INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

I F O R S
International Federation of Operational Research Societies

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 16 (2009) 561–584
DOI: 10.1111/j.1475-3995.2009.00716.x

The pure parsimony haplotyping problem: overview and computational advances

Daniele Catanzaro and Martine Labbé

Graphs and Mathematical Optimization (G.O.M.), Computer Science Department, Université Libre de Bruxelles (U.L.B.), Boulevard du Triomphe, CP 210/01, B-1050 Brussels, Belgium
E-mail: dacatanz@ulb.ac.be [Catanzaro]; mlabbe@ulb.ac.be [Labbé]

Received 18 November 2008; accepted 24 March 2009

Abstract

Haplotyping estimation from aligned single-nucleotide polymorphism fragments has attracted more and more attention in recent years due to its importance in analysis of many fine-scale genetic data. Its application fields range from mapping of complex disease genes to inferring population histories, passing through designing drugs, functional genomics, and pharmacogenetics. The literature proposes a number of estimation criteria to select a set of haplotypes among possible alternatives. Usually, such criteria can be expressed under the form of objective functions, and the sets of haplotypes that optimize them are referred to as optimal. One of the most important estimation criteria is the pure parsimony, which states that the optimal set of haplotypes for a given set of genotypes is that having minimal cardinality. Finding the minimal number of haplotypes necessary to explain a given set of genotypes involves solving an optimization problem, called the pure parsimony haplotyping (PPH) estimation problem, which is notoriously *NP*-hard. This article provides an overview of PPH, and discusses the different approaches to solution that occur in the literature.

Keywords: integer programming; computational biology; pure parsimony haplotyping; SNPs

1. Introduction

The combinations of genetic and environmental factors are the causes of common diseases such as cancer, obesity, diabetes, cardiovascular, and inflammatory diseases (The International HapMap Consortium, 2003). Discovering these genetic factors would provide fundamental new insights into the diagnosis and treatment of human diseases.

The recent completion of the sequencing phase of the Human Genome Project (The International HapMap Consortium, 2007) has shown that any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites (i.e., one variant or

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies
Published by Blackwell Publishing, 9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main St, Malden, MA 02148, USA.

polymorphism per 1000 nucleotides on average) (Wang et al., 1998; Cargill et al., 1999; Haluska et al., 1999; Li and Sadler, 1991). This result has suggested as a possible approach to identifying genetic risk factors, the search for an association between the variant sites of a specific chromosome region and a disease, by comparing a group of affected individuals with a group of unaffected ones (Risch and Merikangas, 1996). A number of association studies, focused on candidate genes, have already led to the discovery of genetic risk factors for several diseases. Examples include type 1 diabetes (Bell et al., 1984; Dorman et al., 1990; Nisticó et al., 1996), type 2 diabetes (Altshuler et al., 2000; Deeb et al., 1998), Alzheimer's disease (Strittmatter and Roses, 1996), deep vein thrombosis (Dahlbäck, 1997), inflammatory bowel disease (Rioux et al., 2001; Hugot et al., 2001; Ogura et al., 2001), hypertriglyceridemia (Pennacchio et al., 2001), schizophrenia (Stefansson et al., 2002), asthma (Van Eerdewegh et al., 2002), stroke (Gretarsdottir et al., 2003), and myocardial infarction (Ozaki et al., 2002).

One approach to doing association studies consists of testing each putative variant site for correlation with the disease (the *direct* or *brute-force* approach; The International HapMap Consortium, 2003). At present, this approach is limited to sequencing the functional parts of suspected genes, selected on the basis of a previous functional or genetic hypothesis (The International HapMap Consortium, 2003). Unfortunately, the direct approach entails the sequencing of numerous patient samples to identify the responsible variant sites, which in turn becomes prohibitively expensive.

An alternative approach (called the *indirect* approach; The International HapMap Consortium, 2003) consists in exploiting human sequence variation as genetic markers. In fact, over 90% of sequence variation among individuals is due to common variant sites (Li and Sadler, 1991), most of which arose from single historical mutation events on the ancestral chromosome (The International HapMap Consortium, 2005). Hence, in a group of people affected by a disease, the variant sites causing the disease will be enriched in frequency compared with its frequency in a group of unaffected ones. This observation was of considerable assistance, for example, in the identification of the genes responsible for cystic fibrosis and diastrophic dysplasia (The International HapMap Consortium, 2003, 2005).

