Haplotype Inference Combining Pedigrees and Unrelated Individuals

Ana Graça[1], Inês Lynce[1], João Marques-Silva[2], and Arlindo L. Oliveira[1]

[1] IST/INESC-ID, Technical University of Lisbon, Portugal
{assg,ines}@sat.inesc-id.pt,aml@inesc-id.pt
[2] School of Computer Science and Informatics, University College Dublin, Ireland
jpms@ucd.ie

Abstract. Haplotype inference is a crucial topic in genetic studies and also represents a challenging computational problem. A significant number of combinatorial approaches tackle the haplotype inference problem either for pedigrees or for unrelated individuals. This work integrates two relevant and well-known constraint based haplotyping approaches. The Minimum Recombinant Haplotyping Configuration (MRHC) problem targets the haplotyping solution which minimizes the number of recombinant events within a pedigree. MRHC only takes into consideration the family information. In contrast, the Haplotype Inference by Pure Parsimony (HIPP) problem aims at finding a solution which minimizes the number of distinct haplotypes. The HIPP approach is adequate for phasing unrelated individuals from the same population. This paper proposes a method for inferring haplotypes for individuals of the same population, although organized in different families, thus combining both MRHC and HIPP approaches. This new method can take into account family information and population information, both important in haplotype inference. Experimental results show that the proposed approach is more accurate, both in terms of switch error rate and missing error rate, than the MRHC approach (performed by the PedPhase tool), on sets of families from the same population.

1 Introduction

Genetic association studies with phenotypic variations are only possible with a deep knowledge of the genetic differences between individuals. A very important and challenging task to understand genetic variations consists of inferring haplotypes from genotypes.

Constraint based methods for haplotype inference have been shown to be a practical and relevant alternative to statistical approaches, either for phasing pedigrees [9, 11] or unrelated individuals [4]. Nonetheless, a study comparing the haplotype inference methods using pedigrees and unrelated individuals [12] points out that a new method which takes into consideration both pedigree and population information is necessary. Indeed, existing haplotyping methods for pedigrees ignore the population information, while haplotyping methods for unrelated individuals do not take into account the pedigree information.

The work presented in this paper was motivated by the comparison study described in [12]. A constraint based model to deal with families and unrelated individuals is proposed. The new method is based on two well-known combinatorial approaches: MRHC and HIPP. The Minimum Recombinant Haplotype Configuration (MRHC) approach is used to phase individuals organized in pedigrees, by minimizing the number of recombination events within each pedigree. In general, a significant number of solutions can be obtained using only the minimum recombinant paradigm, especially when several families are considered. Thus, the Haplotype Inference by Pure Parsimony (HIPP) approach is considered to

choose a solution that uses the minimum number of distinct haplotypes, among all the minimum recombinant solutions. The new method for haplotype inference, named PedRPoly, is shown to be more accurate than the method traditionally used for inferring haplotypes on pedigrees using the minimum recombinant approach, as performed by the PedPhase tool [11]. PedRPoly is also shown to be more accurate than the pure parsimony approach, performed by the RPoly tool [4], when pedigrees are considered. In addition, this paper suggests some reductions on the size of the original integer programming MRHC model.

The paper is organized as follows. The next section describes the haplotype inference problem and overviews the MRHC and HIPP approaches. Section 3 details the new proposed model, PedRPoly, which combines MRHC and HIPP formulations. Afterwards, experimental results comparing the accuracy of PedPhase and PedRPoly are presented and discussed. Finally, the conclusions are presented in section 5.

2 Haplotype Inference

Single Nucleotide Polymorphisms (SNPs) are the most common variations between human beings, which occur when a nucleotide is mutated into another nucleotide at a single DNA position. Haplotypes correspond to the set of closely linked SNPs, within a single chromosome, which tends to be inherited together. However, it is very expensive and time consuming to determine experimentally the haplotypes. Instead, only genotypes, which correspond to the conflated data of two haplotypes on homologous chromosomes are obtained. The haplotype inference problem consists in determining the haplotypes which originate a given genotype.

