BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i49–i56
doi:10.1093/bioinformatics/btm219

Reconstructing sibling relationships in wild populations

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

[1]Department of Computer Science, [2]Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60607 and [3]Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854

ABSTRACT

Reconstruction of sibling relationships from genetic data is an important component of many biological applications. In particular, the growing application of molecular markers (microsatellites) to study wild populations of plant and animals has created the need for new computational methods of establishing pedigree relationships, such as sibgroups, among individuals in these populations. Most current methods for sibship reconstruction from microsatellite data use statistical and heuristic techniques that rely on a priori knowledge about various parameter distributions. Moreover, these methods are designed for data with large number of sampled loci and small family groups, both of which typically do not hold for wild populations. We present a deterministic technique that parsimoniously reconstructs sibling groups using only Mendelian laws of inheritance. We validate our approach using both simulated and real biological data and compare it to other methods. Our method is highly accurate on real data and compares favorably with other methods on simulated data with few loci and large family groups. It is the only method that does not rely on a priori knowledge about the population under study. Thus, our method is particularly appropriate for reconstructing sibling groups in wild populations.

Contact: tanyabw@uic.edu

1 INTRODUCTION

For wild populations, the growing development and application of molecular markers provide new possibilities for establishing kinship and reconstructing pedigrees in species where such information cannot be obtained from field observations alone. Knowledge of kinship in wild or experimental populations of non-model organisms allows the investigation of many fundamental biological phenomena, including mating systems, selection and adaptation, kin selection and dispersal patterns. The power and potential of the genotypic information obtained in these studies often rests in our ability to reconstruct genealogical relationships among individuals (Garant and Kruuk, 2005). These relationships include parentage, full and half-sibships and higher order aspects of pedigrees (Blouin, 2003; Butler *et al.*, 2004; Jones and Ardren, 2003). In this article, we are only concerned with full sibling relationships.

While there are several potential molecular markers that could be applied to pedigree reconstruction, microsatellites (also known as SSRs, STRs, SSLPs and VNTRs) are the most widely used marker and offer several advantages. Unlike dominant markers such as AFLPs and ISSRs, microsatellite alleles are codominant, so inference of genotypes and allele frequencies at each locus are straightforward. Development of SNPs is more difficult and expensive than microsatellite development for species not subject to large-scale genome projects. More importantly, the power to identify related individuals depends mainly on the number of alleles per locus and their heterozygosity, and microsatellites are clearly superior to other markers in both regards, with 5–20 alleles and heterozygosities of >0.700 being typical, as reported in many wild populations. Finally, many field studies wish to estimate population parameters as well as individual relationships, so development and application of microsatellites is the best investment of resources for accomplishing such multiple goals. Because of these advantages of microsatellite over other markers, together with their current widespread use, we focus our development of sibship reconstruction methods to unlinked, multi-allelic, codominantly inherited markers, as these features describe microsatellite markers. Generally, phase or haplotype information is not available for microsatellite loci in non-model organisms.

While several methods for sibling reconstruction from mutli-allelic microsatellite data have been proposed (Almudevar, 2003; Almudevar and Field, 1999; Beyer and May, 2003; Konovalov *et al.*, 2004; Painter, 1997; Smith *et al.*, 2001; Thomas and Hill, 2002; Wang, 2004), most have not been 'ground-truthed' (Butler *et al.*, 2004) and have received relatively limited application. The majority of the kinship and pedigree reconstruction methods rely on the knowledge about typical allele distribution and frequency, family sizes, etc. and use statistical likelihood models to infer genealogical relationships (Blouin, 2003). We build on our earlier work (Berger-Wolf *et al.*, 2005; Chaovalitwongse *et al.*, 2007) and propose a new algorithm for sibship reconstruction using combinatorial optimization. There have been no truly combinatorial methods for kinship reconstruction problems (Almudevar and Field, 1999; Beyer and May, 2003). Combinatorial methods have been very successful in closely related molecular genetics questions, such as haplotype reconstruction (Eskin *et al.*, 2003; Li and Jiang, 2003). Our approach uses the simple Mendelian inheritance rules to impose constraints on the genetic content possibilities of a sibling group. We formulate the inferred combinatorial constraints and, under the parsimony assumption, use a provably correct algorithm to construct the smallest number of groups of individuals that satisfy these

*To whom correspondence should be addressed.

constraints. We test our approach on both simulated and real biological data.

