a schematic example of a microsatellite marker with three alleles and the resulting genotypes. Because there is a relatively high rate of mutation for adding or subtracting repeat units, microsatellite loci have high numbers of alleles and high levels of heterozygosity. PCR-based microsatellite analysis provides co-dominant, unlinked markers where alleles and genotypes can be scored precisely by size. These are the characteristics that make them especially useful for estimating kinship and relatedness. There are some technical problems associated with scoring microsatellites, and any method of sibling reconstruction with microsatellites needs to be able to accommodate a low frequency of scoring errors or artifacts, in addition to occasional mutation.

Microsatellites have been successfully applied to a wide range of non-model organisms, including vertebrates, invertebrates, plants, and fungi, and are used to infer large-scale population structure as well individual kinship. For kinship studies, microsatellites have been used more commonly for parentage than for sibship reconstruction, but there are an increasing number of studies that have attempted to reconstruct sibships with partial or no parental sampling. In lemon sharks, cohorts of juvenile sharks were sampled annually from nursery lagoons, and sibship reconstruction was used to infer the mating system and fertility of adults [17]. Sibship reconstruction was used to infer patterns of brood parasitism for individual female cowbirds, who lay their eggs in the nests of other birds [45, 46]. In a study of wood frogs, tadpoles were sampled from ponds and sibgroups reconstructed to study their spatial distribution and the potential for kin selection [22]. Such studies have employed a variety of methods to reconstruct sibling groups from microsatellite data because there was no widely accepted or easily implemented software available.

In addition to microsatellites, which assay DNA repeat variation, several PCR-based methods are available to assay variation in DNA sequence. RAPDs (randomly amplified polymorphic DNA), ISSRs (inter-simple sequence repeats), and AFLPs (amplified frag-

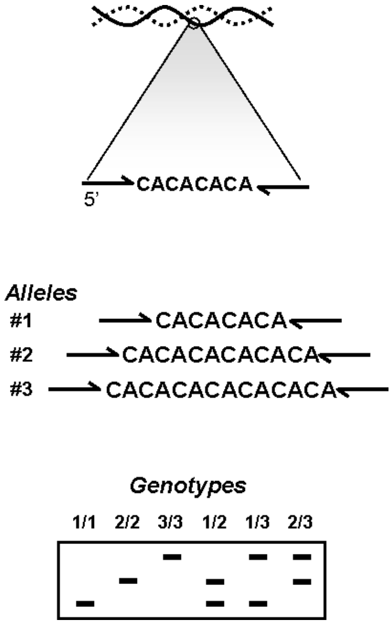

Figure 1. A schematic example of a microsatellite marker.

ment length polymorphisms) are dominant, multi-locus techniques which are problematic for kinship inference. SNPs (single nucleotide polymorphisms) are single locus markers that focus on a variable single nucleotide position in the genome. While they are numerous in the genome and, once identified, easy to score, they have limitations in the area of kinship reconstruction. The power to identify related individual depends mainly on the number of alleles per locus and their heterozygosity. SNPs are usually biallelic, whereas microsatellites may have 10 or more alleles per locus and typically have high heterozygosities. It appears for at least the next few years, microsatellites will remain the marker of choice for estimating relatedness in wild populations. We thus focus our efforts on developing and comparing methods of sibling reconstruction that are applicable to microsatellites or, more generally, codominant, multiallelic markers.

2. Sibling Reconstruction Problem

In order to reason about the inherent computational properties of the problem of reconstructing sibling relationships and to compare the accuracy and performance of various computational methods for solving the problem, we must define it formally. The problem of siblings reconstruction was first formally defined in [5] and is restated here.

Definition 1. Let U be a population of n diploid individuals of the same generation genotyped at at l microsatellite loci:

$$U \ = \{X_1, ...X_n\}, \qquad \text{where } X_i = (\langle a_{i1}, b_{i1}\rangle, ..., \langle a_{il}, b_{il}\rangle)$$

and a_{ij} and b_{ij} are the two alleles of the individual i at locus j represented as some identifying string. The goal of the *Sibling Reconstruction Problem* is to reconstruct the full sibling groups (groups of individuals with the same parents). We assume no knowledge of parental information. Formally, the goal is to find a partition of individuals $P_1, ...P_m$ such that

$$\forall 1 \leq k \leq m, \quad \forall X_p, X_q \in P_k : \quad Parents(X_p) = Parents(X_q)$$

Note, that we have not defined the function $Parents(X)$. This is a biological objective. Computational approaches use the formalization of various biological assumptions and constraints to achieve a good estimate of the biological sibling relationship. We describe the fundamental genetic properties that serve as a basis for most computational approaches in the next section.

3. Genetics of Sibship

3.1. Mendelian Genetics

Mendelian genetics lay down a very simple rule for gene inheritance in diploid organisms: *an offspring inherits one allele from each of its parents for each locus*. This introduces two overlapping necessary (but not sufficient) constraints on full sibling groups in absence of genotyping errors or mutations: the 4-allele property and the 2-allele property [5, 10].

4-Allele Property: The total number of distinct alleles occurring at any locus may not exceed 4.

Formally, a set of individuals $S \subseteq U$ has the 4-allele property if

$$\forall 1 \leq j \leq l : \quad \left| \bigcup_{i \in S} \{a_{ij}, b_{ij}\} \right| \leq 4.$$

Clearly, the 4-allele property is necessary since a group of siblings can inherit only combinations of the 4 alleles of their common parents. The 4-allele property is effective for identifying sibling groups where the data are mostly heterozygous and the parent individuals share few common alleles. Generally, as in Table 1, a set consisting of any two individuals trivially satisfies the 4-allele property. The set of individuals 1, 3 and 4 from Table 1 satisfies the 4-allele property. However, the set of individuals 2, 3 and 5 fails to satisfy it as there are five alleles occurring at the first locus: $\{12, 28, 56, 44, 51\}$.

2-Allele Property: There exists an assignment of individual alleles within a locus to maternal and paternal such that the number of distinct alleles assigned to each parent at this locus does not exceed 2.

Formally, a set of individuals $S \subseteq U$ has the 2-allele property if for each individual X_i in each locus there exists an assignment of $a_{ij} = c_{ij}$ or $b_{ij} = c_{ij}$ (and the other allele assigned to $\bar{c}_{ij}$) such that

$$\forall 1 \leq j \leq l : \quad \left| \bigcup_{i \in S} \{c_{ij}\} \right| \leq 2 \quad \text{and} \quad \left| \bigcup_{i \in S} \{\bar{c}_{ij}\} \right| \leq 2$$

The 2-allele property is clearly stricter than the 4-allele property. Looking at the Table 1, our previous 4-allele set of individuals 1, 3 and 4 fails to satisfy the 2-allele property since there are more than two alleles on the left side of locus 1: $\{44, 28, 13\}$. Moreover, there is no swapping of the left and right sides of alleles that will bring down the number of alleles on each side to two: individuals 1 and 4 with their alleles 44/44 and 13/13 already fill the capacity. Again, any two individuals trivially satisfy the 2-allele property.

Table 1. An example of input data for the sibling reconstruction problem. The five individuals have been sampled at two genetic loci. Each allele is represented by a number. Same numbers within a locus represent the same alleles.

Individual	Alleles $\langle a, b\rangle$ at locus 1	Alleles $\langle a, b\rangle$ at locus 2
1	44, 44	55, 27
2	12, 56	18, 39
3	28, 44	55, 18
4	13, 13	39, 27
5	28, 51	18, 39

Assuming the order of the parental alleles is always the same in the offspring (i.e. the maternal allele is always on the same side), the 2-allele property is equivalent to a biologically consistent full sibling relationship. The parental allele order, however, is not preserved, and an interesting problem arises: given a set of individuals S that satisfies the 4-allele property, does there exist a series of allele reorderings within some loci of individuals in S so that after those reorderings S satisfies the 2-allele property? For example, in Table 1, the individuals 1, 3, and 5 have more than two alleles on the right side of locus 2: $\{27, 18, 39\}$. However, switching the alleles 18 and 39 at locus 2 in the individual 5 will bring the number of alleles on either side down to two. Since the number of alleles on either side of locus 1 is also two, the set of individuals 1, 3, and 5 satisfies the 2-allele property.

