

Family Trios Phasing and Missing data recovery

D. Brinza¹, J.He^{1,2}, W. Mao¹ and A. Zelikovsky^{1,3}

Keywords: haplotypes, genotypes, SNP, family trio data, phasing

1 Introduction

Although there exist many phasing methods for unrelated adults or pedigrees, phasing and missing data recovery for data representing family trios is lagging behind. This work is an attempt to fill this gap by considering the following problem. Given a set of genotypes partitioned into family trios, find for each trio a quartet of parent haplotypes which agree with all three genotypes and recover the SNP values missed in given genotype data. Our contributions include (i) formulating the pure-parsimony trio phasing and the trio missing data recovery problems, (ii) proposing two new greedy and integer linear programming based solution methods, and (iii) extensive experimental validation of proposed methods showing advantage over the previously known methods.

2 Family Trio Phasing Validation

It is clear how to validate a phasing method on simulated data since the underlying haplotypes are known. The validation on real data is usually performed on the trio data. E.g., a phasing method is applied to parents (respectively, to offspring) genotypes and the resulted haplotypes are validated on offspring's (respectively, on parents') genotypes. Unfortunately, in our case, one can not apply such validation since a trio phasing method may rely on both offspring and parents' genotypes. Therefore, we suggest to validate trio phasing by erasing randomly chosen SNP values and recording the errors in the erased SNP sites. In Tables 1, 2, each row corresponds to an instance of real data (Daly et al. or Gabriel et al.) or simulated data (ms) and the column (E) shows the percent of erased data (0% - no data erased, 1%-10% - percent of SNP values erased) .

The value of phasing errors is measured by the Hamming distance from the method's solution to the closest feasible phasing. In Table 1, for parents (P) we report the percent of SNP values that should be inverted out of the total number of SNP values that should be inferred (i.e., number of 2 plus number of unknown values). For offspring (C), we report the percent of SNP which should be inverted with respect to the total number of SNPs. The total number of errors (T) is the percent of SNP's that should be inverted in order to obtain a feasible phasing solution.

3 Missing Data Recovery in Family Trios

Comparison of five methods (ILP, Greedy, Phamily[6], PHASE[5] and HAPLOTYPER[4]) on trio missing data recovery on the real data sets (Daly [1] and Gabriel [2]) and simulated data are shown in the Table 2. We erase random data in trio genotypes with certain

¹Department of Computer Science, Georgia State University, Atlanta, GA 30303.

E-mail: {dima, jingwu, weidong, alexz}@cs.gsu.edu.

²Supported by GSU Molecular Basis of Disease Fellowship

³Partially supported by NIH Award 1 P20 GM065762-01A1.

amount(1%, 2%, 5% and 10%) of the entire data. Instead, We report the error as the number of incorrectly recovered erased positions of the genotypes on offspring (C*), parents (P*) and trios (T*) divided the total number of erased positions in parent genotypes in percentage. We count only half error if the compared paired SNP is 2 and 0 (or 1).

