ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

P. Bertolazzi, A. Godi, M. Labbè, L. Tininini

SOLVING HAPLOTYPING INFERENCE PARSIMONY PROBLEM USING A NEW BASIC POLYNOMIAL FORMULATION

R. 636 2006

Paola Bertolazzi – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: bertola@iasi.cnr.it.

Alessandra Godi – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: godi@iasi.cnr.it.

Martine Labbè – Institut de Statistique et de Recherche Opérationnelle, Université Libre de Bruxelles, Belgium. EMAIL:mlabbe@ulb.ac.be.

Leonardo Tininini – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: tininini@iasi.cnr.it.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

Similarity and diversity among individuals of the same species are expressed in small DNA variations called Single Nucleotide Polymorphism. The knowledge of SNP phase gives rise to the haplotyping problem that in the parsimonious version states to infer the minimum number of haplotypes from a given set of genotype data. ILP technique represents a good resolution strategy for this interesting combinatorial problem whose main limit lies in its NP-hardness. In this paper we present a new polynomial model for the haplotyping inference by parsimony problem characterized by the original use of a maximum formulation jointly with a good heuristic solution. This approach showed to be a robust basic model that can be used as starting point for more sophisticated ILP techniques like branch and cut and polyhedral studies.

1. Introduction

Most vegetal and animal cells are *diploid*, i.e., they have two similar, but not exactly identical, versions (or copies) of each chromosome (*homologous chromosomes*). In general, individuals from the same species are genetically "very similar", as for instance humans: the DNA between two random people is about 99.9% identical. The individual uniqueness lies in a small number of bases that can exist where single base DNA differences occur. Thus a *SNP (Single Nucleotide Polymorphism)* is a single base pair position in genomic DNA at which different nucleotide variants (*alleles*) exist. In humans, SNPs are almost always *biallelic*, that is, only two of the four possible polymorphisms at each site are possible. The knowledge of these two variants is referred to as the *phase* of the SNP. The sequence of alleles along a chromosome copy is called a *haplotype*. Instead, the SNP information of the bases pairs sequence at each site of each chromosome is called a *genotype*, but it does not specify which base (i.e., which allele) occurs on which chromosome. For a given set of SNPs, an individual possesses two haplotypes, one inherited from the paternal genome and the other from the maternal genome and exactly one genotype associated with the chromosome pair. The inheritance process is complicated by a phenomenon known as *recombination* which concerns portion exchanges of the paternal and maternal chromosomes.
A SNP site where both haplotypes have the same variant (nucleotide) is called a *homozygous site*; a SNP site where the haplotypes have different variants is called a *heterozygous site*. Thus, while in haplotype data alleles are completely known, in genotype data the nucleotide variants at homozygous sites are known but the information regarding which heterozygous site SNP variants came from the same chromosome copy is unknown. The determination of the haplotypes within a population is essential. For instance, haplotypes are necessary in evolutionary studies to extract the information needed to detect diseases and to reduce the number of tests to be carried out, in the discovery of a functional gene or in the study of an altered response of an organism to a particular terapy. In human pharmacogenetics, haplotype-SNPs seem to explain why people react differently to different types or amounts of drugs. Indeed, since SNPs can affect the structure and function of proteins and enzymes, they can influence how efficiently a drug is absorbed and metabolized. Unfortunately, experimental techniques to obtain the haplotypes of an individual are very expensive, time consuming and labor intensive. However, it is possible to determine the genotype of an individual quickly and easy. The use of computational techniques joint with specific biological models offers a way of defining the haplotypes from the genotype data (i.e., *Haplotyping or Haplotype Inference (HI)*).

The HI problem has an interesting version with the requirement that the number of inferred haplotypes is minimum. The problem of finding the smallest collection of haplotypes that can explain the genotypic information (a set of input genotypes) of the current population is called the *Haplotype Inference by (maximum) Parsimony (HIP)* problem. Scientists often claim that the *parsimony*[1] of a theory is relevant to decide whether the theory is true, or approximately true, or would make accurate predictions because it is a natural criterion for choosing a solution in many domains. This is particularly true for haplotyping, since the number of distinct haplotypes observed in a population is much smaller than the number of possible haplotypes. Parsimony principle does not erroneously state that haplotypes with high frequency in a population should be preferred in a haplotype reconstruction (in fact, parsimony is affected by haplotype frequencies only in the weakest sense) but means that haplotypes of a population can not be so different from each other, as supported by real data from the practice and by phylogenetic haplotype tree history.

In this paper we describe a new ILP polynomial formulation for HIP problem. This model is first naturally formulated as a minimum problem; then, by using the upper bound from the new heuristic COLLHAPS [1] based on a generalization of Clark's rule [2], it is turned into a maximization problem: fixed the number of generating haplotypes (from the heuristic), find those haplotypes that maximize the number of explained input genotypes. The final model is the result of a strengthening study based on clique and symmetry-breaking inequalities and by dominance relations. Computational experience

[1]The principle of parsimony is also known as *Ockham's razor principle*, named for William of Ockham, the medieval philosopher who said that plurality is not to be assumed without necessity and that what can be done with fewer assumptions is done in vain with more.

shows that our model has good performances respect to the existing "basic" (without any cut addition) polynomial ILP model for HIP known in literature [3].

2. Related works

The HI problem has been studied since nineties and a wide variety of techniques (statistical and combinatorial methods) have been proposed.
Statistical approaches try to iteratively determine the haplotype frequencies, and then infer the haplotype-pairs. In the methods based on expectation-maximization (EM) the haplotype frequency estimates are iteratively updated, starting from an initial guess and trying to maximize a likelihood function [4, 5, 6]. Other statistical methods are based on Bayesian inference and on the adoption of a more or less biologically-based prior, so as to get more accurate estimates of the haplotype frequencies and consequently of the genotype recontructions [7, 8, 9].
Combinatorial methods are mostly inspired to the Clark's "inference rule" [2], based on the principle that, given a genotype and a haplotype compatible with this genotype, the other haplotype can be inferred simply by "difference" between the genotype and the given haplotype. Clark's rule was applied directly giving rise to the first algorithm for haplotyping. The algorithm has good accuracy but two major drawbacks: it could not even start or it can resolve only a subset of the given genotypes.