In [10] we show the connection between the two properties that we restate here:

Theorem 1. Let a be the number of distinct alleles present in a given locus and R be the number of distinct alleles that either appear with three different alleles in this locus or are homozygous (appear with itself). Then, given a set of individuals with the 4-allele property, there exists a series of allele reorderings within loci resulting in a set that satisfies the 2-allele property if and only if for all the loci in the set

$$a + R \leq 4.$$

In our example of individuals 1, 3, and 5 in locus 1, $a = |\{44, 28, 51\}| = 3$ and $R = 1$ since each allele is paired up only with at most two different alleles but 44 is a homozygote. In locus 2, $a = |\{55, 27, 18, 39\}| = 4$ but $R = 0$ since there are no homozygote alleles and no allele appears with more than two different alleles. Thus, the set of individuals 1, 3, and 5 satisfies $a + R \leq 4$ for all loci and, hence, the 2-allele property.

The 2-allele property takes into account the fact that the parents can contribute only two alleles *each* to their offspring. Note, that the 2-allele property is, again, a necessary but not a sufficient constraint for a group of individuals to be siblings (in absence of errors or mutations). The full formalization of the Mendelian inheritance constraints in the context of sibling reconstruction is presented in [5, 10].

3.2. Relatedness Estimators

In the 1980's several statistical coefficients of relatedness were introduced [31, 33, 36]. All methods use observed allele frequencies to define the probabilistic degree of relatedness between two individuals. In 1999, Queller and Goodnight improved on their approach [37] by defining simple statistical likelihood formulae for different types of relationships and used those to infer sibling relationships. The 1999 paper also defines a method to determine the statistical significance, or "p-value", of a relationship estimate. This is done by randomly generating two individuals using the observed allele frequencies and the estimated probabilities of inheriting a shared allele as defined in the paper. Such random pairs of individuals are generated a large number of times, then the likelihood ratio that excludes 95% of the individuals is accepted as being at p-value 0.05. Even though this approach was not presented or aimed as a method for sibship reconstruction, it served as a basis for likelihood methods that followed. A number of assumptions are made by all relatedness estimators, including ignoring mutations and genotyping errors. More importantly, the methods assumes that a sample representative of the population has been scored, and there is accurate estimates of allele frequencies for the entire population. If these assumptions do not hold, results will be biased [34]. Finally, any method relying purely on a pairwise genetic distance may lead to inconsistent results, *i.e.* the transitivity of the sibling relationship may not hold. Moreover, as mentioned before, any pair of individuals can be siblings yet no pairwise distance estimate method cannot exclude that possibility [49].

4. Methods for Full Sibling Reconstruction

As more microsatellite markers become available for wild species there is a growing interest in the possibility of inferring relatedness among individuals when part or all of the pedigree information is lacking [43]. The majority of the available software requires parental data. However, recently there have been several methods attempting to reconstruct sibship groups from genetic data without parental information [1, 2, 6, 8, 29, 32, 43, 49, 53]. Fernandez and Toro [18] and Butler et al. [9] review many of the methods discussed here.

In their survey, Butler et al. [9] classified sibship reconstruction methods into two main groups: (1) methods that generate complete genealogical structures and, thus, require explicit pedigree reconstruction, and (2) pairwise methods that do not imply such pedigree reconstruction. This latter group can be subdivided into methods that estimate pairwise relatedness based on genotypic similarity and likelihood approaches that classify pairs of individuals into different types of relationships based on marker information.

In one of the earlier examples of the first type of method, Painter [32] used a Bayesian approach to calculate relationship likelihood and then an exhaustive search to find the most likely sibship in a small population of 9 individuals. He identified the need for using better optimization techniques for larger populations. Among the methods that followed, some use Markov-Chain Monte Carlo (MCMC) techniques to locate a partition of individuals that maximizes the likelihood of the proposed family relationship, such as COLONY [53] software and Almudevar's method [1]. Smith [43] has developed an approach that maximizes a relatedness configuration score derived form the pairwise relatedness likelihood ratio. Almudevar and Field [2] used an exclusion principle that looks for the largest full-sibling families, using partial likelihoods to pick between families of the same size. Another approach is based on Simpson's index of concentration [9], where groups that conform to Mendelian inheritance rules are formed according to marker information. One of the advantages of these methods is that they avoid the inconsistency problems of pairwise estimators described below. However, the statistical likelihood methods still depend on the knowledge of population allelic data (to calculate likelihoods) which is typically unavailable or inaccurate. Moreover, since most of these methods employ global optimization at their core, they are usually computationally demanding.

As described above, a second type of approach, pairwise methods, are widely used for sibship reconstruction. While these methods are typically simple and fast they suffer several disadvantages. First, they can lead to incongruous assignments because only two individuals are considered at a time and transitivity is not preserved. Second, like all statistical methods, they are dependent on the knowledge of allelic frequencies of the population considered. Third, if multiple definite relationships exist, such as full siblings, half siblings, or unrelated, arbitrary thresholds have to be defined to decide the category to which a particular pair is assigned [18].

Here, we consider a different classification of sibling reconstruction methods, based on the computational approach a method employs as the basis for reconstruction. SIBSHIP [49], Pedigree [43], KINGROUP [29], and COLONY [53] rely on statistical estimates of relatedness [37] and reconstruct the maximum likelihood sibling groups. Family Finder [8] and Almudevar [1] mix statistical and combinatorial approaches. Finally, Almudevar and Field [2], 2-allele Minimum Set Cover [5, 6, 10, 41] and Sheikh et al. [40] use only the fundamental Mendelian constraints and combinatorial techniques to reconstruct sibling groups.

A common assumption of all but two (Sheikh et al. [40] and COLONY [53]) of the sibship reconstruction methods is that the molecular data is error and mutation free [18]. Data that contain errors test the robustness of these methods and are a major problem of the estimators involving pedigree reconstruction [9].

Following our computationally based classification, we now describe some of the methods in more detail, providing deeper analysis of the two best-performing methods (see Section 5. for experimental comparison), the likelihood based COLONY and the combinatorial 2-allele Minimum Set Cover.

4.1. Statistical Likelihood Methods

As Painter's [32] first likelihood-based sibling reconstruction method exemplified, likelihood maximization methods require sophisticated optimization techniques to find the most likely sibship partition for datasets of size greater than 10 individuals.

In 2000, Thomas and Hill [49] introduced a Markov Chain Monte Carlo (MCMC) approach to find the maximum likelihood of a sibship reconstruction. The method compares the likelihood ratio of two individuals being siblings to that of the the pair being unrelated [36]. Starting with a random partition of individuals into potential sibling groups, the method uses a "hill-climbing" approach to explore different sibship reconstructions, reassigning individuals into sibling groups to improve the likelihood of all pairs being siblings. The process continues until one of the halting conditions is reached: either the number of iterations exceeds a threshold, or the sibling reconstruction stabilizes, *i.e.* the likelihood value reaches a fixed point. The algorithm was not computationally efficient and was subsequently improved. Like most likelihood based methods, the main assumption of the approach is that the sample at hand is representative of the entire population in terms of allele frequencies and, thus, the relatedness probabilities. More detrimentally, the method also assumes that the population contains only full siblings and unrelated individuals which typically does not hold for any population.

In 2002, Thomas and Hill [50] extended their approach by adding half sibling relationships, thus creating a limited family hierarchy. The algorithm is similar to their previous approach in [49], with the addition that an individual could be assigned to either a half sibling group or a full sibling group at every iteration. Half sibling groups were randomly created every few hundred iterations to ensure that a hierarchical structure existed in the population. In that paper, Thomas and Hill also explored the effects of population size, population structure, and the allelic information available on the performance of their MCMC approach. Typical of the statistical approaches, the accuracy of the reconstruction improved with the increase of available marker information and the nestedness of the full siblings within half sibling groups but decayed with the increase of the population size.