4 Results

Data	ILP				Greedy			Phamily[6]			PHASE[5]			HAPLOTYPYER[4]		
	E	C	P	T	C	P	T	C	P	T	C	P	T	C	P	T
Daly et al. [1]	0	0.0	0.0	0.0	4.9	16.2	3.8	1.3	0.0	0.7	1.1	0.0	0.6	2.2	0.0	1.2
	1	0.2	0.5	0.2	4.8	16.8	3.8	1.2	1.4	0.7	1.3	0.2	0.7	2.1	1.0	1.6
	2	0.3	0.7	0.4	5.0	16.9	4.0	1.3	1.8	0.9	1.3	0.5	0.8	2.2	2.3	1.7
	5	0.8	2.6	1.2	5.3	17.1	4.0	1.3	1.0	1.0	1.6	0.9	1.0	2.3	7.0	2.9
Gabriel et al. [2]	0	0.0	0.0	0.0	2.9	11.5	2.2	3.0	0.0	2.0	2.2	0.0	1.3	4.4	0.0	2.7
	1	0.2	0.6	0.2	2.9	12.1	2.3	3.1	0.2	2.0	2.8	0.2	1.7	4.6	1.7	1.5
	2	0.3	1.2	0.5	3.2	12.2	2.4	3.3	0.4	2.1	2.9	0.6	1.8	4.9	3.1	1.6
	5	0.8	3.4	1.1	3.4	12.2	2.9	3.4	1.3	2.5	3.0	1.4	1.6	5.4	6.3	2.1
ms [3]	0	0.0	0.0	0.0	2.6	13.2	1.9	9.4	0.0	4.7	5.6	0.0	6.5	8.1	0.0	5.4
	1	0.3	1.0	0.4	2.9	13.5	1.9	10.1	0.8	4.3	5.8	1.2	5.4	8.4	2.2	5.6
	5	1.3	3.8	1.9	4.3	13.9	3.1	10.6	3.8	7.6	6.1	4.7	5.9	9.2	10.2	7.0
	10	2.5	7.7	3.6	5.3	14.0	4.4	11.9	9.5	9.2	6.9	10.5	6.0	11.5	17.1	8.0

Table 1: The results for five phasing methods on the real data sets of Daly et al.[1] and Gabriel et al. [2] and on simulated data. The E column corresponds to the ratio of erased data, C to the error of offspring, P to the error of parents, T the total error.

Data	E	ILP			Greedy			Phamily[6]			PHASE[5]			HAPLOTYPYER[4]		
		C*	P*	T*	C*	P*	T*	C*	P*	T*	C*	P*	T*	C*	P*	T*
Daly et al. [1]	1	2.3	7.8	5.7	3.9	6.0	5.2	0.3	2.3	1.5	0.3	3.1	2.0	1.9	26.1	16.7
	5	3.9	9.9	7.8	4.5	4.8	4.7	0.2	3.6	2.5	0.1	3.4	2.3	1.3	20.5	13.9
	10	5.7	13.5	10.8	4.6	5.8	5.4	0.6	4.4	3.1	0.5	4.0	2.8	1.5	21.8	14.8
Gabriel et al. [2]	1	7.7	8.0	7.9	5.6	6.4	6.1	0	2.5	1.6	0.4	3.1	2.1	1.6	21.8	14.5
	5	7.9	8.7	8.4	5.6	5.8	5.7	0	2.3	1.5	0.1	3.3	2.2	2.5	20.7	14.6
	10	7.4	9.5	8.8	6.1	6.6	6.5	0.1	2.1	1.5	0.3	3.1	2.1	2.3	25.1	17.5
ms [3]	1	10.9	13.3	12.4	11.5	9.2	10.1	1.0	16.0	10.2	0.7	15.2	9.6	4.3	26.4	17.9
	5	13.1	12.1	12.4	12.3	7.8	9.3	0.9	14.8	10.0	0.7	14.9	10.0	3.6	23.1	16.4
	10	12.0	12.4	12.3	11.6	8.9	9.8	2.3	14.4	10.3	0.7	13.9	9.3	3.4	21.9	15.5

Table 2: The results for missing data recovery on the real and simulated data sets with five methods. The E column corresponds to the ratio of erased data, C* to the error of offspring, P* to the error of parents, T* to the total error.

References

- [1] M. Daly, et al. High resolution haplotype structure in the human genome. *Nature Genetics*, 2001.
- [2] G. Gabriel, et al. The structure of haplotype blocks in the human genome. *Science*, 296:2225, 2002.
- [3] R. Hudson. *Gene genealogies and the coalescent process*. Oxford Survey of Evolutionary Biology, 1990.
- [4] T. Niu, et al. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, 2002.
- [5] M. Stephens, et al. A new statistical method for haplotype reconstruction from population data, 2001.
- [6] H.Ackerman, et al. Haplotypic analysis of the TNF locus by association efficiency and entropy, 2003.