A second step is due to Gusfield who first used integer programming for haplotyping problem and formulated two different optimization problems: the former [10] looks for the best sequence of application of the Clark's rule to solve the maximum number of genotypes; the latter [11] is the formulation of the haplotyping inference by parsimony problem. Hubbell [12] showed that the latter version of the problem is, in general, NP-hard by a reduction from the minimum clique cover problem. Recently, Lancia et al. [13] showed that the problem is APX-hard and provided a 2^{k-1}-approximation algorithm for data sets in which each genotype has at most k ambiguous positions[2]. Integer Linear Programming was employed also in other works [3, 13] to find the most parsimonious phasing: Brown et al. [3] proposed another model to solve the HIP problem via ILP. The most relevant difference between Gusfiled's and Brown's models concerns the dimensions: the first model presents an exponential-size formulation whereas the second has polynomial dimensions.
Concerning the *heuristic approaches* for HIP problem, Bayesian inference models provide implicit notions of parsimony, via the implicit "Ockham factor" of the Bayesian formalism [8, 14]. On the other hand, several non-statistical-based approximation algorithms exist: besides a first one with performance guarantee 2^{k-1} [13] for the case in which each genotype has at most k heterozygous positions, Lancia and Rizzi [15] recently presented a polynomial time algorithm for the HIP problem when each genotype has at most two heterozygous positions.

3. The model as minimum problem

Let $G = \{g_1, ..., g_m\}$ be the set of input genotypes represented by an n-dimensional vector (the number of SNPs) where each component $g_i(p) \in \{0, 1, 2\}$, for each $i \in \{1, ..., m\}$ and for each $p \in \{1, ..., n\}$: 0 and 1 are related to homozygous sites, while heterozygous sites are denoted by 2. Thus, an instance of the Haplotype Inference Problem is represented by an $m \times n$ matrix G such that each row g_i is an n-dimensional genotype. The output is a $2m \times n$ binary matrix with a minimum number of distinct rows where each row is a haplotype h_i, i.e., a binary vector of length n and for $p \in \{1, ..., n\}$:

$$g_i(p) = h_{2i-1}(p) \oplus h_{2i}(p) \qquad (i = 1, \ldots, m)$$

where the conflate operator $\oplus : \{0, 1\} \rightarrow \{0, 1, 2\}$ is defined as follows:

$$\begin{cases} 0 \oplus 0 = 0 \\ 0 \oplus 1 = 1 \oplus 0 = 2 \\ 1 \oplus 1 = 1. \end{cases}$$

[2] the algorithm is tightly connected to the ILP formulation by Gusfield [11]

Let UB denote an upper bound on the number of haplotypes needed to generate the input of genotypes under the maximum parsimony hypothesis (as initial value, we consider $UB = 2m$). Denote by I_{UB} the set of indices for the unknown haplotypes: $I_{UB} = \{1, ..., UB\}$ and let h_i, with $i \in I_{UB}$, indicate a generic haplotype. Like in the polynomial formulation [3], we are not interested in enumerating all possible explaining haplotype-pairs for each genotype, but we want to introduce instead binary variables representing the solutions for the problem.
Thus let us define the following 0-1 VARIABLES for haplotypes, pairs and positions:

$$x_i = \begin{cases} 1 & \text{if haplotype } h_i \text{ is chosen in a solution} \\ 0 & \text{otherwise;} \end{cases}$$

for all $i \in I_{UB}$;

$$y^k_{\{i,j\}} = \begin{cases} 1 & \text{if the haplotype-pair indexed by } \{i,j\} \text{ is chosen to explain genotype } g_k \\ 0 & \text{otherwise;} \end{cases}$$

for all $k \in G$ and $i, j \in I_{UB}$, $i \neq j$;

$$z_{i,p} = \begin{cases} 1 & \text{if 1 is the value of haplotype } h_i \text{ in position } p \\ 0 & \text{if 0 is the value of haplotype } h_i \text{ in position } p \end{cases}$$

for all $i \in I_{UB}$, $p \in \{1, ..., n\}$. In the following, we identify the haplotype h_i with its index i.
We now need to define CONSTRAINTS ensuring that solutions to the integer problem properly explain every genotype:

$$\sum_{\{i,j\}} y^k_{\{i,j\}} \geq 1 \qquad \forall g_k \in G; \tag{1}$$

this is equivalent to the "covering" constraint in Gusfield's formulation [11]: each genotype must be covered by at least one haplotype-pair;

$$\sum_{j \neq i} y^k_{\{i,j\}} \leq x_i \qquad \forall g_k \in G, \ \ \forall i \in I_{UB}; \tag{2}$$

even this family of inequalities has the same meaning of the "activation" constraints in the exponential model by Gusfield [11]: since a pair $\{i, j\}$ unambiguously determines a genotype, we can introduce here the sum; the constraint means that if we select the pair of g_k containing haplotype i, then we must also choose haplotype i;

$$z_{i,p} + \sum_{j \neq i} y^k_{\{i,j\}} \leq x_i \qquad \begin{array}{c} \forall\ g_k \in G, \forall\ i \in I_{UB} \\ \forall\ p\ \in \{1, ..., n\}\ s.t.\ g_k(p) = 0 \end{array} \tag{3}$$

establishes the relation among z, y and x variables for each position where genotypes have value 0;