In 2001 Smith et al. [43] presented two different MCMC methods for sibship reconstruction. One of the methods is very similar to [50], while the other aims to maximize the joint likelihood of the entire sibship reconstruction rather than pairwise relatedness ratio. The methods performed very well for the Atlantic salmon dataset the authors used in the original publication. The software PEDIGREE is now available for general use as an online service. Smith *et al.* have also assayed the dependency of the accuracy of reconstruction various data parameters. In general, the methods suffer from typical assumptions of other statistical methods. The accuracy of reconstruction decreases when there is insufficient allelic diversity per locus or the sample is not representative of the population.

Konovalov et al. [29] introduced KINGROUP, available as an open source Java™ program. KINGROUP uses the relatedness estimators of [37] with additional algorithms designed for the reconstructions of groups of kin that share a common relationship.

Family Finder [8] was introduced in 2003. It is a very efficient method that uses a combination of statistics and graph theory. This approach constructs a graph with individuals as vertices. Edges represent pairwise sibling relationship and are weighted using, again, the likelihood ratio of individuals being siblings to their being unrelated [37]. After constructing this graph "clusters", or components, corresponding sibling groups are identified by finding light edge cuts. Cuts with the number of edges less than one third of the edges in the graph are chosen. It is a simple and efficient method that can be effective if enough loci are available and allelic diversity is high. While there is some theoretical basis, usage of the likelihood ratio implies the same assumptions as [37]. Furthermore, it assumes that sibling groups are roughly equally sized, which is a dubious assumption and often does not hold, especially for wild population samples.

4.1.1. COLONY

A different likelihood maximization approach was used by Wang [53]. COLONY is a comprehensive statistical approach that uses the simulated annealing heuristic to find a (local) likelihood maximum of a sibship reconstruction. The algorithm starts with known full and half siblings (if any are available) and places the rest into singleton sibling groups, along with the computed likelihood of each group. A proposed alternate solution at every iteration is created by moving a random number of individuals from one full sibling group to another (both groups must not be one of the known full sibling groups). For half siblings, a random number of entire full sibling groups are moved from one half sibling group to another. As before, these must not be the original known half sibling families. After generating a new proposed solution, the likelihood of the old and new configurations of the altered families is calculated. The new configuration is accepted or rejected based on a threshold which depends on the ratio of the new and old likelihoods.

COLONY is the first method to fully accommodate sampling bias and genotyping errors, although it relies on many user input parameters to do so. Errors are estimated using the calculated probability of observing the given allele assuming the actual allele is different. The probabilities of allelic dropouts and other typing of errors are based on [19], allelic dropout is considered to be twice as likely as other errors.

Simulated annealing relies on random numbers and explores a vast solution space. COLONY can be quite slow, and its performance both in terms of time and accuracy depends drastically on the amount of microsatellite information available. COLONY was designed for both diploid and haplodiploid species. It is perhaps the most comprehensive and sophisticated method currently available for full sibling reconstruction with a strong theoretical basis. However, in addition to other disadvantages common to all statistical sibship reconstruction methods, it also assumes that one of the parents is monogamous which, unfortunately, renders it inappropriate for many species that have promiscuous mating systems.

4.2. Combinatorial Approaches

Combinatorial approaches to sibling reconstruction use Mendelian constraints to eliminate sibling groups that are infeasible and to form potential sibling groups that conform to these constraints. Various methods then use different objectives to choose from among these the groups to form the solution.

Almudevar and Field [2] were the first to introduce a combinatorial approach. They formulated the Mendelian properties in form of graphs and constructed all maximal feasible sibling groups. They then performed an exhaustive search to select the minimal number of these groups using maximum likelihood of the reconstruction as the guide. Their approach yielded reasonably good results but was computationally very expensive, often resulting in the system running out of memory in our experiments (see Section 5.). Almudevar presented a “hybrid” approach in [1] that used simulated annealing based on MCMC methods to find a locally optimal solution. The method generates putative triplets of parents and children, and then uses simulated annealing to explore the space of different possible pedigrees. The exploration is similar to the approach taken by COLONY described above and uses the likelihood of the sibling group configuration as a guide. Such a heuristic approach is not guaranteed to find a globally minimum number of sets. This new version of the method allows for the use of other information in the reconstruction, such as multiple generations of siblings, parental genotypes and sex where available. All the information is translated into constraints that guide the formation of the potential feasible solution.

4.2.1. 2-Allele Minimum Set Cover

The 2-Allele Minimum Set cover approach [5, 6, 10, 41], like Almudevar and Field's, uses Mendelian constraints, specifically the 2-allele property, to form all maximal feasible sibling groups. The goal, then, is to find the smallest number of these that contain all individuals. Unlike Almudevar and Field, this approach finds the true global, rather than local, minimum. We describe the technical details of the approach and the computational complexity of this formulation of the problem below.

Recall that we are given a population U of n diploid individuals sampled at l loci

$$U \ = \{X_1, ...X_n\}, \text{ where } X_i = (\langle a_{i1}, b_{i1}\rangle, ..., \langle a_{il}, b_{il}\rangle)$$

and a_{ij} and b_{ij} are the two alleles of the individual i at locus j.

The goal of the Minimum 2-Allele Set Cover problem is to find the smallest number of subsets $S_1, ..., S_m$ such that each $S_i \subseteq U$ and satisfies the 2-allele constraint and $\bigcup S_i = U$. We shall denote the Minimum 2-Allele Set Cover on n individuals with l sampled loci as 2-ALLELE $_{n,\ell}$.

Of all the sibling reconstruction problem formulations, this is the only one for which its computational complexity is known.

Computational Complexity

The Minimum 2-Allele Set Cover problem is a special case of the MINIMUM SET COVER problem, a classical NP-complete problem [28]. MINIMUM SET COVER is defined as follows: given a universe U of elements $X_1, ..., X_n$ and a collection of subsets $\mathcal{S}$ of U, the goal is to find the minimum collection of subsets $C \subseteq \mathcal{S}$ whose union is the entire universe U.

Recall, that a $(1+\varepsilon)$-*approximate solution* (or simply an $(1+\varepsilon)$-approximation) of a minimization problem is a solution with an objective value no larger than $1+\varepsilon$ times the value of the optimum, and an algorithm achieving such a solution is said to have an *approximation ratio* of at most $1+\varepsilon$. To say that a problem is r-inapproximable under a certain complexity-theoretic assumption means that the problem does not have a r-approximation unless that complexity-theoretic assumption is false.

MINIMUM SET COVER cannot be approximated in polynomial time to within a factor of $(1-\epsilon)\ln n$ unless $NP \subseteq DTIME(n^{loglogn})$ [16]. Johnson introduced a $1 + \ln n$ approximation in 1974 [27].

In the 2-ALLELE $_{n,\ell}$ the problem the elements are the sampled individuals and the sets $\mathcal{S}$ are the groups of individuals that satisfy the 2-allele property. The main difference between MINIMUM SET COVER and 2-ALLELE $_{n,\ell}$, or more generally k-ALLELE $_{n,\ell}$ problem for $k \in \{2, 4\}$, is that the latter add the 2-allele or the 4-allele restriction on the structure of the subsets $\mathcal{S}$. We show that this restriction does not make the problem computationally easier and k-ALLELE $_{n,\ell}$ remains NP-complete.

A natural parameter of interest in this class of problems is the maximum size (number of elements) a in any set in $\mathcal{S}$. We denote the corresponding problem of finding the minimum set cover when the size of sibling sets is at most a as a-k-ALLELE$_{n,\ell}$ in the subsequent discussions. For example, 2-4-ALLELE $_{n,\ell}$ and 2-2-ALLELE $_{n,\ell}$ are the problem instances where each subset contains at most two individuals. Recall, that any pair of individuals necessarily satisfies both the 2-allele and the 4-allele properties. Thus, the collection $\mathcal{S}$ for 2-k-ALLELE $_{n,\ell}$ consists of all possible pairs of individuals and the smallest number of subsets that contain all the individuals are any $n/2$ disjoint pairs.