2 METHODS

2.1 Microsatellite genetic markers

Microsatellites, also known as short tandem repeats (STR), simple tandem repeats (STR), simple sequence repeats (SSR), simple sequence length polymorphisms (SSLP) or variable number of tandem repeats (VNTR), are short sequences of repeated DNA (typically two to four base pairs). Different individual organisms can have microsatellites with different number of repeats at the same *locus* (part of DNA). In fact, this variability is what makes the microsatellites so useful for genetic analysis. In diploid organisms, an individual will have two copies of each microsatellite sequence, one from the mother, one from the father, called *alleles*. The two copies may differ in the number of repeats of the same segment, depending on the parental DNA. For example, if the mother has 'CA' repeated 8 times and 12 times, and the father has 10 and 13 repeats, then the offspring may have 12 and 10 'CA' repeats at that locus.

Finding each new microsatellite locus is time and resource consuming. Thus, microsatellite markers for non-model species typically consist of very few, 2–20 loci. Yet, once a locus is identified and the specific PCR primers are designed, screening each individual is relatively quick and cheap. Together with the high variability (high number of alleles per locus), this makes microsatellites the marker of choice for genetic research of wild populations.

2.2 Sibling reconstruction problem statement

The main focus of our article is to design a method that accurately reconstructs sibling groups from microsatellite data of a single generation. We now define the sibling reconstruction problem more formally. Given a genetic (microsatellite) sample at l loci from a population of n diploid individuals of the same generation, U, the goal is to reconstruct the full sibling groups (groups of individuals with the same parents). We assume no knowledge of parental information.

$$U = \{X_1, \ldots X_n\}, \text{ where } X_i = (<a_{i1}, b_{i1}>, \ldots, <a_{il}, b_{il}>)$$

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

The goal is to find a partition of individuals $P_1, \ldots P_m$ such that

$$\forall 1 \leq k \leq m, \forall X_u, X_v \in P_k : Parents(X_u) = Parents(X_v)$$

Notice, here that we have not defined the function $Parents(x)$. This is a biological objective. We will discuss computational approaches to achieve a good estimate of the biological sibling relationship.

2.3 2-Allele and 4-allele properties

Mendelian genetics lay down a very simple rule for 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 siblings groups: 4-allele property and 2-allele property (Berger-Wolf *et al*., 2005).

2.3.1 4-Allele property. The total number of distinct alleles occurring at any locus may not exceed four.

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

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

Clearly, the 4-allele property is necessary since a group of siblings can inherit only combinations of the four alleles of their common parents. The four-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 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 the alleles occurring at the first locus are {12, 31, 56, 44, 51}.

Table 1. An example of input data for the sibling reconstruction problem

Individual	Alleles (a/b) at locus1	Alleles (a/b) at locus2
Radish 1	44/44	55/23
Radish 2	12/56	14/31
Radish 3	31/44	55/14
Radish 4	13/13	31/23
Radish 5	31/51	14/31

The five individuals have been sampled at two genetic loci. Each allele is represented by a number. Same numbers represent the same alleles.

2.3.2 2-Allele property There exist 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 two.

Formally, a set $S \subseteq U$ has the 2-allele property if for each 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

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

2-Allele property is clearly stricter than 4-allele property. Looking at the Tables 1, our previous 4-allele set of individuals 1, 3 and 4 fails to satisfy the stricter 2-allele property as the alleles appearing on the left side at locus 1 {44, 31, 13} are more than two. Moreover, there is no swapping of alleles that will bring down the number of alleles on each side to two: the first and fourth individuals with alleles 44/44 and 13/13 already fill the capacity.

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. Notice, also, that *any* two individuals necessarily satisfy the 2-allele property as well since by default the number of alleles on each side of any locus is at most two.

The 2-allele property reduces the possible combinations of alleles at a locus in a group of siblings down to a few canonical options (modulo the numbering of the alleles). Assuming the alleles are numbered 1 through 4, Table 2 lists all different types of sibling groups possible with the 2-allele property. We do this by listing all possible pairs of parents whose alleles are among 1, 2, 3 and 4 and all the 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.