$$z_{i,p} \geq \sum_{j \neq i} y^k_{\{i,j\}} \qquad \begin{array}{c} \forall\ g_k \in G, \forall i \in I_{UB} \\ \forall\ p\ \in \{1, ..., n\}\ s.t.\ g_k(p) = 1 \end{array} \tag{4}$$

establishes the relation among z and y variables for each position where genotypes have value 1;

$$z_{i,p} + z_{j,p} \geq y^k_{\{i,j\}} \qquad \begin{array}{l} \forall\ g_k \in G, \forall i,j \in I_{UB}, i \neq j \\ \forall p \in \{1,...,n\}\ s.t.\ g_k(p) = 2 \end{array} \tag{5}$$

this constraint gives a lower bound on z variables for positions where genotypes have value 2: if $\{i,j\}$ is chosen, then $z_{i,p} = 1$ and $z_{j,p} = 0$, or viceversa;

$$z_{i,p} + z_{j,p} \leq x_i + x_j - y^k_{\{i,j\}} \qquad \begin{array}{l} \forall\ g_k \in G, \forall\ i,j \in I_{UB},\ i \neq j \\ \forall\ p\ \in \{1,...,n\}\ s.t.\ g_k(p) = 2 \end{array} \tag{6}$$

this constraint gives an upper bound on z variables for positions where genotypes have value 2;
The parsimony condition leads to the following OBJECTIVE FUNCTION:

$$\min \sum_{i \in I_{UB}} x_i. \tag{7}$$

Then, summarizing, from (1) to (7), we get the first basic formulation as follows:

P_{min} :

$$\min \textstyle\sum_{i \in I_{UB}} x_i$$

$$\textstyle\sum_{\{i,j\}} y^k_{\{i,j\}} \geq 1 \qquad \forall\ g_k \in G$$

$$\textstyle\sum_{j \neq i} y^k_{\{i,j\}} \leq x_i \qquad \forall\ g_k \in G,\ \ \forall\ i \in I_{UB}$$

$$z_{i,p} + \textstyle\sum_{j \neq i} y^k_{\{i,j\}} \leq x_i \qquad \begin{array}{l} \forall\ g_k \in G,\ \ \forall\ i \in I_{UB} \\ \forall p \in \{1,...,n\}\ s.t.\ g_k(p) = 0 \end{array}$$

$$z_{i,p} \geq \textstyle\sum_{j \neq i} y^k_{\{i,j\}} \qquad \begin{array}{l} \forall\ g_k \in G,\ \ \forall\ i \in I_{UB} \\ \forall p \in \{1,...,n\}\ s.t.\ g_k(p) = 1 \end{array}$$

$$z_{i,p} + z_{j,p} \geq y^k_{\{i,j\}} \qquad \begin{array}{l} \forall\ g_k \in G, \forall\ i,j \in I_{UB},\ i \neq j \\ \forall\ p \in \{1,...,n\}\ s.t.\ g_k(p) = 2, \end{array}$$

$$z_{i,p} + z_{j,p} \leq x_i + x_j - y^k_{\{i,j\}} \qquad \begin{array}{l} \forall\ g_k \in G, \forall\ i,j \in I_{UB},\ i \neq j \\ \forall\ p \in \{1,...,n\}\ s.t.\ g_k(p) = 2 \end{array}$$

$$x_i,\ z_{i,p} \in \{0,1\} \qquad \begin{array}{l} \forall i\ \in I_{UB} \\ \forall\ p \in \{1,...,n\} \end{array}$$

$$y^k_{\{i,j\}} \in \{0,1\} \qquad \begin{array}{l} \forall\ g_k \in G \\ \forall\ i,j \in I_{UB},\ i \neq j. \end{array}$$

Concerning the size of P_{min}, we have:

- UB x variables, $\frac{UB(UB-1)}{2}$ y variables and $n \cdot UB$ z variables;
- m constraints of kind (1);
- $m \cdot UB$ constraints of kind (2);
- $m \cdot UB \cdot n_0$ constraints of the form (3), where n_0 is the number of positions which are 0 in the genotype input;
- $m \cdot UB \cdot n_1$ constraints of the form (4), where n_1 is the number of positions which are 1 in the genotype input;
- $2 \cdot m \cdot UB \cdot n_2$ constraints of the form (5) and (6), where n_2 is the number of positions which are 2 in the genotype input.

Thus the number of variables and constraints is polynomial in the input size.

4. A good UB from *COLLHAPS*

The choice of $2m$ haplotypes as upper bound for the previous model is, of course, feasible; but it is possible to do better than this by using a good heuristic solution of HIP problem. In particular, we start with the solution from the new rule-based heuristic COLLHAPS [1] that we are briefly illustrating in the following.

Let us consider, as before, a set of genotypes $G = \{g_1, ..., g_m\}$. Let us call the haplotype-pair h^*_{2i-1} and h^*_{2i} the *genotype solution* of g_i.
Now given a set of genotypes, we associate a distinct variable w_p to each '2' in it. For each genotype g_i two *symbolic haplotypes* h_{2i-1} and h_{2i} are derived, and, for each position $j \in \{1, ..., n\}$, they are defined by:

$$h_{2i-1}(j) = \begin{cases} 0 & \text{if } g_i(j) = 0 \\ 1 & \text{if } g_i(j) = 1 \\ w_p & \text{if } g_i(j) = 2 \end{cases} \qquad h_{2i}(j) = \begin{cases} 0 & \text{if } g_i(j) = 0 \\ 1 & \text{if } g_i(j) = 1 \\ \bar{w}_p & \text{if } g_i(j) = 2 \end{cases}$$

where the variables w_p take values from $\{0, 1\}$ and $\bar{w}_p$ is a shorthand for $1 - w_p$ (obviously $\bar{\bar{w}}_p = w_p$). The variables w_p and $\bar{w}_p$ are called *complementary* and the $2m \times n$ matrix H, whose rows are the symbolic haplotypes h_i, is called *symbolic solution*. According to the maximum parsimony principle we want to determine a variable assignment for the variables w_p such that the resulting number of distinct haplotypes is minimum.