In general, if a is a constant, then a-k-ALLELE $_{n,\ell}$ can be posed as a minimum set cover problem with the number of subsets polynomial in n and the maximum set size being a. This problem has a natural $(1 + \ln a)$-approximation using the standard approximation algorithms for the minimum set cover problem [51]. For a general a, the same algorithm guarantees a $\left(\frac{a}{c} + \ln c\right)$-approximation for any constant $c > 0$. Recently, Ashley et al. [4] have been able to obtain several non-trivial computational complexity results for these problems which we restate here.

For the smallest non-trivial value of $a = 3$, the 3-k-ALLELE $_{n,n^3}$ problem is 1.0065-inapproximable unless $RP = NP$. This was proved by a reduction from the TRIANGLE PACKING problem [20, p. 192]. A $\left(\frac{7}{6} + \varepsilon\right)$-approximation for any $\ell > 0$ and any constant $\varepsilon > 0$ is easily achieved using the results of Hurkens and Schrijver [25].

For the second smallest value of $a = 4$ and $l = 2$, 4-k-ALLELE $_{n,2}$ is 1.00014-inapproximable unless $RP \neq NP$, proved by a reduction from the MAX-CUT problem on cubic graphs via an intermediate novel mapping of a geometric nature. The $\left(\frac{3}{2} + \epsilon\right)$-approximation can be achieved for $a = 3$ by using the result of Berman and Krysta [7].

The n^{ϵ}-inapproximability result under the assumption of ZPP$\neq$NP was proved for all sufficiently large values of a, that is $a = n^{\delta}$, where ϵ is any constant strictly less than δ. This result was obtained by reducing a suitable hard instance of the graph coloring problem.

In all the reductions above additional loci play an important role of adding complexity to the problem to ensure the inapproximability result. Thus, interestingly and somewhat counterintuitively, while sampling more loci provides more information and typically improves the accuracy of most sibling reconstruction methods, it also adds computational complexity and increases the computational time needed to construct the solution, even beyond the scope of practical computability.

The Algorithm

In [6] we have presented a fully combinatorial solution for the siblings reconstruction problem based on the 2-Allele Minimum Cover formulation. We briefly describe the 2-ALLELE COVER algorithm here. The algorithm works by first generating all maximal sibling groups that obey the 2-allele property and then finds the optimal minimum number of sibling groups necessary to explain the data. The algorithm maintains a complete enumeration of canonical possible sibling groups, called the possibilities table, shown in Table 2. Each potential sibling group is mapped to a set of possible canonical representations. Genetic feasibility of membership of each new individual in a sibling group is checked using this mapping. The intricate process of generating the maximal feasible 2-allele sets is described in detail in [6].

The 2-allele property reduces the possible combinations of alleles at a locus in a group of siblings down to a few canonical options, assuming that the alleles in the group are renumbered 1 through 4. Table 2 lists all different types of sibling groups possible with the 2-allele property using such a numbering. We do this by listing all possible pairs of parents whose alleles are among 1,2,3, and 4 and all the genetically different offspring they can produce. However, in any sibling group with a given set of parents only a subset of the offspring possibilities from the table may be present.

Table 2. Canonical possible combinations of parent alleles and all resulting offspring allele combinations

Parents	Offspring	
	allele a	allele b
(1, 2) and (3, 4)	1	3
	1	4
	2	3
	2	4
	3	1
	4	1
	3	2
	4	2
(1, 2) and (1, 3)	1	1
	1	3
	2	1
	2	3
	3	1
	1	2
	3	2
(1, 2) and (1, 2)	1	1
	1	2
	2	1
	2	2
(1, 1) and (1, 1)	1	1
(1, 1) and (1, 2)	1	1
	1	2
	2	1
(1, 1) and (2, 3)	1	2
	1	3
	2	1
	3	1
(1, 1) and (2, 2)	1	2
	2	1

The maximal feasible 2-allele sets are generated using the canonical possibilities in Table 2 in a way which provably produces *all maximal* such sets and does it in provably *fewest* number of queries per individual. After that, the minimum set cover is constructed as the solution to the sibling reconstruction problem. Note, that since 2-allele minimum cover and Minimum Set Cover are both NP-complete problems, the solution time is not guaranteed to be polynomial. We use the commercial mixed integer linear program solver CPLEX[1] to solve the problem to optimality. On datasets with several hundreds individuals it may take several hours to days to obtain a solution.

Subsequently, Chaovalitwongse et al. [10] have presented a full mathematical optimization formulation for the Minimum 2-allele Cover problem. We shall briefly describe the 2-ALLELE OPTIMIZATION MODEL (2AOM) here. The formulation directly models the objective of finding the minimum number of 2-allele sets that contain all individuals, rather than using the intermediate steps of generating all maximal 2-allele sets and finding the minimum set cover of those.

Individual	Locus 1 alleles $\langle a, b\rangle$	Locus 2 alleles $\langle a, b\rangle$	...
1	44, 44	55, 27	
2	12, 56	18, 39	
3	28, 44	55, 18	
4	13, 13	39, 27	
5	28, 51	18, 39	
⋮			

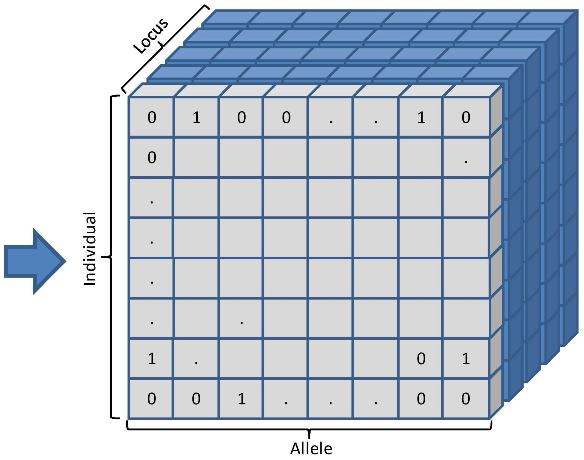

Figure 2. A multidimensional matrix representation of a dataset of microsatellite samples.

Recall, that U is the set of individuals, $\mathcal{S}$ is a set of sibling groups, and $\mathcal{C} \in \mathcal{S}$ is the reconstructed set of sibling groups which is returned as the solution. Let K be the set of possible observed alleles and L be the set of sampled loci. As the input, we are given $|U| = n$ individuals sampled at $|L| = l$ loci. We represent the data as a multidimensional 0-1 matrix M shown in Figure 2. The matrix entry $M(i, k, l) = 1$ if the individual $i \in U$ has the allele $k \in K$ in locus $l \in L$.

From the input matrix, a^l_{ik} is defined as an indicator variable and equals to 1 if the first allele at locus l of individual i is k. Similarly, b^l_{ik} is an indicator variable for the second allele at locus l of individual i is k. $f^l_{ik} = \max\{a^l_{ik} + b^l_{ik}\}$ is an indicator of whether k appears at locus l of individual i, that is, $M(i, k, l) = f^l_{ik}$. Finally $h^l_{ik} = a^l_{ik} \cdot b^l_{ik}$ is an indicator of whether the individual i is homozygous (allele k appears twice) at locus l. The following decision variables are then used:

- $z_s \in \{0, 1\}$: indicates whether any individual is selected to be a member of sibling group s;

- $x_{is} \in \{0, 1\}$: indicates whether the individual i is selected to be a member of sibling group s;

[1] CPLEX is a registered trademark of ILOG

- $y^l_{sk} \in \{0,1\}$: indicates whether any member of sibling group s has the allele k at locus l;
- $w^l_{sk} \in \{0,1\}$: indicates whether there is at least one homozygous individual in sibling group s with the allele k appearing twice at locus l;
- $v^l_{skk'} \in \{0,1\}$: indicates whether the allele k appears with allele k' in sibling group s at locus l.

With these variable, the mathematical representation of the objective function and the constraints of the 2AOM problem are as follows.

Objective function: The overall objective function is to minimize the total number of sibling groups:

$$\min \sum_{\forall s \in \mathcal{S}} z_s$$

The minimization objective is subject to three types of constraints stated below.