The indirect approach is generally preferred to the direct one because it requires neither sequencing multiple patient samples nor prior knowledge of putative functional variant sites. However, in order to be applicable, the indirect approach requires determination of the common patterns of DNA sequence variation in the human genome (also called *haplotypes*; see Fig. 1), by characterizing sequence variants, their frequencies, and correlations between them (The International HapMap Consortium, 2003). In general this is not an easy task, because the current molecular sequencing methods only provide information about the combination (or *conflation*) of the paternal and maternal haplotypes of an individual (also called *genotype*) (Halldórsson et al., 2003). When the family-based genetic information of a population is available, haplotypes can be retrieved experimentally (Lu et al., 2003). However, in the most general case the experimental approach is laborious, cost-prohibitive, requires advanced molecular isolation strategies (Clark et al., 2001), and is sometimes not even possible (Lancia et al., 2004). In the absence of family-based genetic information, a valid alternative to the experimental approach is provided by computational methods.

In order to rebuild the haplotype map of a population, computational methods have to solve an optimization problem (called the *haplotype inference problem*) consisting of finding the set of

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Population	Molecular sequence of a specific chromosome region
Individual 1 (father)	... A C A C G C C A ... T T C G G G G T C ... A G T C G A C C G ...
Individual 1 (mother)	... A C A C G C C A ... T T C G A G G T C ... A G T C A A C C G ...
Individual 2 (father)	... A C A T G C C A ... T T C G G G G T C ... A G T C A A C C G ...
Individual 2 (mother)	... A C A C G C C A ... T T C G G G G T C ... A G T C A A C C G ...
...	...
...	...
Individual k (father)	... A C A T G C C A ... T T C G G G G T C ... A G T C A A C C G ...
Individual k (mother)	... A C A C G C C A ... T T C G A G G T C ... A G T C G A C C G ...

Haplotypes	Molecular Sequence
Haplotype 1	C T C A C G G C G T A ...
Haplotype 2	T T G T C A A C A T A ...
Haplotype 3	C C C A T G A T G T G ...
...	...
Haplotype q	T C G T C G G C A A A ...

Fig. 1. Any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites. In this example, most of the DNA sequence is identical in these chromosomes, but there are three nucleotides where variation occurs. A pattern of DNA sequence variation defines a haplotype.

haplotypes that, opportunely conflated, generate the set of genotypes observed (Lancia et al., 2004). Versions of this problem depend on the nature of the criterion used to choose a set of haplotypes among possible alternatives (Zhang et al., 2006). Two main families of haplotyping criteria have been proposed in the literature: the *likelihood criterion* (Excoffier and Slatkin, 1995; Fallin and Schork, 2000; Niu et al., 2002a) and the *parsimony criterion* (Clark, 1990).

The likelihood criterion states that under many plausible explanations of an observed phenomenon, the one with the highest probability of occurring should be preferred (Catanzaro, 2008). Hence, under the likelihood criterion, a set of haplotypes is defined to be optimal (or the most likely) if it has the highest probability of explaining the observed genotypes (Halldórsson et al., 2003). The likelihood criterion is well suited when the genotypes are characterized by a high variability (Excoffier and Slatkin, 1995). As drawbacks, the corresponding optimization problem is *NP*-hard (Halldórsson et al., 2003) and in some circumstances may provide misleading results (Gusfield and Orzack, 2005). A recent extension of the likelihood criterion, *Bayesian inference* (Erixon et al., 2003; Huelsenbeck et al., 2001; Larget and Simon, 1999), also uses biologically based prior probabilities to obtain a more accurate estimate of haplotype frequencies (Lancia and

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Rizzi, 2006; Niu et al., 2002b; Stephens and Donnelly, 2003; Stephens et al., 2001). Just as with the likelihood criterion, however, finding the optimal phylogeny using Bayesian inference is *NP*-hard (Halldórsson et al., 2003).