It is important to note that Table 2 gives an exhaustive list of canonical possibilities of allele combinations at a given locus in a group of siblings without violating the 2-allele property. Without the loss of generality, we assume that the alleles at each locus are numbered 1 through 4. This is sufficient since according to the 4-allele property, the number of alleles in any sibling group cannot exceed four. Further, there are $4! = 24$ possible mappings of any four alleles onto numbers 1–4. However, we list only the canonical minimal options (parents' alleles being numbered sequentially). It is not hard to check that the list of parents is exhaustive. Hence, Table 2 presents an exhaustive

Table 2. Canonical possible combinations of parent alleles and all resulting offspring allele combinations at a single locus

Parents	Offspring	
	allele *a*	allele *b*
Set parents (1/2) (3/4)		
	1	3
	2	4
	1	4
	2	3
	3	1
	4	2
	4	1
	3	2
Set parents (1/2) (1/3)		
	1	1
	2	3
	1	3
	2	1
	3	2
	3	1
	1	2
Set parents (1/2) (1/2)		
	1	1
	1	2
	2	1
	2	2
Set parents (1/1) (1/1)		
	1	1
Set parents (1/1) (1/2)		
	1	1
	1	2
	2	1
Set parents (1/1) (2/3)		
	1	2
	1	3
	2	1
	3	1
Set parents (1/1) (2/2)		
	1	2
	2	1

canonical list of possible sibling groups. It is also easy to verify that the resulting sibling groups indeed confirm to the 2-allele property.

2.4 Minimum 2-allele set cover

As we have mentioned, the biological function *Parents*(x) cannot be defined mathematically. We model the objective of reconstructing the sibling relationships mathematically by assigning individuals parsimoniously into the smallest number of (possibly overlapping) groups that satisfy the necessary 2-allele constraint. Formally, recall that we are given a population U of n diploid individuals sampled at l loci

$$U = \{X_1, \ldots X_n\}, \quad \text{where } X_i = (<a_{i1}, b_{i1}>, \ldots, <a_{il}, b_{il}>)$$

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 $S_1, \ldots, S_m$ such that each $S_i \subseteq U$ and satisfies the 2-allele constraint and $\bigcup S_i = U$.

We conjecture that the MINIMUM 2-ALLELE SET COVER is NP-complete. A simple corollary of the following theorem from Berger-Wolf *et al.* (2005) shows that it is in NP.

THEOREM 1 (Berger-Wolf *et al.*, 2005). *Let R be the number of alleles that are homozygous or appear with 3 other distinct alleles in a given locus and A be the total number of distinct alleles at a locus. Then a set of individuals satisfies the 2-allele property if and only if for every locus it satisfies the constraint*

$$A + R \leq 4$$

It is easy to see that given a set of individuals we can verify that it satisfies the 2-allele property in $O(nl)$ time using the constraint above. Thus, MINIMUM 2-ALLELE SET COVER is in NP.

Since the MINIMUM 2-ALLELE SET COVER is likely to be NP-hard, one approach is to design approximation algorithms or heuristics that will produce suboptimal solutions. Instead, we use commercial MIP solver CPLEX[1] to solve the problem to optimality.

2.5 Minimum 2-allele set cover algorithm

We now present our algorithm for solving the sibling reconstruction problem abstracted as the MINIMUM 2-ALLELE SET COVER. Our algorithm uses the 2-allele and 4-allele properties (specifically, Table 2) to generate all maximal potential sibling sets. We then restate the problem as a MINIMUM SET COVER to find the minimum number of sibling sets containing all the individuals. Thus, the algorithm has two steps:

(1) Create potential sibling sets based on the 2-allele property for each locus and maximally assign individuals to each set without violating the 2-allele property in any locus.

(2) Use minimum set cover to find the minimum number of the 2-allele sets from step 1 whose union contains all the individuals.

We now explain the algorithm in more detail. In step 1, we build on the approach presented in Berger-Wolf *et al.* (2005) and Chaovalitwongse *et al.* (2007) by generating sets that satisfy the 2-allele property. In the implementation of the algorithm, we use the complete version of Table 2 with all 24 possible mappings of alleles to numbers 1–4, to generate all maximal possible sets. Since the list is exhaustive, if a set does not match one of the patterns in Table 2 under some mapping of its alleles onto numbers 1–4, it cannot possibly be a sibling group. During both steps of our algorithm, we maintain an index or lookup of all sets to ensure there are no duplications.

2.5.1 Algorithm 2-allele. Recall that any pair of individuals necessarily satisfies the 2-allele property. Thus, initially we use all $\binom{n}{2}$ pairs of n individuals to generate the candidate *set*s. Each *set* is generated using the initial possible canonical sets from Table 2 for each locus j. Each allele is assigned a number between 1 and 4 based on the order of its occurrence. Then, for each pair of individual alleles we search for all matching canonical sets in Table 2 to determine the set of possibilities, *PossibilitiesSet*.

After generating these initial sets based on pairs of individuals, the algorithm repeatedly iterates through all the individuals, testing each set for a possible assignment of the individual to the set. In each cycle of the iterations, only the sets that were present at the beginning of the cycle are considered for each individual. An individual is assigned to

[1]CPLEX is a registered trademark of ILOG.

a set if its alleles match the possibilities of the set as defined by the extended Table 2.