Cover and logical constraints: Ensure that every individual is assigned to at least one sibling group:

$$\sum_{\forall s \in \mathcal{S}} x_{is} \geq 1, \quad \forall i \in U$$

The binary sibling group variable s is activated for the assignment of any individual i to the sibling group s:

$$x_{is} \leq z_s, \quad \forall i \in U, \forall s \in \mathcal{S}$$

2-allele constraints: Activate the binary indicator variable for alleles y^l_{sk} with the assignment of any individual i to the sibling set s. Here C_1 is a large constant which can be defined as $C_1 = 2|U| + 1$:

$$\sum_{\forall i \in U} f^l_{ik} x_{is} \leq C_1 y^l_{sk}, \quad \forall s \in \mathcal{S}, \forall k \in K, \forall l \in L$$

Activate the binary indicator variables for homozygous individuals with allele k appearing twice at locus l in sibling group s. Here C_2 is a large constant which can be defined as $C_2 = |U| + 1$:

$$\sum_{\forall i \in U} h^l_{ik} x_{is} \leq C_2 w^l_{sk}, \quad \forall s \in \mathcal{S}, \forall k \in K, \forall l \in L$$

Activate the binary indicator variable for allele pair $v^l_{skk'}$ for any assignment to the sibling group s of the individual i with alleles $\langle k, k' \rangle$ at locus l. Here C_3 is a large constant and can be defined as $C_3 = |U| + 1$:

$$\sum_{\forall i \in U} f^l_{ik} h^l_{ik} x_{is} \leq C_3 v^l_{skk'}, \quad \forall s \in \mathcal{S}, \forall k \neq k' \in K, \forall l \in L$$

Ensures that the number of distinct alleles plus the number of homozygous alleles does not exceed 4, conforming to Theorem 1:

$$\sum_{\forall k \in K} y^l_{sk} + w^l_{sk} \leq 4, \quad \forall s \in \mathcal{S}, \forall l \in L$$

Every allele in the set should not appear with more than two other alleles (excluding itself), also conforming to Theorem 1:

$$\sum_{\forall k' \in K \setminus k} v^l_{skk'} \leq 2, \quad \forall s \in \mathcal{S}, \forall k \in K, \forall l \in L$$

Binary and nonnegativity constraints:

$$z_s, x_{is}, y^l_{sk}, w^l_{sk} \in \{0, 1\}, \quad \forall i \in U, \forall s \in \mathcal{S}, \forall k \in K, \forall l \in L$$

The total number of discrete variables in the 2AOM is $O(|U||K||\mathcal{S}|)$ and so is the total number of constraints. Thus, the 2AOM formulation of the 2-allele minimum cover problem is a very large-scale mixed integer program problem and may not be easy to solve in large instances. The main justification for a formal mathematical model of the problem is that it allows for the theoretical investigation of its computational properties and guides approximation approaches.

4.3. Consensus-based Approach

Among all the methods for sibling reconstruction, only COLONY [53] is designed to tolerate genotyping errors or mutation. Yet, both errors and mutations cannot be avoided in practice and identifying these errors without any prior kinship information is a challenging task. A new approach for reconstructing sibling relationships from microsatellite data designed explicitly to tolerate genotyping errors and mutations in data based on the idea of a consensus of several partial solutions was proposed by Sheikh et al. in [40, 42]

Consider an individual X_i which has some genotyping error(s). Any error that is affecting sibling reconstruction must be preventing X_i's sibling relationship with at least one other individual X_j, who in reality is its 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. 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 core of the consensus-based 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.

Definition 2. A *consensus* method for the sibling reconstruction problem is a computable function f that takes k solutions $\mathcal{S} = \{S_1, ..., S_k\}$ as input and computes *one* final solution.

The *strict consensus* places two individuals into a sibling groups only if they are together in all input solutions. While it always results in a consistent solution, it also produces many singleton sibling groups. In [40, 42] a *distance based consensus* for sibling reconstruction was introduced. Starting with a strict consensus of the input solutions, distance based consensus iteratively merges two sets until the quality of the solution cannot be improved. The computational complexity and the algorithms change depending on the cost of the merging operations and the function that defines the quality of the solution. The approach taken in [40, 42] uses the number of the sibling groups in the resulting solution as the measure of the quality of the solution, that is, it seeks to minimize the number of groups. The cost of the merging operation is based on the size of the groups being merged and errors that need to be corrected for the 2-allele property to be preserved in the combined group.

Any method or a mix of methods for sibling reconstruction can be used as the base to produce the input solution for the consensus method. The running time of the consensus method depends on the running times of the base methods. In our experiments (see Section 5.) consensus based on 2-allele minimum cover algorithm typically achieved over 95% accuracy.

5. Experimental Validation

To assess and compare the accuracy of various sibling reconstruction methods we used datasets with known genetics and genealogy. Since most sibling reconstruction methods do not tolerate errors in data, we first used error free datasets. However, biological datasets containing no errors are rare. Thus, in addition to biological datasets, we created simulated sets using a large number of parameters over a wide range of values. We compare the performance of five sibling reconstruction methods, spanning the variety of computational techniques: Almudevar and Field [2], Family Finder [8], KINGROUP [29], COLONY [53], and 2-allele Minimum Cover [6].

In addition, we used the same datasets with introduced errors to assess the performance of COLONY and the distance-based consensus of the 2-allele Minimum Cover when errors are present.

We measure the error by comparing the known sibling sets with those generated by various sibling reconstruction methods, and calculating the minimum partition distance [21]. The error is the percentage of individuals that would need to be removed to make the reconstructed sibling sets equal to the true sibling sets. Note, we are computing the error in terms of individuals, not in terms of the number of sibling groups reconstructed incorrectly. Thus, the accuracy is the percent of individuals correctly assigned to sibling groups.

The experiments were run on a combination of a cluster of 64 mixed AMD and Intel Xeon nodes of 2.8 GHz and 3.0GHz processors and a single Intel Xeon Quad Core 3.2 GHz Intel processor with 24 GB RAM memory.

5.1. Biological Datasets

For validation of our methods, both the 2-allele and the consensus extension, we used biological datasets of offspring that resulted from one generation of controlled crosses, thus the identity of the parents and their microsatellite genotypes were known.

Radishes. The wild radish *Raphanus raphanistrum* dataset is a subsample of [11]. It consists of samples from 64 radishes from two families with 11 sampled loci. Close to 53% of allele entries are missing.

Salmon. The Atlantic salmon *Salmo salar* dataset comes from the genetic improvement program of the Atlantic Salmon Federation [23]. We use a truncated sample of 351 individuals from 6 families and 4 loci. There are no missing alleles at any locus. This dataset is a subset of one of the samples of genotyped individuals used by [2] to illustrate their technique.

Shrimp. The tiger shrimp *Penaeus monodon* dataset [26] consists of 59 individuals from 13 families with 7 loci. There are 16 missing allele entries (3.87% of all allele entries).

Flies. *Scaptodrosophila hibisci* dataset [56] consists of 190 same generation individuals (flies) from 6 families sampled at various number of loci with up to 8 alleles per locus. All individuals shared at least 2 sampled loci which were chosen for our study. 25% of allele entries were missing.

Table 3 summarizes the results of the four algorithms on the biological datasets.

Table 3. Accuracy (percent) of the 2-allele algorithm and the three reference algorithms on biological datasets. Here l is the number of loci in a dataset and "Inds" column gives the number of individuals in the dataset. The three reference algorithms are [2] (A&F), Family Finder by [8] (B&M), and the KINGROUP by [29] (KG).

Dataset	l	Inds	**Ours**	A&F	B&M	KG
Shrimp	7	59	77.97	67.80	77.97	77.97
Salmon	4	351	98.30	Out of memory	99.71	96.02
Radishes	5	64	75.90	Out of memory	53.30	29.95
Flies	2	190	100.00	31.05	27.89	54.73

Almudevar and Field's algorithm ran out of 4GB memory on the salmon and radish datasets.