The *parsimony criterion* states that under many plausible explanations of an observed phenomenon, the one requiring the fewest assumptions should be preferred (Catanzaro, 2008). Hence, under the parsimony criterion, a set $\mathscr{H}$ of haplotypes is defined to be optimal (or the most parsimonious) for the genotypes analyzed if $\mathscr{H}$ is characterized by having the smallest cardinality (Gusfield, 2003; Lancia et al., 2004; Lancia and Rizzi, 2006). The parsimony criterion is well suited when the genotypes are characterized by a low–medium variability (Adkins, 2004; Gusfield and Orzack, 2005), and is at the core of several versions of the haplotype problem, namely Clark's problem (Clark, 1990), the pure parsimony haplotyping (PPH) problem (Lancia et al., 2004), the minimum perfect phylogeny problem (Gusfield and Orzack, 2005), the minimum recombination haplotype configuration problem (Li and Jiang, 2003), the zero recombination haplotype configuration problem (Li and Jiang, 2003), and the k-minimum recombination configuration problem (Li and Jiang, 2003). One drawback is that, apart from some polynomial cases (Gusfield and Orzack, 2005; Li and Jiang, 2003; Zhang et al., 2006), each version of these optimization problems has been proved to be *NP*-hard (Cilibrasi et al., 2005; Halldórsson et al., 2003).

We refer the reader interested in the likelihood criterion to Excoffier and Slatkin's seminal work (Excoffier and Slatkin, 1995), Halldórsson et al. (2003), and Gusfield and Orzack (2005). Similarly, we refer the reader interested in the parsimony criterion to Gusfield and Orzack (2005) and Bonizzoni et al. (2003).

Here, we provide a review of the available literature on the PPH problem (Lancia et al., 2004), arising from a specific version of the parsimony criterion called the *pure parsimony criterion* (Gusfield and Orzack, 2005). In Section 2 we describe in detail the biological reasons at the core of the pure parsimony criterion, and formalize the corresponding optimization problem. In Section 3 we propose a possible taxonomy of the approaches to the solution of PPH. Finally, in Sections 4 and 5 we review in detail the main solution approaches, emphasizing the exact ones and their current computational advances.

2. The PPH problem

The human genome is divided into 23 pairs of chromosomes; thereof, one copy is inherited from the father and the other from the mother. A chromosome is an organized structure of DNA that contains many genes, regulatory elements and other nucleotide sequences.

When a nucleotide site of a specific chromosome region shows a statistically significant variability within a population then it is called single-nucleotide polymorphism (SNP). Specifically, a site is considered an SNP if for a minority of the population a certain nucleotide is observed (called the least frequent allele) while for the rest of the population a different nucleotide is observed (the most frequent allele). The least frequent allele, or *mutant type* (Zhang et al., 2006), is generally encoded as "1", as opposed to the most frequent allele, or *wild type* (Zhang et al., 2006), generally encoded as "0". A haplotype is a set of alleles, or more formally, a string of length p over an alphabet $\Sigma = \{0, 1\}$.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

The diploid nature of humans implies that, for a given SNP, an individual can be either homozygous of type 0/1 (i.e., both the father and the mother alleles are equal) or heterozygous (i.e., the father and the mother alleles are different). When extracting the SNPs from an individual (i.e., when genotyping an individual) the information about which haplotype (maternal or paternal) a given allele belongs to is missed and only the homo- or heterozygous nature of the site is known. Hence, the genotype of an individual can be thought as a string of length p over an alphabet $\Sigma = \{0, 1, 2\}$, where the symbols "0" or "1" are used to denote a homozygous site and the symbol "2" is used to denote a heterozygous site. As an example, the sequence $\langle 0, 1, 2\rangle$ denotes a genotype in which the first SNP is homozygous of wild type, the second SNP is homozygous of mutant type, and finally the third SNP is heterozygous.

Haplotyping a set of genotypes means recovering from a set of genotypes the corresponding generating haplotypes. As an example, the generating haplotypes for the genotype $\langle 0, 1, 2\rangle$ are $\langle 0, 1, 1\rangle$ and $\langle 0, 1, 0\rangle$. It is worth noting that the number of possible generating haplotypes for a given genotype g grows exponentially as a function of the number of heterozygous sites of g. Specifically, if n is the number of heterozygous sites in a genotype g, then there exist 2^{n-1} possible haplotypes that may generate g (Zhang et al., 2006). As an example, genotype $\langle 0, 1, 2, 2\rangle$ may be generated by combining appropriately either the pair of haplotypes $\{\langle 0, 1, 0, 0\rangle, \langle 0, 1, 1, 1\rangle\}$ or the pair $\{\langle 0, 1, 1, 0\rangle, \langle 0, 1, 0, 1\rangle\}$. This insight entails the use of a criterion to select pairs of haplotypes among possible alternatives.