However, adding an individual to a potential sibling set may reduce the set of the matching canonical patterns. For example, adding an individual with alleles 3/1 to a set of two individuals with alleles 1/2 and 2/1 changes the potential set of parents from {(1,1)(2,2); (1,1)(2,2); (1,2)(1,2); (1,1)(2,3); (1,2)(1,3); (1,2)(3,4)} to just {(1,1)(2,3); (1,2)(1,3); (1,2)(3,4)}. Thus, when adding a new individual to a set, we check if a new **valid** *set* can be created to accommodate all of the individuals already assigned to the set as well as the new individual. The validity of the new *set* is determined by the 4-allele property and the extended Table 2. The alleles at every locus of the new individual must match at least one of the canonical patterns that collectively satisfy all the previous individuals assigned to the set. Once we determine that the *set* can be expanded (and its set of possible matching parents reduced) to accommodate the new individual in a valid way, we create a modified copy of the set. The individual is then checked against this new set for all the remaining loci. After we have verified that the new individual does not violate the 2-allele property of the new *set* at every locus, as explained above, and verifying that the set does not already exist, we add the *set* to the collection of potential sibling sets. However, for the remainder of the iteration cycle all the individuals are checked only against the sets that had been present at the beginning of the cycle. This ordering ensures that each individual is checked against each *set* exactly once.

We repeat this process, cycling through all the individuals in the population. Once a set present at the beginning of the cycle has been inspected against all the individuals, the set is marked as *done* and is not revisited. This ensures that all sibling pairs that could possibly occur are evaluated, and that no sibling sets are generated that never occur in data.

The cycles of iterations over the individuals continues until all sets are marked as *done*. As the last step a singleton *set* for each of the elements is added containing just that element to ensure that a family group containing one offspring is possible.

After all the potential sibling sets are generated, we apply the minimum set cover to find the minimum number of sibling groups whose union contains all the individuals.

2.5.2 Proof of correctness and termination First, we note that the algorithm terminates since the sets newly added in each iteration cycle are always bigger than the sets present at the beginning of the iteration cycle and each individual can occur at most once in a set.

We already showed that Table 2 exhaustively lists all the canonical possibilities of sibling groups (modulo the mapping of alleles to the numbers 1–4). We show that our algorithm produces all the sibling groups that confirm to the listings in this table, and no sibling group is generated that does not satisfy one of the canonical possibilities.

THEOREM 2. *Algorithm 2-allele produces all and only the possible 2-allele groups that are supported by the data.*

Proof. As we have stated before, all possible pairs of individuals create minimal (non-singleton) valid sibling groups and must correspond to at least one of the entries in Table 2 by default. The algorithm then exhaustively compares every individual against every such possible sibling set and generates new sets as necessary if the 2-allele property is not violated. Thus, every combination of individuals that can be siblings will be generated. Suppose, to the contrary, there exists a valid maximal sibling group S that has never been generated and consider as such group the smallest. Let X_i be the individual with the highest index i in this group. When we remove the individual X_i from the population, all the individuals that could be siblings before can still be siblings. Thus, $S - X_i$ is still a valid sibling group and, by inductive hypothesis, it must have been generated by the algorithm. We examine X_i against the group $S - X_i$. Adding X_i does not violate the 2-allele property (since it is a sibling group) and therefore there exists a corresponding canonical set in Table 2 that contains S. Thus, we would add the corresponding possible set if it was not already among the *set*s.

Since we check every sibling group at all loci before adding it to the collection of potential sets, we ensure that we never generate an *set* that does not satisfy the 2-allele property at every locus. □

After all possible sibling groups are generated, we use the minimum set cover approach to find the smallest number of sibling groups whose union contains all the individuals. While the minimum set cover problem is NP-complete, modern mixed integer program solvers can solve it to optimally in most instances. Thus, it is not meaningful to discuss the theoretical computational complexity of the algorithm.

2.5.3 Minimum set cover. Minimum set cover problem is a classical NP-complete (Karp, 1972) problem. Minimum set cover is defined as follows: given a universe U of elements $X_1, \ldots, X_n$ 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.

Formally, given: $U = \{X_1, X_2, ..., X_n\}$ and $\mathcal{S} = \{S_1, S_2, ...S_m\}$ find

$$min|C| \; s.t. \; C \subseteq \mathcal{S} \text{ and } \bigcup_{S_i \in C} S_i = U$$

Set cover cannot be approximated in polynomial time to within a factor of $(1-\epsilon)\ln n$ unless $NP \subseteq DTIME(n^{\log\log n})$ (Feige, 1998). Johnson introduced a $1 + \ln n$ approximation in 1974 (Johnson, 1974).