5.2. Synthetic Datasets

To test and compare sibling reconstruction approaches, we also use random simulations to produce synthetic datasets. We first create random diploid parents and then generate complete genetic data for offspring varying the number of males, females, alleles, loci, number of families and number of offspring per family. We then use the 2-allele algorithm described above 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 of algorithm using the Gusfield Partition Distance [21]. In addition, we compare the accuracy of our 2-allele algorithm to the two reference sibling reconstruction methods, [8] and [29], described above. We repeat the entire process for each fixed combination of parameter values 1000 times. We omit the comparison of the results to the algorithm of [2] since the current version of the provided software requires user interaction and therefore it is infeasible to use it in the automated simulation pipeline of 1000 iterations of over a hundred combinations of parameter values.

First, we generate the parent generation of M males and F females with parents with l loci and a specified number of alleles per locus a. We create populations with uniform as well as non-uniform allele distributions. After the parents are created, their offsprings are generated by selecting f pairs of parents. A male and a female are chosen independently, uniformly at random from the parent population. For these parents a specified number of offsprings o is generated. Here, too, we create populations with a uniform as well as a skewed family size distribution. Each offspring randomly receives one allele each from its mother and father at each locus. This is a rather simplistic approach, however, it's consistent with the genetics of known parents and provides a baseline for the accuracy of the algorithm since biological data are generally not random and uniform.

The parameter ranges for the study are as follows:

- The number of adult females F and the number of adult males M were equal and set to 5, 10 or 15.
- The number of loci sampled $l = 2, 4, 6$
- The number of alleles per locus (for the uniform allele frequency distribution) $a = 5, 10, 15$.
- Non-uniform allele frequency distribution (for 4 alleles): 12 - 4 - 1 - 1, as in [1].
- The number of families in the population $f = 2, 5, 10$.
- The number of offspring per mating pair (for the uniform family size distribution) $o = 2, 5, 10$.
- Non-uniform family size distribution (for 5 families): 25 - 10 - 10 - 4 - 1, as in [1]

All datasets were generated on the 64-node cluster running RedHat Linux 9.0. The 2-allele algorithm is used on this generated population to find the smallest number of 2-allele sets necessary to explain this offspring population. We use the commercial MIP solver CPLEX 9.0 for Windows XP on a single processor machine to solve the minimum set cover problem to optimality. The reference algorithms were run on a single processor machine running Windows XP[2].

We measure the reconstruction accuracy of various methods as the function of the number of alleles per each locus, family size (number of offspring), number of families (and polygamy), and the variation in allele frequency and family size distributions.

Figure 3 shows representative results for the accuracy of our 2-allele algorithm and the two reference algorithms on uniform allele frequency and family sizes distributions. Figure 4 shows results for the datasets with skewed family sizes and allele frequency distributions.

[2] The difference in platforms and operating systems is dictated by the available software licenses and provided binary code

Each bar represents the mean value of a 1000 random repetitions and the error bars show the standard deviation.

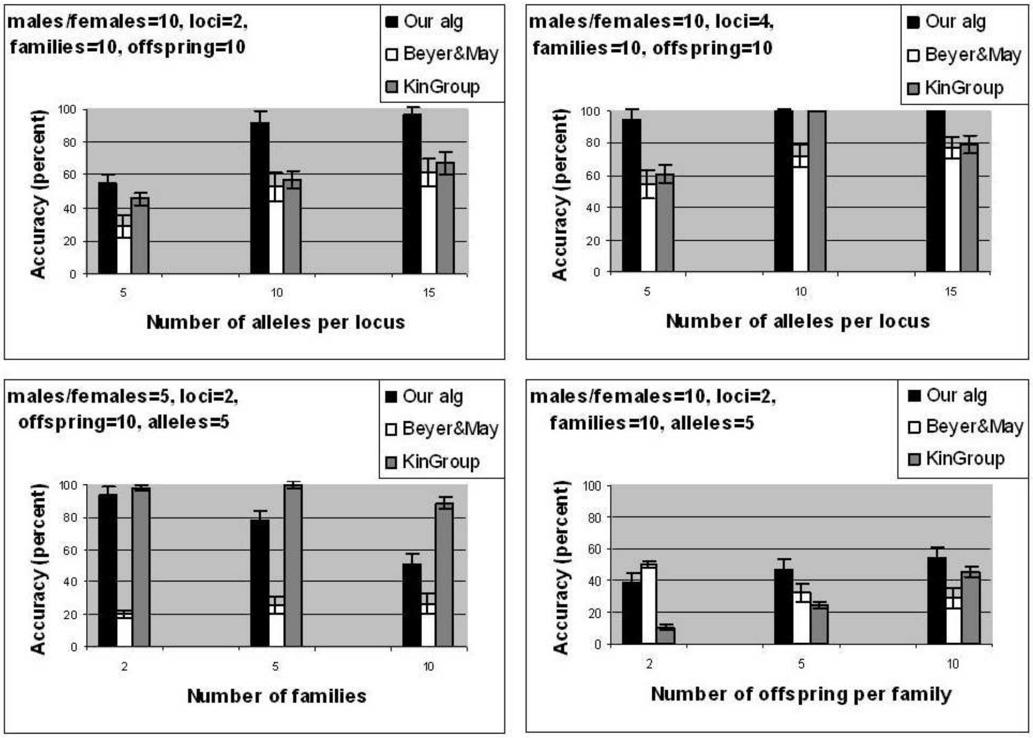

Figure 3. Accuracy of the sibling group reconstruction methods on randomly generated data. The y-axis shows the accuracy of reconstruction as a function of various simulation parameters. The accuracy of our 2-allele algorithm is shown, as well as that of the two reference algorithms: [8] and [29] (KINGROUP). The title shows the value of the fixed parameters: the number of adult males/females, number of families, the number of offspring per family, the number of loci, and the number of alleles per locus.

The results of COLONY and the consensus based 2-allele Minimum Cover on simulated datasets with introduced errors are shown in Figure 5.

Overall, we have compared our 2-allele algorithm as well as the robust consensus approach to the best existing sibling reconstruction methods on biological and synthetic data over a wide range of parameters. We have identified the strengths and weaknesses of various approaches to sibling reconstruction and pinpointed the data parameters under which those are manifested.

6. Conclusion

Full utilization of new genetic tools provided by advances in DNA and genome analysis will only be realized if computational approaches to exploit the genetic information keep pace.

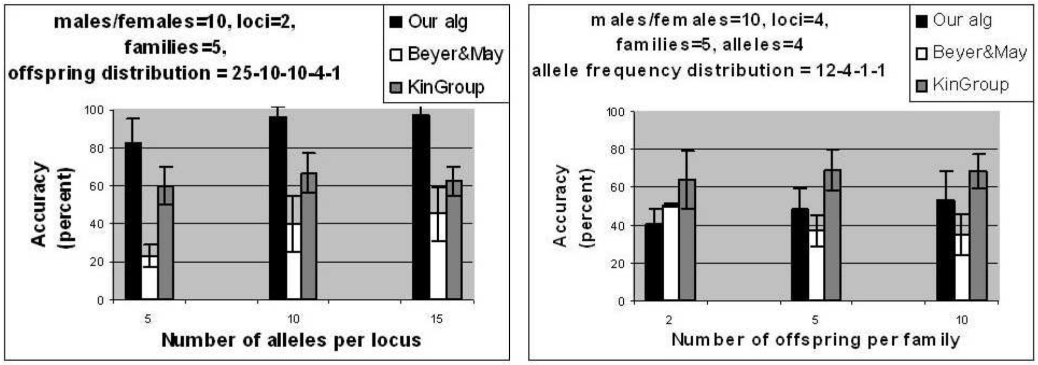

Figure 4. Accuracy of the sibling group reconstruction using our 2-allele algorithm and the two reference methods on the datasets with skewed family sizes and allele frequency distributions.

Pedigree reconstruction in wild populations is an emerging field, made possible by the development of markers, particularly DNA microsatellites, that can be used to genotype any organism, including free-living populations sampled in the field. Rules of Mendelian inheritance and principles of population genetics can be applied to microsatellite genotyping data to infer familial relationships such as parentage and sibship, and thus reconstruct wild pedigrees. Such pedigrees, in turn, can be used to learn about a species' evolutionary potential, their mating systems and reproductive patterns, dispersal and inbreeding (reviewed in [35]). The findings of pedigree reconstruction have been especially notable in the area of paternity assignment, where dozens of examples of previously undocumented multiple paternity have now been reported (*e.g.* [15, 17, 44, 52]).