The analysis of low-rate recombination genes of different molecular functions (e.g., chaperone, ligase, isomerase, kinase, and transferase; see The International HapMap Consortium, 2007) has shown that the number of distinct haplotypes existing in a large population of individuals is generally much smaller than the overall number of distinct genotypes observed in that population (Wang and Xu, 2003). This insight has suggested that, at least for low-rate recombination genes, the criterion of minimizing the overall number of haplotypes necessary to explain a set of genotypes may have good chances to recover the biological haplotype set (Wang and Xu, 2003). This criterion, first introduced by Gusfield (2001), is known as *the pure parsimony criterion of haplotype estimation* and can be formalized as follows.

Given a pair of haplotypes $\{h_i, h_j\}$, define the operator sum $\oplus$ among h_i and h_j as the genotype g whose pth entry is h_{ip} if $h_{ip} = h_{jp}$, and 2 otherwise. As an example, the genotype obtained by summing haplotypes $h_i = \langle 0, 1, 1, 0\rangle$ and $h_j = \langle 1, 1, 0, 0\rangle$ is $g = \langle 2, 1, 2, 0\rangle$. We say that a genotype g_k is *resolved* from a pair of haplotypes $\{h_i, h_j\}$ if $g_k = h_i \oplus h_j$. Haplotyping a set of genotypes under the pure parsimony criterion, hence, consists of solving the following optimization problem:

Problem. *PPH*.

Given a set $\mathcal{G}$ of m genotypes, having p SNPs each, find the minimum set $\mathcal{H}$ of haplotypes such that for each genotype $g_k \in \mathcal{G}$ there exists a pair of haplotypes $\{h_i, h_j\} \in \mathcal{H}$ resolving g_k.

As an example, an instance of PPH and the corresponding solution is shown in Fig. 2.

PPH is known to be polynomially solvable when each genotype has at most two heterozygous sites (Lancia and Rizzi, 2006), and $\mathcal{APX}$-hard when each genotype has at least three heterozygous sites (Lancia et al., 2004). We shall propose now a possible taxonomy of the main approaches to solution of PPH currently available in the literature, and subsequently we shall review them in detail.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Instance of PPH

Genotypes	SNP			
Genotype 1	1	2	2	1
Genotype 2	2	0	2	2
Genotype 3	2	0	1	1
Genotype 4	1	0	2	2
Genotype 5	1	2	0	2

Solution

Haplotypes	SNP			
Haplotype 1	1	0	1	1
Haplotype 2	1	1	0	1
Haplotype 3	0	0	1	1
Haplotype 4	1	0	0	0

Conflation				
Genotype 1	=	**Haplotype 1**	⊕	**Haplotype 2**
Genotype 2	=	**Haplotype 3**	⊕	**Haplotype 4**
Genotype 3	=	**Haplotype 1**	⊕	**Haplotype 3**
Genotype 4	=	**Haplotype 1**	⊕	**Haplotype 4**
Genotype 5	=	**Haplotype 2**	⊕	**Haplotype 4**

Fig. 2. Graphical representation of an instance of pure parsimony haplotyping (PPH) and the corresponding solution.

Table 1
References classified by approaches to solution

Approach to solution	References
Exact	
Combinatorial branch-and-bound algorithms	Wang and Xu (2003)
Exponential ILP models	Gusfield (2003), Gusfield and Orzack (2005) Lancia and Rizzi (2006), Lancia and Serafini (2008)
Polynomial and hybrid ILP models	Brown and Harrower (2004, 2005, 2006), Bertolazzi et al. (2008) Catanzaro et al. (2007), Halldórsson et al. (2004) Lancia et al. (2004)
Pseudo-Boolean optimization algorithms	Lynce and Marques-Silva (2006a, b), Graça et al. (2007)
Non-exact	
Approximation algorithms	Lancia et al. (2004), Lancia and Rizzi (2006)
Greedy heuristics	Godi et al. (2004), Li et al. (2005), Wang and Xu (2003)
Semi-definite programming heuristics	Huang et al. (2005), Kalpakis and Namjoshi (2005)
Metaheuristics	Di Gaspero and Roli (2008), Wang et al. (2005)

ILP, Integer Linear Programming.