In order to solve set cover, we use standard integer programming solvers. The integer program formulation of the set cover problem is as follows: given a matrix A

$$a_{ij} = \begin{cases} 1 & \text{if } i \in S_j \\ 0 & \text{otherwise} \end{cases}$$

the set cover problem is

$$\min \sum_{i=1}^{m} x_i \quad \text{s.t.} \quad \begin{array}{c} Ax \geq 1 \\ x_i \in \{0,1\} \end{array}$$

3 EVALUATION AND RESULTS

To validate and assess the accuracy of our approach, we use datasets with known genetics and genealogy. However, such biological datasets containing no errors are rare. In addition, we create simulated sets using a large number of parameters over a wide range of values. In each instance, we compare our algorithm to other methods for sibship reconstruction.

We measure the error by comparing the known sibling sets with those generated by our algorithm, and calculating the minimum partition distance (Gusfield, 2002). 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 3.0 GHz processors and a single Intel Pentium D Dual Core 3.2 GHz Intel processor with 4 GB RAM memory.

3.1 Sibship reconstruction methods

We compare the performance of our algorithm to three other sibship reconstruction methods. The methods span a variety of approaches and have different behavior on different parameters. We now describe the methods.

3.1.1 Almudevar and Field. Our algorithm is based on a very similar idea proposed by Almudevar and Field (1999) which is a completely combinatorial approach. Here, potential sibling sets are too constructed using the 2-allele property (although the authors do not explicitly state the property). However, these sets are constructed by enumerating exhaustively all combinations of individuals and testing those for the compliance with the 2-allele property. At the end, a maximal, not necessarily optimal, collection of sibling sets is returned as a solution.

3.1.2 Beyer and May. The approach proposed in Beyer and May (2003) 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. Potential families are identified by the connected components in this graph.

3.1.3 KinGroup. Konovalov *et al.* (2004) have proposed an algorithm based entirely on likelihood estimates of partitions of individuals into sibling groups. The individuals are considered one at a time. For each individual, the likelihood of it being part of any existing sibling group, as well as starting its own group, is calculated. The individual is placed into the group it is most likely to belong. Unfortunately, the outcome heavily depends on the order in which the individuals are considered.

3.2 Biological data

We have identified four biological datasets of microsatellite data where sibling groups are known. These are not wild populations since in wild populations we typically do not know the true sibling groups, which is precisely why we need the sibling reconstruction method.

3.2.1 Radishes. The wild radish *Raphanus raphanistrum* dataset (Conner, 2006) consists of samples from 150 radishes from 2 families with 17 sampled loci. There are missing alleles among all the loci. The parent genotypes are available.

3.2.2 Salmon. The Atlantic salmon *Salmo salar* dataset comes from the genetic improvement program of the Atlantic salmon federation (Herbinger *et al.*, 1999). We use a truncated sample of microsatellite genotypes of 351 individuals from 6 families with 4 loci per individual. The data does not have missing alleles at any locus. This dataset is a subset of one of the samples of genotyped individuals used by Almudevar and Field (1999) to illustrate their technique.

3.2.3 Shrimp. The tiger shrimp *Penaeus monodon* dataset (Jerry *et al.*, 2006) consists of 59 individuals from 13 families with 7 loci. There are 16 missing alleles. The parentage is known.

3.2.4 Flies. *Scaptodrosophila hibisci* dataset (Wilson *et al.*, 2002) consists of 190 same generation individuals (flies) from 6 families sampled at various number of loci with up to 8 alleles per locus. Parent genotypes were known. All individuals shared 2 sampled loci which were chosen for our study. Some of the alleles were missing for some of the individuals.

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

3.3 Random simulations

In addition, 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 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 (Gusfield, 2002). In addition, we compare the accuracy of our 2-allele algorithm to the two reference sibling reconstruction methods (Beyer and May, 2003; Konovalov *et al.*, 2004) 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 Almudevar and Field (1999) since the current version of 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, randomly and uniformly 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 is

Table 3. Accuracy (percent) of our algorithm and the three reference algorithms on biological datasets

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	531	75.90	Out of memory	53.30	29.95
Flies	2	190	100.00	31.05	27.89	54.73

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 Almudevar and Field (1999) (A&F), Beyer and May (2003) (B&M) and the KinGroup by Konovalov *et al.* (2004) (KG).
Almudevar and Field's algorithm ran out of 4 GB memory on the salmon and radish datasets.