For example, let us consider the following HIP-instance constituted by the three genotypes: 1022, 2220, 2202. A symbolic solution is given by the three pairs of symbolic haplotpyes:

$$\begin{array}{lcccc} h_1: & 1 & 0 & w_1 & w_2 \\ h_2: & 1 & 0 & \bar{w}_1 & \bar{w}_2 \\ h_3: & w_3 & w_4 & w_5 & 0 \\ h_4: & \bar{w}_3 & \bar{w}_4 & \bar{w}_5 & 0 \\ h_5: & w_6 & w_7 & 0 & w_8 \\ h_6: & \bar{w}_6 & \bar{w}_7 & 0 & \bar{w}_8 \end{array}$$

A candidate solution is an assignment for the variables $w_1, \ldots, w_8$. In particular, the following variable assignment: w_1=0, w_2=0, w_3=1, w_4=0, w_5=0, w_6=1, w_7=0, w_8=0 corresponds to a candidate solution with 4 haplotypes, namely 1000 $\{h^*_1, h^*_3, h^*_5\}$, 1001 $\{h^*_2\}$, 0110 $\{h^*_4\}$ and 0101 $\{h^*_6\}$:

$$\begin{aligned} g_1 = 1022 &= h^*_1 \oplus h^*_2 = 1000 \oplus 1011 \\ g_2 = 2220 &= h^*_3 \oplus h^*_4 = 1000 \oplus 0110 \\ g_3 = 2202 &= h^*_5 \oplus h^*_6 = 1000 \oplus 0101 \end{aligned}$$

while the variable assignment: w_1=0, w_2=1, w_3=1, w_4=0, w_5=1, w_6=0, w_7=1, w_8=0 corresponds to an (optimal) solution with 3 haplotypes, namely 1001 $\{h_1^*, h_6^*\}$, 1010 $\{h_2^*, h_3^*\}$ and 0100 $\{h_4^*, h_5^*\}$:

$$g_1 = 1022 = h_1^* \oplus h_2^* = 1001 \oplus 1010$$
$$g_2 = 2220 = h_3^* \oplus h_4^* = 1010 \oplus 0100$$
$$g_3 = 2202 = h_5^* \oplus h_6^* = 0100 \oplus 1001$$

With the previous notation, a *collapse rule* corresponds to the minimum set of variable assignments, which forces the equality of two symbolic haplotypes. A prerequisite for the application of a collapse rule to a pair of symbolic haplotypes h' and h'' is their compatibility. More formally, two k-dimensional symbolic haplotypes h' and h'' are *compatible (for collapse)* iff for each $j \in \{1, \ldots, k\}$ one of the following holds:

- $h'_j = h''_j = 0$
- $h'_j = h''_j = 1$
- either h'_j or h''_j is a variable (but not both)
- both h'_j and h''_j are variables and not complementary

Let us consider now two compatible symbolic haplotypes h', h''; the *collapse assignment* (for h', h'') is the variable assignment ϑ defined as follows:

- if $h'_j = w_p$ and h''_j is a constant $c \in \{0/1\}$ then $\vartheta(w_p) = c$
- if $h'_j = \bar{w}_p$ and h''_j is a constant c (0/1) then $\vartheta(w_p) = 1 - c$
- if h'_j is a constant c (0/1) and $h''_j = w_q$ then $\vartheta(w_q) = c$
- if h'_j is a constant c (0/1) and $h''_j = \bar{w}_q$ then $\vartheta(w_q) = 1 - c$
- if $h'_j = w_p$ and $h''_j = w_q$ then $\vartheta(w_q) = w_p$
- if $h'_j = \bar{w}_p$ and $h''_j = \bar{w}_q$ then $\vartheta(w_q) = w_p$
- if $h'_j = w_p$ and $h''_j = \bar{w}_q$ then $\vartheta(w_q) = \bar{w}_p$
- if $h'_j = \bar{w}_p$ and $h''_j = w_q$ then $\vartheta(w_q) = \bar{w}_p$
- ϑ is the identity for any other variable

Given a matrix of symbolic haplotypes H and a pair of compatible symbolic haplotypes $h', h'' \in H$, the application of a collapse rule for h', h'' on H is the matrix obtained by applying the collapse assignment for h', h'' on all symbolic haplotypes in H.
Let us remark that given one optimal (in terms of parsimony) solution for a given instance of genotypes, there always exists a collapse rule application sequence, which produces a set of optimal solutions including at least the given one: this is a fundamental property of the collapse rule.

This is basically the idea of COLLHAPS algorithm and the underlying model. Other sophisticated techniques for the solution's progressive improvement have been implemented and can be found in [1].