Considering that mutations are rare, we may assume that each SNP can only have two values. Each haplotype is therefore represented by a binary string, with size $m \in \mathbb{N}$, where 0 represents the wild type nucleotide and 1 represents the mutant type nucleotide. Each site of the haplotype h_i is represented by h_{ij} ($1 \leq j \leq m$). Each genotype is represented by a string, with size m, over the alphabet $\{0, 1, 2\}$, and each site of the genotype g_i is represented by g_{ij}. Each genotype is explained by two haplotypes. A genotype g_i is explained by a pair of haplotypes (h_i^a,h_i^b) such that

$$g_{ij} = \begin{cases} h_{ij}^a \text{ if } h_{ij}^a = h_{ij}^b \\ 2 \quad \text{if } h_{ij}^a \neq h_{ij}^b \end{cases} . \tag{1}$$

A genotype site g_{ij} with either value 0 or 1 is a homozygous site (the same allele is inherited from both parents), whereas a site with value 2 is a heterozygous site (different alleles are inherited from each parent).

Definition 1. *Given a set $\mathcal{G}$ of n genotypes, each with size m, the haplotype inference problem consists in finding a set of haplotypes $\mathcal{H}$, such that each genotype $g_i \in \mathcal{G}$ is explained by two haplotypes $h_i^a, h_i^b \in \mathcal{H}$.*

For each genotype g with k heterozygous sites, there are 2^{k-1} pairs of haplotypes that can explain g. For example, genotype $g_i = 202$ can be explained either by haplotypes (000,101) or by haplotypes (001,100).

Most often genotyping procedures leave a percentage of missing data. To represent missing sites, the alphabet of the genotypes is extended to $\{0, 1, 2, ?\}$.

When the considered individuals are organized in pedigrees, additional information may be associated with the haplotype inference problem. Considering the Mendelian law of inheritance, every site in a single haplotype is inherited from a single parent, assuming no mutations within a pedigree. In a pedigree, an individual is a *founder* if he does not have

parents on the pedigree (and a *non-founder* if he has both parents on the pedigree). We assume that haplotype h^a is inherited from the father and h^b is inherited from the mother, thus breaking a symmetry on the pairs of haplotypes for non-founder individuals. However, a recombination may occur, where the two haplotypes of a parent get shuffled and the shuffled haplotype is passed on to the child. For example, suppose a father has the haplotype pair $(011, 100)$ and the haplotype that he passed on to his child is 111. Hence one recombination event must have occurred: haplotypes 011 and 100 have mixed together and originated a new haplotype $h = 111$. Although every site of the child's haplotype h was inherited from the father, the first site came from the paternal grandmother, while the second and third sites came from the paternal grandfather.

2.1 Minimum Recombinant Haplotype Configuration

Recombination events are rare in DNA regions with high linkage disequilibrium. Therefore, most rule-based haplotype inference methods for pedigrees assume no recombination among SNPs within each pedigree [19, 20, 13]. Although the assumption of no recombination is valid in many cases, this assumption can be violated even for some dense markers [9]. Therefore, the problem of minimizing the number of recombinations was suggested [6, 18]. The Minimum Recombinant Haplotype Configuration (MRHC) problem is a well-known approach to solve the haplotype inference problem in pedigrees. The MRHC problem is NP-hard [9, 10, 15].

Definition 2. *The Minimum Recombinant Haplotype Configuration (MRHC) problem aims at finding a haplotype inference solution for a pedigree which minimizes the number of required recombination events [6, 18].*

The PedPhase tool [9] implements an Integer Linear Programming (ILP) model for MRHC with missing alleles.

2.2 Haplotype Inference by Pure Parsimony

The Haplotype Inference by Pure Parsimony (HIPP) approach aims at finding a minimum-cardinality set of haplotypes $\mathcal{H}$ that can explain a given set of genotypes $\mathcal{G}$. The idea of searching for the solution with the smallest number of haplotypes is biologically motivated by the fact that individuals from the same population have the same ancestors and mutations do not occur often. Moreover, it is also well-known that the number of haplotypes in a population is much smaller than the number of genotypes. It has been shown that the HIPP problem is NP-hard [7].