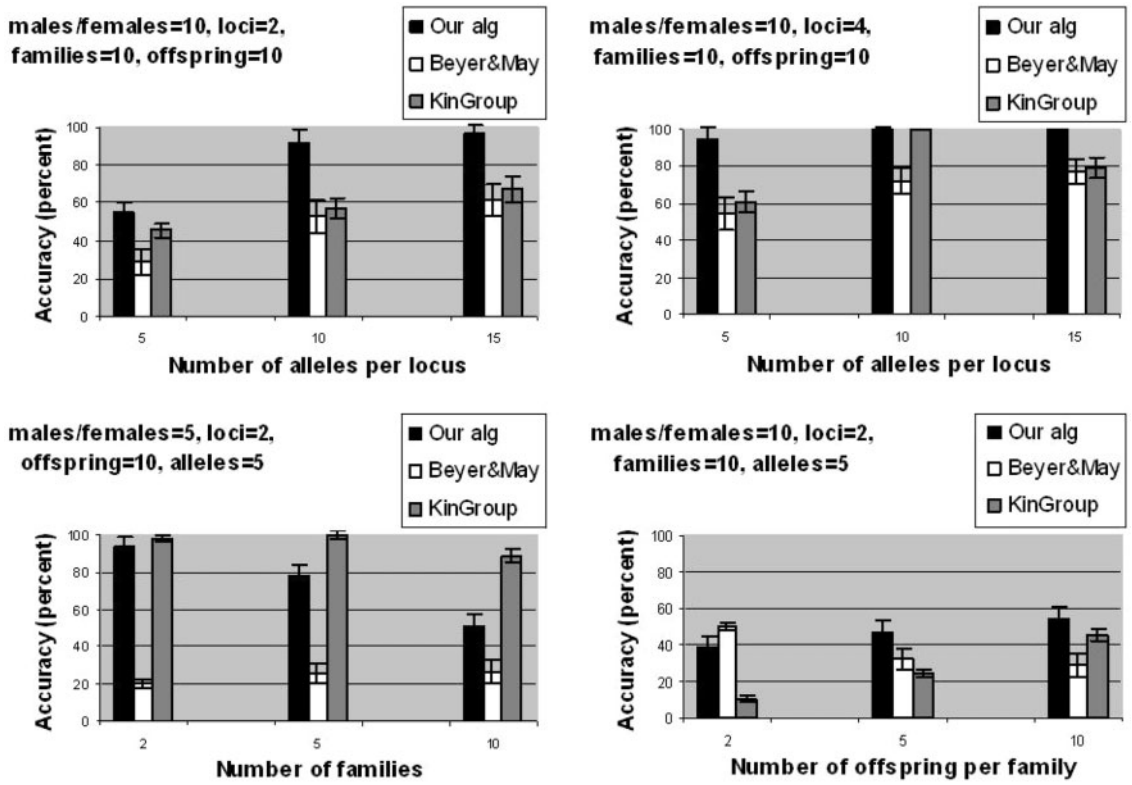

Fig. 1. Accuracy of the sibling group reconstruction using the 2-allele algorithm on randomly generated data. The *y*-axis shows the accuracy of reconstruction as a function of various simulation parameters. The accuracy of our algorithm is shown, as well as that of the two reference algorithms: Beyer and May (2003) and Konovalov *et al.* (2004) (KingGroup). 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.

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 Almudevar (2003).
- The number of families in the population $f = 2, 5, 10$.
- The number of offspring per couple (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 Almudevar (2003).

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 juvenile 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 the 2-allele algorithm 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 1 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 2 shows results for the datasets with skewed family sizes and allele frequency distributions. Each bar represents the mean value of a 1000 random repetitions, and the error bars show the standard deviation.

4 DISCUSSION AND CONCLUSIONS

We have proposed a new fully combinatorial algorithm for the problem of reconstruction of sibling relationships from single generation microsatellite genetic data. We have implemented and tested our approach on both biological and simulated data.