5. Turning P_{min} into a maximization problem

We could certainly solve directly P_{min} getting a final solution as a solution for HIP. But we can do more than this turning P_{min} into a maximization problem on $y^k_{\{i,j\}}$ and $z_{i,p}$ variables, using as upper bound on the optimal solution for the HIP problem, the heuristic value from COLLHAPS, z^*_{HEUR}, and removing the x variables [16].
Let us define $UB_{max} = z^*_{HEUR} - 1$, fix the $UB := UB_{max}$, set $x_i = 1$ for $i = 1, ..., UB$, and consider the

following problem in which we substitute x variables with the assigned values and adapt constraints (1) and (2) to this maximization model:

$$P_{max}^{UB}:$$

$$\max \sum_{g_k \in G} \sum_{\{i,j\}} y^k_{\{i,j\}} \tag{8}$$

$$\sum_{\{i,j\}} y^k_{\{i,j\}} \leq 1 \qquad \forall\ g_k \in G \tag{9}$$

$$\sum_{g_k \in G} y^k_{\{i,j\}} \leq 1 \qquad \forall\ \{i,j\}, i,j \in I_{UB} \tag{10}$$

$$z_{i,p} + \sum_{j \neq i} y^k_{\{i,j\}} \leq 1 \qquad \forall\ g_k \in G, \forall\ i \in I_{UB},\ \forall\ p \in \{1,...,n\}\ s.t.\ g_k(p) = 0 \tag{11}$$

$$z_{i,p} \geq \sum_{j \neq i} y^k_{\{i,j\}} \qquad \forall\ g_k \in G, \forall\ i \in I_{UB},\ \forall\ p \in \{1,...,n\}\ s.t.\ g_k(p) = 1 \tag{12}$$

$$z_{i,p} + z_{j,p} \geq y^k_{\{i,j\}} \qquad \forall\ g_k \in G, \forall\ i,j \in I_{UB},\ i \neq j,\ \forall\ p \in \{1,...,n\}\ s.t.\ g_k(p) = 2 \tag{13}$$

$$z_{i,p} + z_{j,p} \leq 2 - y^k_{\{i,j\}} \qquad \forall\ g_k \in G, \forall\ i,j \in I_{UB},\ i \neq j,\ \forall\ p \in \{1,...,n\}\ s.t.\ g_k(p) = 2 \tag{14}$$

$$z_{i,p} \in \{0,1\} \qquad \forall i \in I_{UB},\ \forall\ p \in \{1,...,n\}$$

$$y^k_{\{i,j\}} \in \{0,1\} \qquad \forall\ g_k \in G,\ \forall\ i,j\ \in I_{UB},\ i \neq j.$$

The idea is that given the "good" value (i.e., a good number of haplotypes) from heuristic COLLHAPS, we can formulate a problem where we try to cover all genotypes with this number of haplotypes.
In other words, in order to solve the original problem P_{min}, we can work as follows:

1. Define P_{max}^{UB} as described above.
2. Solve P_{max}^{UB}. Let $opt_{P_{max}^{UB}}$ the integer optimal solution of P_{max}^{UB}.
3. If $opt_{P_{max}^{UB}} < m$, z^*_{HEUR} is the optimal solution of P_{min}. STOP.
4. Otherwise, set $UB := UB - 1$ and go to step 1.

Since typically ILP problems are solved by exhaustive enumeration techniques (like $B\&B$), if the value of linear relaxation of P_{max}^{UB} [3], let us say $LP_{P_{max}^{UB}}$, is such that $LP_{P_{max}^{UB}} < z^*_{HEUR}$, we could stop the procedure and conclude that z^*_{HEUR} is the optimal solution of P_{min} already at root node, before starting the branching procedure.

Thus, solving P_{max}^{UB} represents an alternative and in several cases very efficient way to get an optimal solution of P_{min}.

[3] which represents an upper bound on $opt_{P_{max}^{UB}}$.

10.

5.1. A final strengthened formulation

We complete P^{UB}_{max} formulation with some additional constraints that dominate the previous ones and then can be used to strengthen them and to improve the model.
Given a set of genotypes $G = \{g_1, ..., g_m\}$, let us define the *Conflict Graph (CG)* associated with G: $CG = (G, E)$, where G is the set of genotypes and $(i, j) \in E$ iff g_i and g_j are in conflict, i.e., there exists at least one position p such that $g_i(p) + g_j(p) = 1$. The following inequality holds:

$$\sum_{g_k \in K} \sum_{j \neq i} y^k_{\{i,j\}} \leq 1 \qquad \forall \text{ maximal clique } K \text{ on } CG, \forall \ i \ \in I_{UB},$$

that is, pair $\{i, j\}$ can cover at least one genotype in the clique K.
This inequality is certainly valid, but we can use it to strengthen constraints on z variables ((11)-(14)). For each $p \in \{1, ..., n\}$ such that $g_k(p) = 0$, let us define:

$$G^0_p = \{g \in G : \ g_k(p) = 0\},$$

and construct the conflict graph on this set G^0_p of genotypes (i.e., there is an edge between two genotypes for some coordinate different from p, if one genotype has value 0 and the other value 1).
Let CG^0_p be the conflict graph associated with G^0_p:

$$(g_k, g_{k'}) \in CG^0_p \qquad if \ \ \exists \ \ p' \neq p \ s.t. \ g_k(p') \oplus g_{k'}(p') = 1.$$

Denoted by K^0 a maximal clique on this graph CG^0_p, we can write the following inequality:

$$z_{i,p} + \sum_{g_k \in K^0} \sum_{j \neq i} y^k_{\{i,j\}} \leq 1 \qquad \forall \text{ maximal} \quad \text{clique} \ \ K^0 \ \ \text{on } CG^0_p,$$
$$\forall \ p \in \{1, ..., n\} \ s.t. \ g_k(p) = 0, \forall \ i \in I_{UB}$$

Of course, this is valid for any maximal clique of this conflict graph CG^0_p. Let us note that, given a maximal clique K on CG, a maximal clique K^0 on CG^0_p for a given p is defined as:

$$K^0 = K \cap \{g_k \in G : \quad g_k(p) = 0\}.$$

Figure 1 shows an example of conflict graph, maximal cliques K and K^0 with $p = 1$.