Definition 3. *The haplotype inference by pure parsimony (HIPP) problem consists in finding a solution to the haplotype inference problem which minimizes the number of distinct haplotypes [5].*

RPoly [4] is a state-of-the-art solver implementing a 0-1 ILP model for solving the HIPP problem.

3 PedRPoly: Minimum Recombinant Maximum Parsimony

This section describes the PedRPoly model which aims at finding a haplotype inference solution for sets of pedigrees from the same population. The PedRPoly model is a combination of the MRHC PedPhase model [10] and the HIPP RPoly model [4].

Table 1. The PedRPoly Model: Minimum Recombinant Maximum Parsimony.

minimize:	$(2n+1) \times \sum_{non-founder\ i} \sum_{j=1}^{m-1} (r_{i\,j}^1 + r_{i\,j}^2) + \sum_{i=1}^{n} (u_i^a + u_i^b)$	
subject to:		
Equation	Constraint	Indexes
	Mendelian Law of Inheritance rules (Table 2)	
(2)	$-r_{i\,j}^l + g_{i\,j}^l - g_{i\,j+1}^l \leq 0$ $-r_{i\,j}^l - g_{i\,j}^l + g_{i\,j+1}^l \leq 0$	$l = 1,\ 2$ $1 \leq i \leq n$, i non-founder $1 \leq j \leq m$
(3)	$\neg(R \Leftrightarrow S) \Rightarrow x_{i\,k}^{p\,q}$ (Table 3)	$p, q \in \{a, b\}$ $1 \leq k < i \leq n$
(4)	$\sum_{k<i\,;\,q\in\{a,b\}} x_{i\,k}^{p\,q} - u_i^p \leq 2i - 3$	$1 < i \leq n$ $p \in \{a, b\}$

Definition 4. *The Minimum Recombinant Maximum Parsimony model aims at finding a haplotype inference solution which minimizes the number of recombination events within pedigrees and the number of distinct haplotypes used.*

Example 1. Consider two trios (father, mother and child), from two families A and B, with the following genotypes: $g_A^{father} = 102$, $g_A^{mother} = 222$, $g_A^{child} = 202$, $g_B^{father} = 211$, $g_B^{mother} = 202$ and $g_B^{child} = 222$. Consider the following haplotype inference solutions.

Solution 1: $h_A^{father} = (100, 101)$, $h_A^{mother} = (001, 110)$, $h_A^{child} = (100, 001)$, $h_B^{father} = (011, 111)$, $h_B^{mother} = (000, 101)$ and $h_B^{child} = (111, 000)$. Solution 1 is a 0-recombinant solution with 7 distinct haplotypes $(100, 101, 000, 111, 011, 001, 110)$.

Solution 2: $h_A^{father} = (101, 100)$, $h_A^{mother} = (000, 111)$ and $h_A^{child} = (101, 000)$, $h_B^{father} = (011, 111)$, $h_B^{mother} = (000, 101)$ and $h_B^{child} = (011, 100)$. Solution 2 is a 1-recombinant solution (there is one recombination event in family B) and uses 5 distinct haplotypes $(100, 101, 000, 111, 011)$.

Solution 3: $h_A^{father} = (101, 100)$, $h_A^{mother} = (000, 111)$ and $h_A^{child} = (101, 000)$, $h_B^{father} = (011, 111)$, $h_B^{mother} = (000, 101)$ and $h_B^{child} = (111, 000)$. Solution 3 is a 0-recombinant solution using 5 distinct haplotypes $(100, 101, 000, 111, 011)$.

Clearly, solution 3 is preferred to the other solutions. Solution 3 is both a MRHC and a HIPP solution. Consequently, solution 3 is a Minimum Recombinant Maximum Parsimony solution. If there exists no solution that minimizes both criteria, then preference is given to the MRHC criterion and hence, the MRHC solution which uses the smallest number of distinct haplotypes would be chosen.