On biological data, our algorithm performed as well or better than other sibling reconstruction methods. The difference is

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

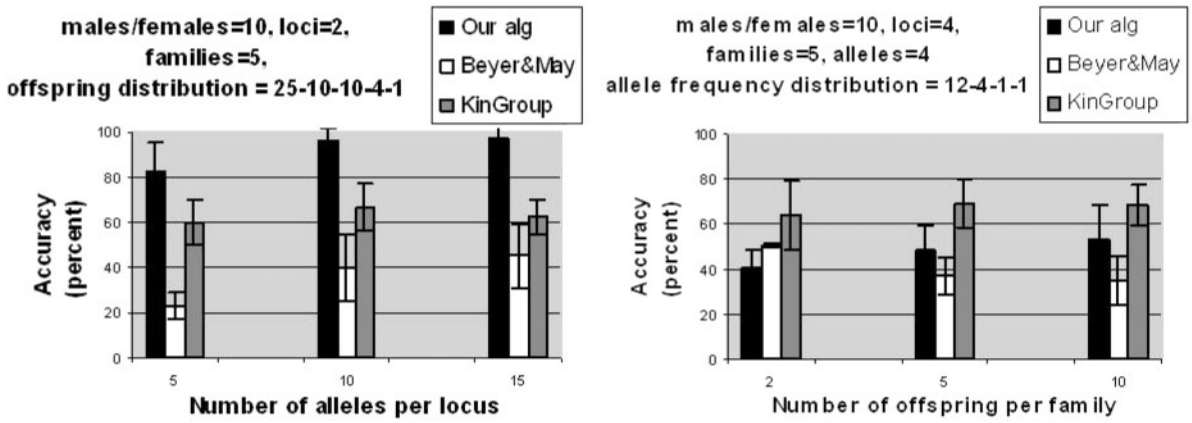

Fig. 2. Accuracy of the sibling group reconstruction using our 2-allele algorithm and the two reference methods on the datasets with skewed family sizes and allele frequency distributions.

particularly striking for the flies dataset with 2 loci. Our algorithm accurately reconstructed the sibling groups despite some missing alleles while other methods all have over 45% error rate. The radish dataset presented a problem for all methods since it has partial self-reproduction which none of the methods take into account. Offspring of a selfed individual are hard to separate from their half-siblings produced by that and any other individual. Still, even on this dataset our method performed significantly better than others.

The simulated data provides a base line for the accuracy estimate of our algorithm, with real biological data likely to have better reconstruction accuracy, as indicated by the results on the biological datasets. For the datasets with the uniform allele frequency and family sizes distributions, for the number of alleles per locus above 5 and the number of offspring per family above 5, the accuracy of our algorithm is above 50% in most cases, rapidly increasing as the number of offspring or alleles increases. Our algorithm performs significantly better than other methods when the number of loci is very small and there is reasonable diversity of alleles. In fact, the algorithm of Beyer and May (2003) has high error rates specifically for those parameter values. Thus, our algorithm is particularly well suited for natural populations of plants and animals with large family groups and few sampled loci.

However, we have conducted a very limited set of experiments on datasets with non-uniform parameter distributions. To obtain conclusive results, we need to explore a wider range of non-uniform distributions. To fully evaluate the performance of our algorithm, we need to validate the results on other biological datasets and more realistic simulated populations. In addition, we have yet to address the possibilities of errors in data. The fact that our algorithm accurately reconstructed sibling groups on biological data with missing alleles is encouraging. However, our algorithm would need to be modified to address errors from mistyped and misidentified alleles.

It is impossible in the current setting of the experiments to accurately compare running times of the algorithms. However, the algorithm of Almudevar and Field (1999), which uses exhaustive enumeration of all potential sibling sets, unsurprisingly ran out of memory on datasets >200 individuals. It took on the order of hours to complete on the dataset of 190 individuals. Moreover, since the current implementation requires user interaction, the performance of the algorithm could not be evaluated in the random simulations. The likelihood approaches of Beyer and May (2003) and Konovalov *et al*. (2004) are very fast. Both produced answers in a matter of seconds on datasets of a 100 individuals and in less than 2 min on datasets of 500 individuals. For our algorithm, we first (in matter of seconds) formulate the 2-allele set cover problem, then this formulation is imported into CPLEX and solved as a set cover problem. Recall that the set cover problem is NP-complete and CPLEX is a commercial software designed specifically to solve such computationally hard problems to optimality. The entire (automated) process takes ~2 h on datasets of 500 individuals. At this point, our focus has been on evaluating the accuracy and viability of our approach. Now that our approach has proven viable, we will concentrate on improving the running time and the overall usability of our method.

The main advantage of the combinatorial approach is its lack of reliance on a priori knowledge about various population parameters, such as allele frequency and the degree of inbreeding. However, Mendelian constraints do not provide any basis for distinguishing between family groups when the groups are small, or when individuals share many common alleles. Additional information, such as *relative* allele frequency within the sample can be easily added to generate combinatorial constraints on the potential sibling sets. Unlike likelihood methods, combinatorial approaches use that information only for comparison purposes, and do not require a background data model or an accurate estimate value for any of the parameter. Thus, we believe that combinatorial approaches are particularly appropriate for analysis of natural animal and plant populations where background information is difficult to obtain.