Our focus has been on a more challenging computational problem than paternity (or parentage) assignment, that of sibling reconstruction. Sibling reconstruction is needed when wild samples consist primarily of offspring cohorts, in cases where it is logistically difficult or impossible to sample the parental generation. We first develop a formal definition of the sibling reconstruction problem and formalize the genetics of sibship. Sibling reconstruction methods can be divided into three categories depending on their approach, methods that rely only statistical estimates of relatedness [29, 32, 43, 49, 50, 53], those that combine statistical and combinatorial approaches [8], and those that use only Mendelian constraints and combinatorial techniques [1, 2, 5, 6, 10, 41]. Statistical methods rely on estimates of pairwise relatedness and typically reconstruct maximum likelihood sibling groups. The performance of statistical methods depends upon an accurate estimate of underlying allele frequencies *within the sampled populations*, rather than the observed sample. Furthermore, they are often computationally demanding. Combinatorial approaches offer the advantage that sibling groupings are based only on Mendelian constraints without needing information on population allele frequencies. A new method we describe here, the 2-allele minimum set cover, generates all sibling groups that obey the 2-allele property and then finds the optimal minimum number of sibling groups needed to explain the data. To accommodate genotyping errors and mutations, we also describe a new consensus-based approach applied here to the 2-allele minimum cover algorithm.

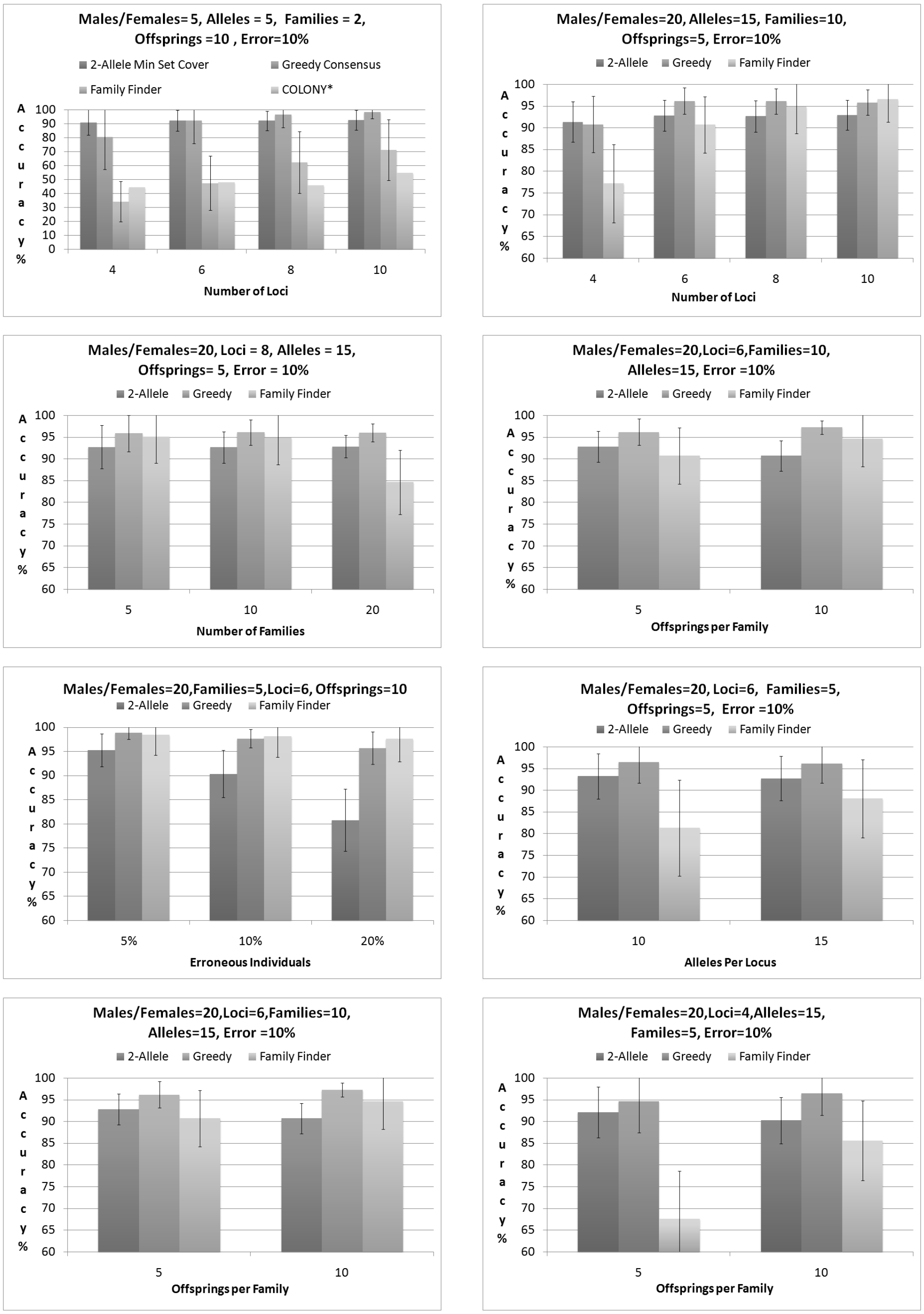

Figure 5. Results on simulated datasets with errors. Only 50 iterations were used for the COLONY algorithm due to its computational inefficiency and time constraints.

We tested the performance of various sibling reconstruction methods using both real biological data and synthetic data sets. For real data, the actual pedigree and sibgroups were known from controlled crosses, and we tested the accuracy of five different methods in recovering the known sibgroups. We found that our 2-allele distance-based consensus method performed very well, recovering over 95% of the known sibgroups. We also produced synthetic datasets which simulated a variety of mating systems, family structures, and genetic data. Again, our method produced very good results. Of the other methods tested, COLONY [53], a statistical approach, also performed very well when the assumptions of monogamy held and there were a sufficient number of loci and accurate estimates of allele frequencies.

There is no one method that is guaranteed to provide the correct answer, since samples of different populations suffer from different sampling biases and all methods make assumptions that may not hold for a specific dataset. We favor the 2-allele method for this very reason: it makes the fewest assumptions. Also, the 2-allele algorithm overall performs well over a wide range of data parameters, thus making it a good general method, especially when few loci are sampled or the allelic variation is low. Our current recommendation is to use the proposed concensus approach on the 2-allele method in combination with other available methods, keeping in mind aspects of the study organism's biology or sampling biases, as a way to achieve confidence in sibling reconstruction.

Another consideration is presentation and implementation of the methods. Most molecular ecologists do not have a background in computer science, and will opt for a method that is easily accessible, user-friendly, and produces results that can be readily interpreted, regardless of the underlying mathematical or computational elegance. COLONY is available as a Windows executable. However, it is computationally intensive and as such, is impractical to run on a personal computer. Our method does not require installation on a user's computer but provides a web-based service. It only requires an Internet connection to send the dataset for analysis using a web interface[3]. Our software accepts any file formatting using Excel software which is widely used by biologists.

Sibling reconstruction is among the first kinship reconstruction problems that have generated a variety of computational methods. However, more complicated pedigrees and genealogical relationships await computational solutions. Computationally, kinship reconstruction in wild populations is not only a rich source of interesting problems, but one that poses a particular challenge of testing the accuracy of devised solutions. Real biological data must be used to conduct comparisons of feasibility and accuracy of different methods. More benchmark data is needed to ground truth algorithms and software. Finally, novel approaches must be developed to assess accuracy of the resulting solutions and confidence in the answers provided.

Acknowledgments

This research is supported by the following grants: NSF IIS-0612044 and IIS-0611998 (Berger-Wolf, Ashley, Chaovalitwongse, DasGupta), Fulbright Scholarship (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 Dmitry Konovalov for sharing their software.

[3] See http://compbio.cs.uic.edu for more details

References

[1] A. Almudevar. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, **63**(2):63–75, 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**(2):136–165, 1999.