PedRPoly is a 0-1 ILP model which combines the ILP PedPhase and the RPoly models. The constraints of the model are detailed in Table 1. Following the RPoly model, PedRPoly associates two haplotypes, h_i^a and h_i^b, with each genotype g_i, and these haplotypes are required to explain g_i. Moreover, PedRPoly associates a variable $t_{i\,j}$ with each heterozygous site $g_{i\,j}$, such that $t_{i\,j} = 1$ indicates that the mutant value was inherited from the father ($h_{i\,j}^a = 1$) and the wild value was inherited from the mother ($h_{i\,j}^b = 0$) whereas $t_{i\,j} = 0$ indicates that the wild value was inherited from the father ($h_{i\,j}^a = 0$) and the mutant value was

Table 2. Mendelian Law of Inheritance Rules (regarding variables g^1_{ij}). The constraints involving variables g^2_{ij} are defined similarly ($1 \leq i \leq n$, i non-founder, $1 \leq j \leq m$).$f(i)$ corresponds to the father of i.

Condition	Constraint
$g_{ij} = 0 \wedge g_{f(i)j} = 2$	$t_{f(i)j} \Leftrightarrow g^1_{ij}$
$g_{ij} = 0 \wedge g_{f(i)j} = ?$	$(g^1_{ij} \vee \neg t^a_{f(i)j}) \wedge (\neg g^1_{ij} \vee \neg t^b_{f(i)j})$
$g_{ij} = 1 \wedge g_{f(i)j} = 2$	$t_{f(i)j} \Leftrightarrow \neg g^1_{ij}$
$g_{ij} = 1 \wedge g_{f(i)j} = ?$	$(g^1_{ij} \vee t^a_{f(i)j}) \wedge (\neg g^1_{ij} \vee t^b_{f(i)j})$
$g_{ij} = 2 \wedge g_{f(i)j} = 0$	$\neg t_{ij}$
$g_{ij} = 2 \wedge g_{f(i)j} = 1$	t_{ij}
$g_{ij} = 2 \wedge g_{f(i)j} = 2$	$(g^1_{ij} \vee t_{ij} \vee \neg t_{f(i)j}) \wedge (g^1_{ij} \vee \neg t_{ij} \vee t_{f(i)j}) \wedge$ $(\neg g^1_{ij} \vee t_{ij} \vee t_{f(i)j}) \wedge (\neg g^1_{ij} \vee \neg t_{ij} \vee \neg t_{f(i)j})$
$g_{ij} = 2 \wedge g_{f(i)j} = ?$	$(g^1_{ij} \vee t_{ij} \vee \neg t^a_{f(i)j}) \wedge (g^1_{ij} \vee \neg t_{ij} \vee t^a_{f(i)j}) \wedge$ $(\neg g^1_{ij} \vee t_{ij} \vee \neg t^b_{f(i)j}) \wedge (\neg g^1_{ij} \vee \neg t_{ij} \vee t^b_{f(i)j})$
$g_{ij} = ? \wedge g_{f(i)j} = 0$	$\neg t^a_{ij}$
$g_{ij} = ? \wedge g_{f(i)j} = 1$	t^a_{ij}
$g_{ij} = ? \wedge g_{f(i)j} = 2$	$(g^1_{ij} \vee t^a_{ij} \vee \neg t_{f(i)j}) \wedge (g^1_{ij} \vee \neg t^a_{ij} \vee t_{f(i)j}) \wedge$ $(\neg g^1_{ij} \vee t^a_{ij} \vee t_{f(i)j}) \wedge (\neg g^1_{ij} \vee \neg t^a_{ij} \vee \neg t_{f(i)j})$
$g_{ij} = ? \wedge g_{f(i)j} = ?$	$(g^1_{ij} \vee t^a_{ij} \vee \neg t^a_{f(i)j}) \wedge (g^1_{ij} \vee \neg t^a_{ij} \vee t^a_{f(i)j}) \wedge$ $(\neg g^1_{ij} \vee t^a_{ij} \vee \neg t^b_{f(i)j}) \wedge (\neg g^1_{ij} \vee \neg t^a_{ij} \vee \neg t^b_{f(i)j})$

Table 3. Definition of predicates R and S, accordingly to index values.