3. A possible taxonomy of the literature

The main perspective to classify the literature on PPH is the solution approach. Specifically, we can distinguish between exact and non-exact approaches. Exact approaches include combinatorial branch-and-bound algorithms, integer linear programming (ILP) models and pseudo-Boolean optimization (PBO) algorithms. Non-exact approaches include approximation algorithms, greedy heuristics, semi-definite programming heuristics, and metaheuristics. Table 1 lists the references according to this classification.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Historically, exact approaches were the first to be proposed in the literature on PPH. Apart from the combinatorial branch-and-bound algorithms, exact approaches can be subdivided into three groups according to the size of the model proposed. Specifically, we can distinguish among models of exponential size, models of polynomial size, and hybrid models. As a general trend, polynomial models outperform exponential and hybrid models. However, almost all models run into problems when trying to solve difficult instances, i.e., instances having a large number of heterozygous sites per genotype (Lancia, 2008). In fact, exponential models imply the creation of too many variables and/or constraints for obtaining a solution within a reasonable time. On the other hand, apart from some exceptions (see Catanzaro et al., 2007), polynomial and hybrid models use quite weak lower bounds, so that closing the gap and terminating the branch-and-bound search is impractical within a reasonable time. We shall discuss exact approaches in Section 4.

Non-exact approaches have been developed more recently and are far less numerous than exact approaches. They are typically heuristics, i.e., algorithms that produce reasonably good solutions in short computing time (Papadimitriou and Steiglitz, 1998, p. 401). Although in general heuristics do not provide any formal guarantee of the quality of the solution found, for some of them (hereafter referred to as *approximation algorithms*), it is possible to prove that the solution found is optimal up to a small constant factor. The presence (or absence) of such certificate of quality suggests, as possible classification of non-exact approaches, the subdivision into approximation algorithms and heuristics. The class of approximation algorithms for PPH has more a theoretical importance and no computational aspect has ever been analyzed as opposed to the class of heuristics, which is more of practical interest. Heuristics can be subdivided into greedy, semi-definite programming, and metaheuristics. Greedy heuristics are typically constructive in nature and perform at each iteration locally optimal choices. Semi-definite programming heuristics exploit the linear relaxation of quadratic formulations of PPH and provide solutions by means of rounding procedures. Metaheuristics are algorithms that combine a number of heuristics to tackle the problem in a more efficient way. As a general trend, metaheuristics outperform greedy and semi-definite programming heuristics providing qualitative better solutions in reasonable computing times (Di Gaspero and Roli, 2008; Wang et al., 2005). We shall discuss non-exact approaches in Section 5.

4. Exact approaches

4.1. Combinatorial branch-and-bound algorithms

The first exact approach to the solution of PPH was proposed by Wang and Xu (2003) and dates back to 2003. The approach is based on a simple combinatorial branch-and-bound algorithm in which the solution of the problem is built by enumerating all possible resolutions for each genotype in $\mathscr{G}$. The upper bound proposed is computed by means of a greedy heuristic that generates, for each genotype $g \in \mathscr{G}$, a set of possible resolving haplotypes and subsequently selects from those sets the smallest number of haplotypes that resolve the genotype set. Because of the weakness of the upper bound, Wang and Xu's algorithm is unable to solve instances of size comparable to other exact approaches. Even instances containing 20 genotypes of 20 SNPs each can take an extremely long time to be solved (Lancia, 2008).