By symmetry, define $G^1_p = \{g \in G : \quad g_k(p) = 1\} \quad \forall \ p \in \{1, ..., n\}$; let CG^1_p be the associated conflict graph and K^1 a maximal clique on CG^1_p defined as $K^1 = K \cap \{g_k \in G : \quad g_k(p) = 1\}$.
We have:

$$z_{i,p} \geq \sum_{g_k \in K^1} \sum_{j \neq i} y^k_{\{i,j\}} \qquad \forall \text{ maximal} \quad \text{clique} \ \ K^1 \ \ \text{on } CG^1_p,$$
$$\forall \ p \in \{1, ..., n\} \ s.t. \ g_k(p) = 1, \forall i \in I_{UB}.$$

For the inequalities corresponding to $g_k(p) = 2$ ((13) and (14)), we can replace $y^k_{\{i,j\}}$ by the $\sum_{g_k \in G^2_p} y^k_{\{i,j\}}$ (because we know that at most one genotype is assigned to a given pair $\{i, j\}$). Then from the (13) and (14), we obtain:

$$\overbrace{\sum_{g_k \in G^2_p} y^k_{\{i,j\}}}^{(a)} \ \leq \ \underbrace{z_{i,p} + z_{j,p} \ \leq \ 2 - \sum_{g_k \in G^2_p} y^k_{\{i,j\}}}_{(b)}$$
$$\forall \ p \in \{1, ..., n\}, \quad \forall \ i, j \in I_{UB}, \ i \neq j.$$

Figure 1: Conflict graph on the set $G = \{010222, 001220, 222221, 011222, 120220\}$ and a maximal clique on CG_p^0.

But we can further strengthen these inequalities; since we know that $\sum_{g_k \in G_p^1} y^k_{\{i,j\}} \leq z_{i,p} \leq 1 - \sum_{g_k \in G_p^0} y^k_{\{i,j\}}$ for each position $p \in \{1, ..., n\}$, for each $i, j \in I_{UB}$, with $i \neq j$, we have for (a):

$$(a) \rightarrow \sum_{g_k \in G_p^2} y^k_{\{i,j\}} + 2 \sum_{g_k \in G_p^1} y^k_{\{i,j\}} \leq z_{i,p} + z_{j,p}$$

i.e., we know that at most one variable of this sum will be equal to one (at most one genotype is assigned to $\{i, j\}$); if the $y^k_{\{i,j\}} = 1$ corresponds to a genotype g_k in G_p^1, both $z_{i,p}$ and $z_{j,p}$ must be equal to 1 and it is valid.
In a similar way, for (b):

$$(b) \rightarrow z_{i,p} + z_{j,p} \leq 2 - \sum_{g_k \in G_p^2} y^k_{\{i,j\}} - 2 \sum_{g_k \in G_p^0} y^k_{\{i,j\}}$$

The model is now the following: P^{UB}_{max} :

$$\max \sum_{g_k \in G} \sum_{\{i,j\}} y^k_{\{i,j\}} \tag{15}$$

$$\sum_{\{i,j\}} y^k_{\{i,j\}} \leq 1 \qquad \forall\ g_k \in G \tag{16}$$

$$\sum_{g_k \in G} y^k_{\{i,j\}} \leq 1 \qquad \forall\ i, j \in I_{UB},\ i \neq j \tag{17}$$

$$z_{i,p} + \sum_{g_k \in K^0} \sum_{j \neq i} y^k_{\{i,j\}} \leq 1 \qquad \forall\ K^0 \text{ max. cliq. on } CG_p^0 \tag{18}$$
$$\forall\ p \in \{1, ..., n\}, \forall\ i \in I_{UB}$$

$$z_{i,p} \geq \sum_{g_k \in K^1} \sum_{j \neq i} y^k_{\{i,j\}} \qquad \forall\ K^1 \text{ max. cliq. on } CG_p^1 \tag{19}$$
$$\forall\ p \in \{1, ..., n\}, \forall\ i \in I_{UB}$$

$$\sum_{g_k \in G_p^2} y^k_{\{i,j\}} + 2 \sum_{g_k \in G_p^1} y^k_{\{i,j\}} \qquad \forall\ p \in \{1, ..., n\}, \tag{20}$$
$$\leq z_{i,p} + z_{j,p} \qquad \forall\ i, j \in I_{UB},\ i \neq j$$

$$z_{i,p} + z_{j,p} \leq 2 - \sum_{g_k \in G_p^2} y^k_{\{i,j\}} + \qquad \forall\ p \in \{1, ..., n\}, \tag{21}$$
$$-2 \sum_{g_k \in G_p^0} y^k_{\{i,j\}} \qquad \forall\ i, j \in I_{UB},\ i \neq j$$

$$z_{i,p} \in \{0, 1\} \qquad \forall i \in I_{UB},\ \forall p \in \{1, ..., n\}$$

$$y^k_{\{i,j\}} \in \{0, 1\} \qquad \forall g_k \in G,\ \forall i, j \in I_{UB},\ i \neq j.$$

Let us note that adding (20) and (21), we obtain

$$\sum_{g_k \in G_p^1} y^k_{\{i,j\}} + \sum_{g_k \in G_p^0} y^k_{\{i,j\}} + \sum_{g_k \in G_p^2} y^k_{\{i,j\}} \leq 1$$

which dominates the (17) which can therefore be removed from the formulation.