Condition	Constraint
$g_{ij} \neq 2 \wedge g_{kj} = 2$	$R = (g_{ij} \Leftrightarrow (q \Leftrightarrow a))$ and $S = t_{kj}$
$g_{kj} \neq 2 \wedge g_{ij} = 2$	$R = (g_{kj} \Leftrightarrow (p \Leftrightarrow a))$ and $S = t_{ij}$
$g_{ij} = 2 \wedge g_{kj} = 2$	$R = (p \Leftrightarrow q)$ and $S = (t_{ij} \Leftrightarrow t_{kj})$
$g_{ij} = ? \wedge g_{kj} \notin \{2, ?\}$	$R = t^p_{ij}$ and $S = g_{kj}$
$g_{kj} = ? \wedge g_{ij} \notin \{2, ?\}$	$R = t^q_{kj}$ and $S = g_{ij}$
$g_{ij} = ? \wedge g_{kj} = 2$	$R = (q \Leftrightarrow a)$ and $S = (t^p_{ij} \Leftrightarrow t_{kj})$
$g_{kj} = ? \wedge g_{ij} = 2$	$R = (p \Leftrightarrow a)$ and $S = (t^q_{kj} \Leftrightarrow t_{ij})$
$g_{ij} = ? \wedge g_{kj} = ?$	$R = t^p_{ij}$ and $S = t^q_{kj}$

inherited from the mother ($h^b_{ij} = 1$). In addition, PedRPoly associates two variables with each missing site. Variable t^a_{ij} is associated with the paternal haplotype site h^a_{ij}, whereas variable t^b_{ij} is associated with the maternal haplotype site h^b_{ij}. The values of h^a_i and h^b_i at homozygous sites are implicitly assumed.

To analyze the recombination events within pedigrees, not only the paternal and maternal haplotypes are considered, but also the grand-paternal and grand-maternal origin of each allele in the haplotypes. Following the PedPhase MRHC model, for each non-founder individual i and site j, two variables are defined: g^1_{ij} and g^2_{ij}. The assignment $g^1_{ij} = 0$ ($g^1_{ij} = 1$) represents that the paternal allele of individual i at site j comes from the paternal grandfather (grandmother). In a similar way, $g^2_{ij} = 0$ ($g^2_{ij} = 1$) represents that the maternal allele of individual i at site j comes from the maternal grandfather (grandmother). Constraints to ensure that the Mendelian law of inheritance is satisfied are defined in Table 2. Note that PedRPoly only associates variables with heterozygous and missing sites (inspired by RPoly), while PedPhase associates variables with both homozygous and heterozygous sites. The new definition of variables associated with sites requires the redefinition of the constraints related with Mendelian laws. For instance, consider the first constraint of Ta-

ble 2, $t_{f(i)\,j} \Leftrightarrow g^1_{i\,j}$, for the case $g_{i\,j} = 0$ and $g_{f(i)\,j} = 2$. Clearly, if $t_{f(i)\,j} = 1$ (which represents that individual $f(i)$ has inherited value 1 from his father and value 0 from his mother) then $g^1_{i\,j} = 1$ (which represents that individual i must have inherited the value 0 from his paternal grandmother) and vice-versa.

Furthermore, variables are defined to count the number of recombinations. For each non-founder individual i, variable $r^1_{i\,j}$ ($r^2_{i\,j}$) is assigned value 1 if a recombination took place at site j, to create the paternal (maternal) haplotype of individual i. Thus, $r^l_{i\,j} = 1$ if $g^l_{i\,j} \neq g^l_{i\,j+1}$, for $l = 1, 2$ and $1 \leq j \leq m-1$, which is ensured by constraints (2). Here, a simplification to the original MRHC is considered. Actually, in the original model, $r^l_{i\,j} = 1$ *if and only if* $g^l_{i\,j} \neq g^l_{i\,j+1}$. Observe that an implication, instead of an equivalence, is sufficient for correctness and reduces in half the number of these constraints.

Furthermore, the model should be able to determine the number of distinct haplotypes used. Once more following the RPoly model, let $x^{p\,q}_{i\,k}$, with $p, q \in \{a, b\}$ and $1 \leq k < i \leq n$, be 1 if haplotype p of genotype g_i and haplotype q of genotype g_k are different. The conditions on the $x^{p\,q}_{i\,k}$ variables are based on the values of variables $t_{i\,j}$ and $t_{k\,j}$ for heterozygous sites and of variables $t^a_{i\,j}$, $t^b_{i\,j}$, $t^a_{k\,j}$ and $t^b_{k\,j}$ for missing sites, and are described by equations (3).

