

# Reconstructing parents from genotyping data in wild <sup>1</sup> populations

Saad Sheikh<sup>1</sup>, Tanya Y. Berger-Wolf<sup>1</sup>, Bhaskar DasGupta<sup>1</sup>, Ashfaq A. Khokhar<sup>1</sup>, Mary V. Ashley<sup>2</sup>, Wanpracha Chaovalitwongse<sup>3</sup>, Isabel C. Caballero<sup>2</sup>

## 1 Introduction

Many wild species, from large marine animals to small terrestrial plants, are often difficult to study in their natural settings. In particular, it is challenging to track the mating patterns of such species, yet understanding these patterns is essential for understanding many aspects of population biology, ecology, and evolution. The goal of the kinship reconstruction problem is, given genotypic data of a population sample, infer the low order genealogical relationships among individuals. Our work is focused on reconstructing parents of a cohort of individuals by inferring the full-sibling groups, therefore, for the remaining of the paper we refer to full-sibling groups as *sibling groups*, or *sibgroups*. In [1] we presented a novel parsimony-based formulation for reconstructing parents: Find the sibling reconstruction for a given cohort that minimizes the number of parents necessary to explain a given population. We also showed that it is NP-Hard and inapproximable to  $2^{\log^\epsilon n}$  unless  $NP \subseteq DTIME(n^{\text{polylog}(n)})$  for every constant  $0 < \epsilon < 1$ . In this abstract we present a non-trivial template algorithm that can be used to get both exact or approximate solutions to the problem.

## 2 Meta-Approach for Min-Parents Sibling Reconstruction

Given a microsatellite (multiallelic) sample of a cohort of diploid individuals at a (small) number of loci, we use Mendelian inheritance rules to evaluate whether any given set of individuals  $c$  is a feasible sibling group. We now present a meta-approach to solving this problem. The approach is composed of three steps:

- 1) Generate a set  $\mathcal{S}$  of full-sib groups for the individuals
- 2) Generate a parent graph  $G = (V, E)$  from  $\mathcal{S}$  where every vertex represents a possible parent and every edge represents a mating between two parents. Associated with each vertex and edge are children of that parent and a mating, respectively.
- 3) Find a minimal subset of vertices,  $\mathcal{P}$ , such that the set of individuals associated with subset of edges with both endpoints in  $\mathcal{P}$  contains all the offspring individuals.

In step 1, all maximal feasible sibling groups must be used to ensure an exact algorithm. However, this results in the problem becoming prohibitively large. The *Greedy Cover* heuristic first generates all maximal feasible sibling groups  $\mathcal{M}$  and then uses the *Greedy Set Cover* [7] algorithm to select a subset containing all individuals from the set of the feasible groups as the basis for the parent graph. Alternatively, using *Greedy k-Covers* heuristic  $k$  minimal set covers are extracted using the classical greedy set cover algorithm [7] on the set of all maximal feasible sibling groups  $\mathcal{M}$ . Lastly, *Random k-Minimal Covers* is a randomized approach geared for a quicker solution. Instead of computing all maximal feasible sibling groups, which is computationally very demanding,  $k$  random minimal sibling reconstructions are used.

In step 2, all  $2^{|\text{loci}|}$  parents must be generated and added to the graph in order to get an exact algorithm. Alternatively, as a heuristic association rule mining can be used to find out pairs of alleles across loci are used to find the correct parents. It can be shown that this association rule mining gives the correct result in case of polygamous parents, and in case of monogamous parents the result is unaffected by this choice.

---

<sup>1</sup>Department of Computer Science, University of Illinois at Chicago E-mail:{ssheikh,tanyabw,dasgupta,ashfaq}@cs.uic.edu

<sup>2</sup>Department of Biological Sciences, University of Illinois at Chicago. E-mail:{ashley,icabal2}@uic.edu

<sup>3</sup>Department of Industrial Engineering, Rutgers University. E-mail:wchaoval@rci.rutgers.edu

In step 3, the problem can be solved to optimality using Integer Linear Programming. As a heuristic it suffices to greedily select one vertex at a time that promises to cover the most uncovered individuals.

### 3 Validation and Testing

We compared the performance of the combinations of different heuristics described above as well as the minimum sibling groups formulation[2]. We used datasets where the original sibship composition was known and compared that to the sibship outcome given by each method using the Gusfield Partition Distance [3]. We use three biological datasets of microsatellite data where sibling groups are known. We did not use wild populations since the true sibgroups composition is unknown, which is precisely why we need the kinship reconstruction method. We were able to collect 3 datasets: Shrimp [6], Salmon [5], and Ants[4], the results are described in Figure 1. We generated random simulated data similar to our approach in [2] and compared the performance of the heuristics.

### 4 Discussion and Conclusions

We have compared qualitatively our previous parsimony-based approach and the one presented in this paper. The advantage of a purely parsimony-based approach is its wide applicability as no prior knowledge about population parameters is required. We presented a meta-approach to solving this problem which can be used to give both exact and inexact solutions. Since the problem is NP-Complete and just as hard to approximate, we did not implement and test the exact algorithm. On the practical side, we proposed several heuristics that can replace steps in the exact algorithm for the Minimum Parents Sibling Reconstruction. Overall the  $k$ -greedy heuristic with an optimal graph cover yields the better results among the various heuristics. Its performance should improve as we increase the value of  $k$ .

Even though we were unable to provide effective heuristics for achieving the minimum number of parents, we have broken the problem into substeps and thus laid ground-work for using better heuristics in future.

Dataset			Accuracy(%)					Estimated Number of Parents				
Name	$l$	Inds	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
Shrimp	7	59	84.8	83	42.4	71.2	83	26	28	44	30	28
Salmon	4	351	98.3	98.3	98.3	86.9	98.3	14	14	14	15	14
Ants	6	377	99.7	99.7	94.2	92.3	92.8	20	20	22	23	22

Legend:

- M1:  $k$ -greedy cover with optimal graph cover
- M2: greedy set cover with optimal graph cover
- M3: Randomized cover with optimal graph cover
- M4:  $k$ -greedy with graph heuristics
- M5: greedy set cover with graph heuristic

Table 1: Accuracy % and Number of parents given by heuristics on biological datasets. Here  $l$  is the number of loci in a dataset and “Inds” column gives the number of individuals in the dataset.

### References

- [1] M. V. Ashley, T. Y. Berger-Wolf, Wanpracha Chaovalitwongse, B. DasGupta, and S. Sheikh. On approximating an implicit cover problem in biology. In *Proceedings of the 5<sup>th</sup> International Conference on Algorithmic Aspects of Information and Management*, (to appear).
- [2] 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.
- [3] D. Gusfield. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164, May 2002.
- [4] R. L. Hammond, A. F. G. Bourke, and M. W. Bruford. Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum*. *Molecular Ecology*, 10(11):2719–2728, 1999.
- [5] 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):105–116, March 1999.
- [6] 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, 2006.
- [7] D. S. Johnson. Approximation algorithms for combinatorial problems. *J. Comput. System Sci.*, 9:256–278, 1974.