It is clear that there are many symmetries among solutions in the problem: for instance, haplotype set $\{1,2,3,4,5\}$ is exactly equivalent to haplotype set $\{4,1,5,2,3\}$, and in our model we can find both of them. In order to break symmetries in the formulation, we impose a lexicographic ordering on z variables, so that we order haplotypes in a lexicographic way.
Given haplotypes h_1 and h_2, we want that $h_1 \preceq h_2$, in any feasible solution, i.e.,

$$z_{1,1} < z_{2,1}.$$

Furthermore, we need a constraint for the second position such that in case of a tie on the first one, the second position will determine the order: if $z_{1,1} = z_{2,1}$, then

$$z_{1,2} < z_{2,2};$$

and hence

$$z_{1,1} + z_{1,2} < z_{2,1} + z_{2,2}.$$

If there is a tie for the first position, the first term of the left- and right-hand side will cancel and consequently only the number of times position two appears will be decisive. If there is no tie, we only want that the number of times position one appears to be decisive. We ensure this by multiplying by a specific factor the number of times position one is included. The smallest factor that gives the appropriate result is the number of possible values for any position ($\{0,1\}$), i.e. 2. In fact, for the third position, we have:

$$2\ z_{1,1} + z_{1,2} + z_{1,3} < 2\ z_{2,1} + z_{2,2} + z_{2,3}$$

Extending the reasoning on the other positions, we get the general formula:

$$\sum_{i=1}^{p} 2^{p-i} z_{j,i} \leq \sum_{i=1}^{p} 2^{p-i} z_{j+1,i} \quad \forall\ \ j \in \{1,...,UB-1\}, \forall\ \ i \in \{1,...,n\} \tag{22}$$

We can use the (5.1) to further improve the model. Let $\bar{K} = \{g_1,...,g_k\}$ be one of the largest maximal cliques on CG. We know that all genotypes in $\bar{K}$ cannot share any haplotype. Thus, we can fix all variables associated with $\bar{K}$ as follows:

$$\begin{cases} y^1_{1,2} = 1 \\ y^2_{3,4} = 1 \\ \vdots \\ y^k_{2k-1,2k} = 1 \end{cases}$$

i.e., we arbitrarily assign a couple of haplotypes to each genotype in the largest clique.
As a consequence, due to constraints in the model, we have the following relations:

$$\begin{array}{l} z_{1,p} = 0/1 \text{ if } g_1(p) = 0/1 \\ z_{2,p} = 0/1 \text{ if } g_1(p) = 0/1 \\ \vdots \\ z_{2k-1,p} = 0/1 \text{ if } g_k(p) = 0/1 \\ z_{2k,p} = 0/1 \text{ if } g_k(p) = 0/1 \end{array}$$

and

$$\begin{array}{l} y^c_{1,i} = 0 \quad \forall\ i \in I_{UB},\ \forall\ g_c\ \in G \text{ in conflict with } g_1 \\ y^c_{2,i} = 0 \quad \forall\ i \in I_{UB},\ \forall\ g_c\ \in G \text{ in conflict with } g_1 \\ \vdots \\ y^c_{2k-1,i} = 0 \quad \forall\ i \in I_{UB},\ \forall\ g_c\ \in G \text{ in conflict with } g_k \\ y^c_{2k,i} = 0 \quad \forall\ i \in I_{UB},\ \forall\ g_c\ \in G \text{ in conflict with } g_k \end{array}$$

Let us note that now we cannot add directly the (22), since it would be infeasible. But we can impose the lexicographic ordering on variables which are not in the largest clique $\bar{K}$.
Moreover, it is also possible to fix the first ambiguous position in each haplotype assigned to genotypes in the largest maximal clique $\bar{K}$:

$$z_{i,p(k)} \leq 1 - \sum_{\{i,j\}} y^k_{\{i,j\}} \qquad \forall \;\; i \in I_{UB} \tag{23}$$

for all $g_k \in \bar{K}$, for the first ambiguous position, let us say $p(k)$ s.t. $g_k(p(k)) = 2$.

5.2. Computational Experience

We compare our final model ((15), (16), (18), (19), (20), (21), (22), (23)) with the "basic" polynomial model by Brown et al. [3] that is the current polynomial model known in literature for the HIP problem. Let us note that in their paper Brown et al. first introduce a basic model for HIP as a minimum ILP problem. Then, since this model does not behave well under LP relaxation (some variables often assume value 0.5 in the LP solution), the authors use a Branch and Cut procedure to overcome that drawback: basically, they find new valid inequalities and add them during branching phase with a separation technique.
Our aim consists in comparing Brown's model and P^{UB}_{max} at the same level in order to propose a new formulation that can represent a good and robust starting point for more sophisticated techniques like cutting, column generation and polyhedral studies. Thus, we just consider the two basic formulations (note that the one of Brown is a minimum problem and ours is a maximum problem), give them to a MIP solver package and wait for a final integral solution (or an infeasibility test, as we will see for the maximum problem), without any cutting procedure.

We implemented the two models with XPRESS-MOSEL that is a comprehensive, powerful and flexible algebraic modeling language for the linear, nonlinear and integer programming problems which can be used in conjunction with any XPRESS solver. Since in our model we need to compute all maximal cliques in the conflict graph associated with genotypes in the input set, we wrote a program in C implementing the algorithm of Tsukiyama et al. [17]. Furthermore, we emphasize here that for the processed instances the COLLHAPS heuristic (implemented in C) provided a solution in negligible time (on average, less than one second) and for all data sets it found the optimal (regard to the parsimony objective) solution: consequently, we only solved one P^{UB}_{max} since UB haplotypes were not able to cover all input genotypes.

Let us remark that the advantage of our approach (a "reformulation" from P_{min} to P^{UB}_{max} and the use of a good upper bound on the number of haplotypes) allows us to avoid to solve P^{UB}_{max} to the optimum: we can stop solving it as soon as we find a feasible (not necessarily optimal) integer solution with objective value greater or equal to the number of genotypes (it implies that we are able to cover all genotypes). Thus, we add to the model the constraint:

$$\sum_{g_k \in G} \sum_{\{i,j\}} y^k_{\{i,j\}} \geq m \tag{24}$$

and set the XPRESS control parameter MAXMIPSOL to 1. In this way, XPRESS stops the first time it finds an integral-feasible solution to P^{UB}_{max} (which implies objective value greater or equal to m, given the new constraint) or it stops if it finds the problem infeasible.