In addition, the model uses variables u to denote when one of the haplotypes, associated with a given genotype, is different from all previous haplotypes. Hence, u^p_i, with $p \in \{a, b\}$ and $1 \leq i \leq n$, is 1 if haplotype p of genotype g_i is different from all previous haplotypes. Then, the conditions on the u^p_i variables are based on the conditions for the $x^{p\,q}_{i\,k}$ variables, with $1 \leq k < i$ and $q \in \{a, b\}$ and correspond to equation (4).

Finally, the cost function minimizes the number of recombination events, which is given by the sum of variables r, and the number of distinct haplotypes used in the solution, which is given by the sum of variables u,

$$\textit{minimize } ((2n+1) \times \textstyle\sum_{non-founder\ i} \sum_{j=1}^{m-1} (r^1_{i\,j} + r^2_{i\,j})) + \sum_{i=1}^{n} (u^a_i + u^b_i).$$

A larger weight is given to the number of recombinations, because we want to guarantee that a minimum recombinant solution is chosen. Note that $2n$ is a trivial upper bound on the number of haplotypes in the solution, and therefore giving weight $2n + 1$ to the number of recombinations implies that a minimum recombinant solution is always preferred. The idea of giving more weight to the number of recombinations is biological motivated by the fact that recombination events within haplotypes are rare. Moreover, the number of recombination events is related with the number of distinct haplotypes. In fact, a larger number of recombination events suggests a larger number of haplotypes. In general, a recombination event generates a new haplotype, whereas if no recombination occurs, then the haplotypes of the child are exact copies of the parents haplotypes.

4 Experimental Results

The experimental data was simulated using the SimPed program [8]. SimPed generates haplotypes for families, given the pedigree structure, as well as the haplotypes and their frequencies for founders. Three different sets of haplotypes were considered. These sets of haplotypes are real data for which haplotypes have been experimentally identified [1, 17], and correspond to the A, B and C data sets used in [2]. Haplotypes in set A have 9 SNPs, haplotypes in set B have 5 SNPs and haplotypes in set C have 17 SNPs. Three different pedigree structures, taken from [10], were considered: pedigree 1 with 15 individuals, pedigree 2 with 29 individuals and pedigree 3 with 17 individuals (with a mating loop). Each

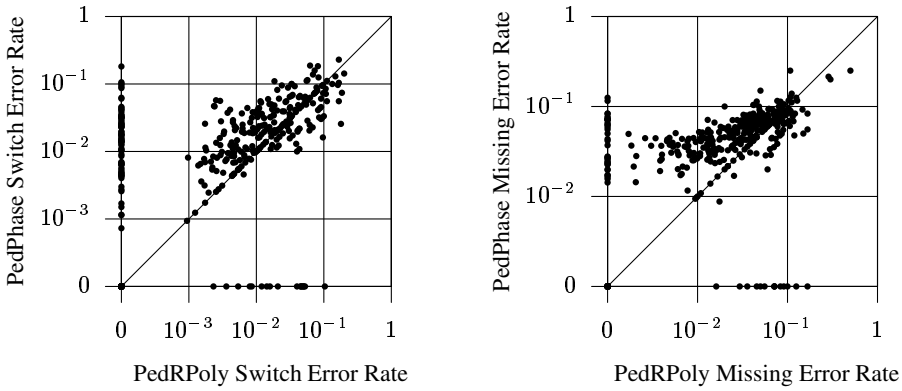

Fig. 1. Comparison of PedRPoly and PedPhase error rates

simulated instance consists of 10 replicates of the given pedigree, simulating 10 different families from the same population. Recombination events between alleles were considered with probabilities 0.1%, 0.5% and 1%. Three variations on missing rates were considered: 1%, 10% and 20%. For each combination of parameters, 5 independent replicates were selected, resulting in a total of 405 ($= 3^4 \times 5$) input trials. MiniSat+ [3] was used as a 0-1 ILP tool to solve the PedRPoly model.