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

4.2. Exponential ILP models

Gusfield (2003) and Gusfield and Orzack (2005) proposed the first exact approach to the solution of PPH based on ILP. Specifically, the author introduced an ILP model, called TIP, based on an implicit enumeration of all possible 2^{k-1} haplotypes that resolve each input genotype. TIP can be summarized as follows. Denote H as the set of all possible haplotypes explaining the set of genotypes $\mathcal{G}$, and let x_h be a decision variable associated with each haplotype $h \in H$. Fix any total order on H and denote H^2 as the set of those pairs of haplotypes (h_1, h_2) such that $h_1, h_2 \in H$, $h_1 < h_2$. Finally, for every genotype $g \in \mathcal{G}$, let $H_g^2 = \{(h_1, h_2) \in H^2 : h_1 \oplus h_2 = g\}$. Let $y_{h_1h_2}$ be a decision variable associated with every pair $(h_1, h_2) \in H^2$. Then, the following model is a valid formulation of PPH:

Formulation 1. *Gusfield's (2003) model*

$$\min \sum_{h \in H} x_h, \tag{1}$$

$$\text{s.t.} \sum_{(h_1,h_2) \in H_g^2} y_{h_1h_2} \geqslant 1, \quad \forall g \in \mathcal{G}, \tag{2}$$

$$y_{h_1h_2} \leqslant x_{h_1}, \qquad \forall (h_1, h_2) \in H^2, \tag{3}$$

$$y_{h_1h_2} \leqslant x_{h_2}, \qquad \forall (h_1, h_2) \in H^2, \tag{4}$$

$$x_h \in \{0, 1\}, \qquad \forall h \in H, \tag{5}$$

$$y_{h_1h_2} \in \{0, 1\}, \qquad \forall (h_1, h_2) \in H^2. \tag{6}$$

Constraints (2) impose that each genotype is explained by at least one pair of haplotypes, and constraints (3)–(4) impose that haplotypes (h_1, h_2) are allowed to explain genotype g only if h_1 and h_2 belong to the solution.

TIP is characterized by an exponential number of variables and constraints that make it impractical to solve even for small instances (Gusfield, 2003; Gusfield and Orzack, 2005). For this reason, Gusfield (2003) proposed a new model, called reduced TIP (RTIP), to provide solutions to PPH in reasonable times. The core of RTIP is based on a preprocessing strategy consisting of removing the decision variables related to those pairs of haplotypes whose sum does not result in a genotype belonging to $\mathcal{G}$. The strategy preserves the optimality of the solution and makes RTIP able to tackle instances containing up to 50 genotypes of 30 SNPs each and characterized by a relatively small numbers of heterozygous sites. No column-generation technique was described in Gusfield's article (Gusfield, 2003) and, according to Lancia and Serafini (2008), none seems suitable for RTIP. Recently, the particular structure of TIP was analyzed by Lancia and Rizzi (2006). The authors proved that when each genotype has at most two heterozygous sites, TIP constraint matrix (3)–(6) is totally unimodular and PPH is reducible to the node cover (NC) problem (Garey and Johnson, 2003) on a bipartite graph. Lancia and Rizzi also provided a polynomial algorithm able to solve this class of PPH instances with an overall computational complexity $O(pm+m^{3/2})$.

An alternative exponential model for PPH was proposed more recently by Lancia and Serafini (2008). The idea at the core of the model is to formulate PPH as a version of the set covering (SC)

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

problem (Garey and Johnson, 2003). Specifically, the authors observed that if $\mathscr{H}$ is a feasible solution to PPH then for each genotype $g \in \mathscr{G}$ there must exist a haplotype $h \in \mathscr{H}$ able to resolve g for all of its homozygous sites. The authors called this property the *covering condition* and described an ILP model able to find a haplotype set $\mathscr{H}$ for $\mathscr{G}$ satisfying the covering condition. The model can be described as follows.

Given a genotype g, let $g(s)$ be the sth SNP of g. Similarly, given a haplotype h, let $h(s)$ be the sth SNP of h. A haplotype h is said to be *compatible* with g if $g(s) = h(s)$ whenever $g(s) \neq 2$. For each genotype $g \in \mathscr{G}$, denote $A(g)$ as the set of heterozygous sites of g, $H(g)$ as the set of haplotypes that are compatible with g, and $H_s^a(g)$ as the sets of haplotypes that are compatible with g and such that their sth SNP is equal to $a \in \{0, 1\}$. Finally, let $H_{\mathscr{G}} = \cup_{g \in \mathscr{G}} H(g)$. Let x_h be a decision variable used to select a haplotype $h \in H_{\mathscr{G}}$. Then, the following ILP model ensures the satisfaction of the covering condition:

Formulation 2. *Lancia and Serafini's (2008) model*

$$\min \sum_{h \in H_{\mathscr{G}}} x_h, \tag{7}$$

$$\text{s.t.} \sum_{h \in H_s^a(g)} x_h \geqslant 1, \quad \forall g \in \mathscr{G},\ s \in A(g),\ a \in \{0, 1\}, \tag{8}$$

$$x_h \in \{0, 1\}, \qquad \forall h \in H_{\mathscr{G}}, \tag{9}$$

where constraints (8) are analogous to constraints (2) in TIP.