Our technique can be extended to solve a number of related problems. The first immediate variation is reconstructing sibship relationships when partial information, such as one of the parents, is available. We have already implemented and

applied the extended version of our algorithm to identifying the minimum number of necessary male oak trees that have pollinated a single female tree to produce the sampled acorns. Our approach provided additional supporting evidence that oak pollen disperses much further than previously thought.

Another simple variation of our algorithm produces half-sibling groups as well as full sibs. In addition, our algorithm can be used to identify the parsimonious set of parental genotypes necessary to produce the sibling groups.

Finally, from the algorithmic perspective, there are a number of alternatives that would improve the performance of our method that we are currently exploring.

To conclude, we have presented a fully combinatorial approach to reconstructing sibling groups from microsatellite data. Our method does not rely on any a priori knowledge about data parameters yet provides results with accuracy comparable to or better than those of likelihood methods.

ACKNOWLEDGEMENTS

This research is supported by the following grants: NSF IIS-0612044 and IIS-0611998 (T.Y.B-W., M.V.A., W.C., B.D.), Fulbright Scholarship (S.I.S.), NSF CCF-0546574 (W.C.). 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 and the anonymous reviewers for very helpful comments.

Conflict of Interest: none declared.

REFERENCES

Almudevar,A. (2003) A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theor. Popul. Biol.*, **63**, 63–75.

Almudevar,A. and Field,C. (1999) Estimation of single generation sibling relationships based on DNA markers. *J. Agric. Biol. Environ. Stat.*, **4**, 136–165.

Berger-Wolf,T.Y. *et al.* (2005) Combinatorial reconstruction of sibling relationships. In *Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics (CBGI 05)*, Utah, 1252–1255.

Beyer,J. and May,B. (2003) A graph-theoretic approach to the partition of individuals into full-sib families. *Mol. Ecol.*, **12**, 2243–2250.

Blouin,M.S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *TRENDS Ecol. Evol.*, **18**, 503–511.

Butler,K. *et al.* (2004) Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Mol. Ecol.*, **13**, 1589–1600.

Chaovalitwongse,A. *et al.* (2007) Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, **22**, 11–24.

Conner,J.K. (2006) Personal communication.

Eskin,E. *et al.* (2003) Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinform. Comput. Biol.*, **1**, 1–20.

Feige,U. (1998) A threshold of ln *n* for approximating set cover. *J. ACM*, **45**, 634–652.

Garant,D. and Kruuk,L.E.B. (2005) How to use molecular marker data to measure evolutionary parameters in wild populations. *Mol. Ecol.*, **14**, 1843–1859.

Gusfield,D. (2002) Partition-distance: a problem and class of perfect graphs arising in clustering. *Inf. Process. Lett.*, **82**, 159–164.

Herbinger,C. *et al.* (1999) Early growth performance of atlantic salmon full-sib families reared in single family tanks or in mixed family tanks. *Aquaculture*, **173**, 105–116.

Jerry,D.R. *et al.* (2006) Development of a microsatellite DNA parentage marker suite for black tiger shrimp monodon. *Aquaculture*, **255**, 542–547.

Johnson,D.S. (1974) Approximation algorithms for combinatorial problems. *J. Comput. Syst. Sci.*, **9**, 256–278.

Jones,A.G. and Ardren,W.R. (2003) Methods of parentage analysis in natural populations. *Mol. Ecol.*, **12**, 2511–2523.

Karp,R.M. (1972) Reducibility among combinatorial problems. In Miller,R.E. and Thatcher,J.W. (eds.), *Complexity of Computer Computations*, Plenum Press, pp. 85–103.

Konovalov,D.A. *et al.* (2004) KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Mol. Ecol. Notes*, **4**, 779–782.

Li,J. and Jiang,T. (2003) Efficient inference of haplotypes from genotype on a pedigree. *J. Bioinform. Comput. Biol.*, **1**, 41–69.

Painter,I. (1997) Sibship reconstruction without parental information. *J. Agric. Biol. Environ. Stat.*, **2**, 212–229.

Smith,B.R. *et al.* (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.

Thomas,C.S. and Hill,W.G. (2002) Sibship reconstruction in hierarchical population structures using markov chain monte carlo techniques. *Genet. Res. Camb.*, **79**, 227–234.

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

Wilson,A. *et al.* (2002) Isolation and characterization of 20 polymorphic microsatellite loci for *Scaptodrosophila hibisci*. *Mol. Ecol. Notes*, **2**, 242–244.