In order to analyze the accuracy of the methods, two different errors were considered. The *switch error rate* measures the percentage of possible switches in haplotype orientation, used to recover the correct phase in an individual [14]. Missing alleles are not considered for computing the switch error. The *missing error rate* is the percentage of incorrectly inferred missing data [16].

Figure 1 presents two scatter plots comparing the switch error and missing error rates for PedRPoly and PedPhase. Each problem instance corresponds to a point in the plot, where the x-axis represents the error rate of the PedRPoly approach and the y-axis represents the error rate of the ILP PedPhase approach.

The switch error rate of the methods is compared in the left plot of Figure 1. The switch error of PedRPoly is smaller than the switch error of PedPhase for 55.3% of the problem instances. The switch error of PedRPoly is larger than the switch error of PedPhase for 16.8% of the problem instances, and for the remaining 27.9% the error is the same for both methods.

With respect to the missing error rate (right plot of Figure 1), it is clear that PedRPoly is more accurate than PedPhase. Indeed, the missing error of PedRPoly is smaller than the missing error of PedPhase for 64.7% of the problem instances. The missing error of PedRPoly is larger than the missing error of PedPhase for 18% of the problem instances, and for the remaining 17.3% instances the error is the same for both methods.

Table 4 presents the accuracy results organized by parameter value. Each value is the average of the error rate for the 135 instances generated with the corresponding parameter value. We conclude that PedRPoly has a smaller (switch and missing) error rate on every class of instances.

In addition, the HIPP solver, RPoly, has also been tested. RPoly does not take into consideration the pedigree information, and therefore, has in general higher error rates which can go up to 50% in some cases. Moreover, RPoly is not able to solve around 25% of the instances within a time out of 10000 seconds (in particular the instances with higher missing rates).

Table 4. Switch Error Rate and Missing Error Rate for PedRPoly and PedPhase in sets of instances with different parameters (n is the number of genotypes of the instance, with $n = 10 \cdot f$ where f is the size of each pedigree, and m is the number of sites of each genotype).

		Error Rate			
Set		**Switch Error Rate**		**Missing Error Rate**	
		PedRPoly	**PedPhase**	**PedRPoly**	**PedPhase**
Missing Rate	1%	0.0115	0.0143	0.0361	0.0411
	10%	0.0190	0.0232	0.0382	0.0488
	20%	0.0304	0.0437	0.0456	0.0625
Recombination Rate	0.1%	0.0157	0.0221	0.0393	0.0472
	0.5%	0.0212	0.0267	0.0398	0.0506
	1%	0.0240	0.0324	0.0407	0.0546
Pedigree	Ped1 (n=150)	0.0194	0.0245	0.0428	0.0469
	Ped2 (n=290)	0.0199	0.0284	0.0355	0.0471
	Ped3 (n=170)	0.0217	0.0283	0.0415	0.0584
Population	A (m=9)	0.0116	0.0210	0.0428	0.0469
	B (m=5)	0.0450	0.0482	0.0355	0.0471
	C (m=17)	0.0044	0.0120	0.0415	0.0584

Finally, we have compared the number of distinct haplotypes in the PedRPoly solution and the PedPhase solution with the number of haplotypes in the real solution. PedRPoly has the same number of haplotypes as the real solution for 64.7% of the instances and for 96.8% of the instances the number of haplotypes in the PedRPoly solution differs by less than 3 haplotypes from the number in the real solution. PedPhase solutions are less similar to the real solutions with respect to the number of distinct haplotypes. For only 23.7% of the instances, the number of haplotypes in the PedPhase solution is equal to the number of haplotypes in the real solution, and for 50% of the instances, the number of haplotypes in the PedRPoly solution differs by more than 2 haplotypes from the real solution.

Regarding the efficiency of PedRPoly and PedPhase methods, while PedPhase is able to solve each instance in a few seconds, PedRPoly can take a few hours. Improving the efficiency of PedRPoly is the main short term goal.

5 Conclusions and Future Work

This paper presents a new method for inferring haplotypes from genotype data of families from the same population. The proposed method (called PedRPoly) integrates the minimum recombinant and the pure parsimony principles, two relevant constraint based haplotyping approaches. Thus, PedRPoly can take into account both the family information and the population information. Experimental results show that PedRPoly is actually more accurate than PedPhase which only uses the minimum recombinant principle.