In our tests, we used a generator of simulated instances known in literature: *ms*-program by Hudson [18]. The program *ms* can be used to generate many independent replicate samples (haplotypes) under a variety of assumptions about migration, recombination rate and population size to aid in the interpretation of polymorphism studies. The samples are generated using a standard coalescent approach in which the random genealogy of the sample is first generated and then mutations are randomly placed on the genealogy. Thus haplotypes were generated and then, in order to obtain genotypes, they were paired up in a random way such that all of them were used. This procedure allows us to know the final optimal haplotype reconstruction.

Hudson's program allows us to set different kinds of biological parameters, in particular the recombination level r: as r increases haplotypes become more different from each others and the number of $'2's$ in genotype samples are larger.

Testing phase is summarized in Table 1 and has been done considering 10 samples for the following kinds of instances: 50 genotypes by 10 SNPs and recombination level set to 0, 4 and 16. The gaps "–" in the table mean that none of the instances was solved in maximum time fixed to one day. We chose to process only instances 50 by 10 in order to put in evidence that even with these dimensions Brown's basic polynomial model was not able to solve them in reasonable time. We also tested instances of 50 genotypes and 30 SNPs: P^{UB}_{max}, on average, produced a solution in less than 2 hours, less than 20000 B&B visited node and only 1 MIP iteration.

50×10	r	Max. num. of '2'	time z_{ILP}	MIP iter.	B&B nodes
P^{UB}_{max}	0	4	0.24	1	20
$Brown\ Model$[3]	0	4	9702	21401	2201
P^{UB}_{max}	4	6	33.17	1	553
$Brown\ Model$[3]	4	6	19985	165010	58028
P^{UB}_{max}	16	8	875	1	12871
$Brown\ Model$[3]	16	8	–	–	–

Table 1: Comparison between P^{UB}_{max} and Brown's model: 50 genotypes and 10 SNPs; times and branch and bound node on average.

6. Conlcusions and future works

We presented a polynomial formulation for HIP problem: the new contribution respect to previous approaches (of pure integer linear programming) consists in the joint use of an ILP polynomial model and a new heuristic. We tested our resolution method with respect to the other existing polynomial model [3] (in its basic version, without any more sophisticated technique, like Branch and Cut or Cutting Plane). A low number of MIP iterations (at most one), of B&B visited nodes and a low computation time for the vast majority of the instances allow us to conclude that the proposal method has good performances.

The next step will consist in strengthening our model with a polyhedral study: we are working on the definition of the convex hull of lexigraphically ordered pair variables of 0/1 vectors that is easy to generate with one very general family of inequalities. Given the complete description of this polytope, the separation can be done in quadratic time. We are also analyzing other kinds of inequalities in order to add them to the model in a cutting procedure and compare it with the complete Branch and Cut approaches proposed by Brown at al [3].

References

[1] L. Tininini, P. Bertolazzi, A. Godi, Haplotype Inference by Parsimony for Large Datasets, Technical Report IASI n.616 (2004).

[2] A.G. Clark, Inference of haplotypes from PCR-amplified samples of diploid populations, Molecular Biol. Evol. 7 (1990) 111-122.

[3] D.G. Brown and I.M. Harrower, A New Integer Programming Formulation for the Pure Parsimony Problem in Haplotype Analysis, Fourth Workshop on Algorithms in Bioinformatics (WABI) (2004) 254-265.

[4] L. Excoffier and M. Slatkin, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, Molecular Biol. Evol. 12 (5) (1995) 921-927.

[5] D. Fallin and N.J. Schork, Accuracy of haplotype frequency estimation for biallelic loci, via the expectation maximization algorithm, for unphased diploid genotype data, Am. J. Hum. Genet. 67 (2000) 947-959.

[6] T. Niu, Z.S. Qin, J.S. Liu, Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms, Am. J. Hum. Genet. 71 (2002) 1242-1247.

[7] D. Greenspan and D. Geiger, Model-based inference of haplotype block variation, Proceedings of RECOMB03 (2003).

[8] T. Niu, Z.S. Qin, X. Xu, J.S. Liu, Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, Am. J. Hum. Genet. 70 (2002) 157-169.

[9] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978-989.

[10] D. Gusfield, Inference of Haplotypes from Samples of Diploid Populations: Complexity and Algorithms, Journal of Computational Biology 8 (3) (2001) 305-323.

[11] D. Gusfield, Haplotype inference by pure parsimony, Technical Report Computational Science and Engineering (CSE-2003-2), University of California, Davis (2003).

[12] E. Hubbell, Finding a parsimony solution to haplotype phase is NP-hard, Personal communication (2002).

[13] G. Lancia, M.C. Pinotti, R. Rizzi, Haplotyping Populations by Pure Parsimony: Complexity of Exact and Approximation Algorithms, INFORMS Journal on Computing 16 (4) (2004) 348-359.

[14] M. Stephens, P. Donnelly, A comparison of bayesian methods for haplotype reconstruction from population genotype data, Am. Journ. Hum. Genet. 73 (2003) 1162-1169.

[15] G. Lancia, R. Rizzi, A polynomial case of the parsimony haplotyping problem, Operat. Res. Letters. 34 (3) (2006) 289-295.

[16] S. Elloumi, M. Labbé, Y. Pochet, A New Formulation and Resolution Method for the p-Center Problem, INFORMS Journal on Computing 16 (1) (2004) 83-94.

[17] S. Tsukiyama, M. Ide, H. Aviyoshi and I. Shirakawa, A new algorithm for generating all the maximum independent sets, SIAM Journal on Computing 6 (1977) 505-517.

[18] R. Hudson, Gene genealogies and the coalescent process, Oxford Survey of Evolutionary Biology 7 (1990) 1-44.