[3] M. V. Ashley and B D. Dow. The use of microsatellite analysis in population biology: background, methods and potential applications. *EXS*, **69**:185–201, 1994.

[4] Mary Ashley, Tanya Y. Berger-Wolf, Piotr Berman, Wanpracha Chaovalitwongse, Bhaskar DasGupta, and Ming-Yang 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] Tanya Y. Berger-Wolf, Saad I. Sheikh, Bhaskar Dasgupta, Mary V. Ashley, Isabel C. Caballero, Wanpracha Chaovalitwongse, and Satya P. Lahari. Reconstructing sibling relationships in wild populations. *Bioinformatics*, **23**(13):49–56, July 2007.

[7] Piotr Berman and Piotr Krysta. Optimizing misdirection. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 192–201, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

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

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

[10] W. Chaovalitwongse, T. Y. Berger-Wolf, B. Dasgupta, and M. V. Ashley. Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, **22**(1):11 – 24, February 2007.

[11] J. K. Conner. personal communication, 2006.

[12] J. L. Constable, M. V. Ashley, J. Goodall, and A. E. Pusey. Noninvasive paternity assignment in gombe chimpanzees. *Molecular Ecology*, **10**(5):1279–1300, 2001.

[13] B. D. Dow and M. V. Ashley. Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, quercus macrocarpa. *Molecular Ecology*, **5**(5):615–627, May 1996.

[14] B. D. Dow and M. V. Ashley. High levels of gene flow in bur oak revealed by paternity analysis using microsatellites. *Journal of Heredity*, **89**:62–70(9), January 1998.

[15] H.L. Dugdale, D. W. MacDonald, L. C. Pop, and T. Burke. Polygynandry, extra-group paternity and multiple-paternity litters in european badger (*Meles meles*) social groups. *Molecular Ecology*, **16**:5294–5306, 2007.

[16] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, **45**:634–652, 1998.

[17] K. A. Feldheim, S. H. Gruber, and M. V. Ashley. Population genetic structure of the lemon shark (negaprion brevirostris) in the western atlantic: DNA microsatellite variation. *Molecular Ecology*, **10**(2):295–303, February 2001.

[18] J. Fernández and M. A. Toro. A new method to estimate likelihood from molecular markers. *Molecular Ecology*, pages 1657–1667, May 2006.

[19] P. Gagneux, C. Boesch, and D. S. Woodruff. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear dna amplified from shed hair. *Molecular Ecology*, **6**(9):861–868, September 1997.

[20] M. R. Garey and D. S. Johnson. *Computers and Intractability* **-** *A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

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

[22] M. A. Halverson., D. K. Skelly, and A. Caccone. Kin distribution of amphibian larvae in the wild. *Molecular Ecology*, **15**(4):1139–1145, 2006.

[23] 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 versus in mixed family tanks. *Aquaculture*, **173**(1–4), March 1999.

[24] J. T. Hogg and S. H. Forbes. Mating in bighorn sheep: frequent male reproduction via a high-risk unconventional tactic. *Journal Behavioral Ecology and Sociobiology*, **41**(1):33–48, July 1997.

[25] C. A. Hurkens and A. Schrijver. On the size of systems of sets every t of which have an sdr with applications to worst-case heuristics for packing problems. *SIAM Journal of Discrete Mathematics*, **2**(1):68–72, 1989.

[26] Dean R. Jerry, Brad S. Evans, Matt Kenway, and Kate Wilson. Development of a microsatellite DNA parentage marker suite for black tiger shrimp *Penaeus monodon*. *Aquaculture*, pages 542–547, May 2006.

[27] D. S. Johnson. Approximation algorithms for combinatorial problems. *J. Comput. System Sci.*, **9**:256–278, 1974.

[28] Richard M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[29] D. A. Konovalov, C. Manning, and M. T. Henshaw. KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, **4**(4):779–82, December 2004.

[30] T. C. Marshall, J. Slate, L. E. B. Kruuk, and J. M. Pemberton. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**(5):639–655, May 1998.

[31] D. E. McCauley, M. J. Wade, F. J. Breden, and M. Wohltman. Spatial and temporal variation in group relatedness: Evidence from the imported willow leaf beetle. *Evolution*, **42**(1):184–192, January 1988.

[32] I. Painter. Sibship reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**(2):212–229, 1997.

[33] P. Pamil. Genotypic Correlation and Regression in Social Groups: Multiple Alleles, Multiple Loci nd Subdivided Populations. *Genetics*, **107**(2):307–320, 1984.

[34] P. Pamilo. Estimating relatedness in social groups. *Trends in Ecology & Evolution*, **4**(11):353–355, 1989.

[35] J. M. Pemberton. Wild pedigrees: the way forward. *Proceedings of Biological Sciences*, 2008.

[36] D. C. Queller and K. F. Goodnight. Estimating relatedness using genetic markers. *Evolution*, **43**(2):258–275, March 1989.

[37] D. C. Queller and K. F. Goodnight. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**(7):1231–1234, July 1999.

[38] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**(4839):487–491, 1988.

[39] C. Schltterer. The evolution of molecular markers–just a matter of fashion? *Nature Review Genetics*, **5**:63–69, January 2004.

[40] S. I. Sheikh, T. Y. Berger-Wolf, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, and B. DasGupta. Error-tolerant sibship reconstruction in wild populations. In *Proceedings of 7th Annual International Conference on Computational Systems Bioinformatics (CSB) (to appear)*, 2008.

[41] S. I. Sheikh, T. Y. Berger-Wolf, W. Chaovalitwongse, and M. V. Ashley. Reconstructing sibling relationships from microsatellite data. In *Proceedings of the European Conf. on Computational Biology (ECCB)*, January 2007.

[42] S. I. Sheikh, T. Y. Berger-Wolf, A. A. Khokhar, and B. DasGupta. Consensus methods for reconstruction of sibling relationships from genetic data. In *Proceedings of the 4th Workshop on Advances in Preference Handling (to appear)*, 2008.

[43] 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.

[44] S.M. Sogard, E. Gilbert-Horvath, E. C. Anderson, R. Fisher, S. A. Berkeley, and J. Carlos Garza. Multiple paternity in viviparous kelp rockfish, *Sebastes atrovirens*. *Environmental Biology of Fishes*, **81**:7–13, 2008.

[45] B. M. Strausberger and M. V. Ashley. Breeding biology of brood parasitic brown-headed cowbirds (*Molothrus ater*) characterized by parent-offspring and sibling-group reconstruction. *The Auk*, **120**(2):433–445, 2003.

[46] B. M. Strausberger and M. V. Ashley. Host use strategies of individual female brown-headed cowbirds molothrus ater in a diverse avian community. *Journal of Avian Biology*, **36**(4):313–321, 2005.

[47] R. Streiff, A. Ducousso, C. Lexer, H. Steinkellner, J. Gloessl, and A. Kremer. Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur L.* and *Q. petraea (Matt.) Liebl.* *Molecular Ecology*, **8**(5):831–841, 1999.

[48] D. Tautz. Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. *Nucl. Acids Res.*, **17**(16):6463–6471, August 1989.

[49] S. C. Thomas and W. G. Hill. Estimating Quantitative Genetic Parameters Using Sibships Reconstructed From Marker Data. *Genetics*, **155**(4):1961–1972, 2000.

[50] S. C. Thomas and W. G. Hill. Sibship reconstruction in hierarchical population structures using markov chain monte carlo techniques. *Genetics Research*, **79**:227–234, 2002.

[51] V. Vazirani. *Approximation Algorithms*. Springer, 2001.

[52] M.J. Vonhof, D. Barber, M. B. Fenton, and C. Strobeck. A tale of two siblings: multiple paternity in big brown bats (*Eptesicus fuscus*) demonstrated using microsatellite markers. *Molecular Ecology*, **15**:241–247, 2006.

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

[54] J. L. Weber and P. E. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American journal of human genetics*, **44**(3):388–396, March 1989.

[55] D. F. Westneat and M. S. Webster. Molecular analysis of kinship in birds: interesting questions and useful techniques. In B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle, editors, *Molecular Ecology and Evolution: Approaches and Applications*, pages 91–128. Basel, 1994.

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