Future work directions include improving the efficiency of the PedRPoly method and testing the method in larger and real data sets.

Acknowledgments

This work is partially supported by Fundação para a Ciência e Tecnologia under research project SHIPs (PTDC/EIA/64164/2006) and PhD grant SFRH/BD/28599/2006.

References

1. A. M. Andrés, A. G. Clark, L. Shimmin, E. Boerwinkle, C. F. Sing, and J. E. Hixson. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genetic Epidemiology*, 31(7):659–671, 2007.
2. S. Climer, G. Jäger, A. Templeton, and W. Zhang. How frugal is mother nature with haplotypes? *Bioinformatics*, 25(1):68–74, 2009.
3. N. Eén and N. Sörensson. Translating pseudo-Boolean constraints into SAT. *Journal on Satisfiability, Boolean Modeling and Computation*, 2:1–26, 2006.
4. A. Graça, J. Marques-Silva, I. Lynce, and A. Oliveira. Efficient haplotype inference with pseudo-Boolean optimization. In *Algebraic Biology (AB'07)*, pages 125–139, 2007.
5. D. Gusfield. Haplotype inference by pure parsimony. In *Symposium on Combinatorial Pattern Matching (CPM'03)*, pages 144–155, 2003.
6. J. L. Haines. Chromlook: an interactive program for error detection and mapping in reference linkage data. *Genomics*, 14(2):517–519, 1992.
7. G. Lancia, C. M. Pinotti, and R. Rizzi. Haplotyping populations by pure parsimony: complexity of exact and approximation algorithms. *INFORMS Journal on Computing*, 16(4):348–359, 2004.
8. S. M. Leal, K. Yan, and B. Mller-Myhsok. SimPed: A simulation program to generate haplotype and genotype data for pedigree structures. *Human Heredity*, 60(2):119–122, 2005.
9. J. Li and T. Jiang. Efficient inference of haplotypes from genotypes on a pedigree. *Journal of Bioinformatics and Computational Biology*, 1(1):41–69, 2003.
10. J. Li and T. Jiang. Efficient rule-based haplotyping algorithms for pedigree data. In *International Conference on Research in Computational Molecular Biology (RECOMB'03)*, pages 197–206, 2003.
11. J. Li and T. Jiang. Computing the minimum recombinant haplotype configuration from incomplete genotype data on a pedigree by integer linear programming. *Journal of Computational Biology*, 12(6):719–739, 2005.
12. X. Li and J. Li. Comparison of haplotyping methods using families and unrelated individuals on simulated rheumatoid arthritis data. In *BMC Proceedings*, pages S1–S55, 2006.
13. X. Li and J. Li. Efficient haplotype inference from pedigree with missing data using linear systems with disjoint-set data structures. In *International Conference on Computational Systems Bioinformatics (CSB'08)*, pages 297–307, 2008.
14. S. Lin, A. Chakravarti, and D. J. Cutler. Haplotype and missing data inference in nuclear families. *Genome Research*, 14(8):1624–1632, 2004.
15. L. Liu, C. Xi, J. Xiao, and T. Jiang. Complexity and approximation of the minimum recombinant haplotype configuration problem. *Theoretical Computer Science*, 378(3):316–330, 2007.
16. J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. Qin, H. Munro, G. Abecassis, P. Donnelly, and International HapMap Consortium. A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics*, 78(3):437–450, 2006.
17. S. H. Orzack, D. Gusfield, J. Olson, S. Nesbitt, L. Subrahmanyan, and J. V. P. Stanton. Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes. *Genetics*, 165(2):915–928, 2003.
18. D. Qian and L. Beckmann. Minimum-recombinant haplotyping in pedigrees. *American Journal of Human Genetics*, 70(6):1434–1445, 2002.
19. E. M. Wijsman. A deductive method of haplotype analysis in pedigrees. *American Journal of Human Genetics*, 41(3):356–373, 1987.
20. K. Zhang, F. Sun, and H. Zhao. HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics*, 21(1):90–103, 2005.