The above model is characterized by an exponential number of variables and a polynomial number of constraints. However, the authors observed that a solution to the above model is not necessarily a solution to PPH. In fact, the covering condition is a necessary but not sufficient condition to guarantee the feasibility of a solution to PPH. For example, consider the instance $\mathscr{G} = \{\langle 1, 2, 2, 2\rangle, \langle 2, 1, 2, 2\rangle, \langle 2, 2, 1, 2\rangle, \langle 2, 2, 2, 1\rangle\}$, the haplotype set $\mathscr{H} = \{\langle 0, 1, 1, 1\rangle, \langle 1, 0, 1, 1\rangle, \langle 1, 1, 0, 1\rangle, \langle 1, 1, 1, 0\rangle\}$ satisfies the covering condition but does not resolve $\mathscr{G}$. Hence, in order to guarantee the feasibility of the solution for PPH, the authors added the following set of exponential cuts to Formulation 2:

$$\sum_{h \in H_{\mathscr{G}} \setminus H'} x_h \geqslant 1, \quad \forall\, g \in \mathscr{G},\ \forall\, H',$$

where H' is an *insufficient* haplotype set, i.e., a haplotype set satisfying the covering condition but not resolving $\mathscr{G}$ (Lancia and Serafini, 2008). The resulting model is characterized by an exponential number of variables and constraints, and can be further strengthened by specific clique inequalities (Lancia and Serafini, 2008). The authors solved the linear programming (LP) relaxation of the model by means of row- and column-generation techniques, and observed that the lower bound so obtained is tight for the set of instances analyzed. A systematic comparison between Formulations 1 and 2 shows that RTIP needs a quantity of memory almost double that for Lancia and Serafini's model, and requires a solution time up to two orders of magnitude larger than Lancia and Serafini's (2008) model.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

4.3. Polynomial and hybrid ILP models

The first polynomial models for PPH date back to 2004 and are based on the following insight (Lancia et al., 2004): if the set of genotypes $\mathcal{G}$ has cardinality m, then at most $2m$ haplotypes are necessary to resolve $\mathcal{G}$ (Lancia et al., 2004). Hence, any solution of PPH may be represented by means of a matrix having at most $2m$ distinct rows and p columns. Halldórsson et al. (2004) and Lancia et al. (2004) exploited this fact to propose (almost simultaneously) a polynomial model for PPH. Because of the similarity of both models, in the following we shall focus on Halldórsson et al.'s (2004) model.

Let H be a $2m \times p$ binary matrix whose ith row represents the ith haplotype used to resolve some genotypes in $\mathcal{G}$. Assume that the rows of H are sorted according to the order of $\mathcal{G}$, i.e., h_{2i-1} and h_{2i} are the rows of H resolving g_i. For the ith genotype $g_i \in \mathcal{G}$, we denote by $g_i(s)$ its sth SNP and by $\mathcal{SNP}$ the set of SNPs of each genotype.

Let z_{is} be a decision variable representing the sth site of ith haplotype in H. For each possible pair of haplotypes in $1 \leqslant i < j \leqslant 2m$, let y_{ij} be a decision variable equal to 1 if haplotype $h_i \neq h_j$, and 0 otherwise. Finally, let x_i be a decision variable equal to 1 if all haplotypes with an index higher than i are different from h_i, and 0 otherwise. Then, the following model is a valid formulation of PPH:

Formulation 3. *Halldórsson et al.'s (2004) model*

$$\min \sum_{i=1}^{2m} x_i, \tag{10}$$

$$\text{s.t. } z_{2i-1,s} = z_{2i,s} = g_i(s), \quad \forall\, i \in \mathcal{G},\ s \in \mathcal{SNP} : g_i(s) \neq 2, \tag{11}$$

$$z_{2i-1,s} + z_{2i,s} = g_i(s) - 1, \quad \forall\, i \in \mathcal{G},\ s \in \mathcal{SNP} : g_i(s) = 2, \tag{12}$$

$$z_{is} - z_{qs} \leqslant y_{iq}, \quad \forall\, 1 \leqslant i < q \leqslant 2m,\ s \in \mathcal{SNP}, \tag{13}$$

$$z_{qs} - z_{is} \geqslant y_{iq}, \quad \forall\, 1 \leqslant i < q \leqslant 2m,\ s \in \mathcal{SNP}, \tag{14}$$

$$x_i \geqslant 1 - \sum_{j=1}^{i-1} (1 - y_{ji}), \quad \forall\, 1 \leqslant i \leqslant 2m, \tag{15}$$

$$x_i, y_{ij}, z_{is} \in \{0, 1\}. \tag{16}$$

Constraints (11)–(12) impose the sum operator between haplotypes. Constraints (13)–(14) force y_{ij} to be 1 if h_i differs from h_j. Finally, constraints (15) force x_i to be 1 if there are no haplotypes with an index higher than i equal to h_i. It is easily seen that the above model is polynomial in terms of both variables and constraints.

Lancia et al. (2004) discussed possible preprocessing techniques to further reduce of one order of magnitude the size of Formulation 3. However, neither Halldórsson et al. (2004) nor Lancia et al. (2004) provided information about the computational performances of their respective models. Only recently, Brown and Harrower (2006) tested the above polynomial model on a set of artificial and biological datasets. Specifically, the authors observed that the lower bound obtained from the LP relaxation of Formulation 3 is quite poor, and even including ad hoc valid inequalities the overall performance is noticeably worse than Gusfield's RTIP (Brown and

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Harrower, 2006). Therefore, Brown and Harrower developed an alternative model, called HybridIP, able to combine both the practical size of Formulation 3 and the reasonable runtime of Gusfield's RTIP. HybridIP can be summarized as follows.

Let H be the $2m \times p$ binary matrix of Formulation 3, and let $H_e = \{h_1, h_2, \ldots, h_q\}$ be a set of q specific haplotypes potentially used to resolve some genotypes in $\mathscr{G}$. Let x_i and x_j be decision variables used to select a haplotype in H and H_e, respectively. Let z_{is} be a decision variable representing the sth SNP of the ith haplotype. For each pair of haplotypes $(h_i, h_j) \in H_e$ and genotype $k \in \mathscr{G}$, let $w_{k(ij)}$ be a decision variable equal to 1 if both haplotypes h_i and h_j are used to resolve k, and 0 otherwise. For each haplotypes $h_i \in H_e$ and genotype $k \in \mathscr{G}$, let $w_{k(i^*)}$ be a decision variable equal to 1 if genotype k is resolved by a pair of haplotypes of which only h_i belongs to H_e, and 0 otherwise. Let W_k be the set of all $w_{k(ij)}$ and $w_{k(i^*)}$ variables. For each genotype $k \in \mathscr{G}$, let u_k be a decision variable equal to 1 if k is resolved by a pair of haplotypes that do not belong to H_e, and 0 otherwise. Finally, let $y_{h'h''}$ be a decision variable equal to 1 if haplotype $h' \neq h''$, and $h', h'' \in H$. Similarly, let $y_{h'h''}$ be a decision variable equal to 1 if haplotype $h' \neq h''$, where at least one between h' and h'' belongs to H_e. Then, the following model is a valid formulation of PPH:

Formulation 4. *Brown and Harrower's (2006) model*

$$\min \sum_{i=1}^{|H_e|} x'_i + \sum_{i=1}^{2m} x_i, \tag{17}$$

$$\text{s.t. } u_k + \sum_{w_{k(i\rho)} \in W_k} w_{k(i\rho)} = 1, \qquad \forall\, k \in \mathscr{G}, \tag{18}$$

$$x'_i \geqslant w_{k(ij)}, \qquad \forall\, k \in \mathscr{G},\ w_{k(ij)} \in W_k, \tag{19}$$

$$x'_j \geqslant w_{k(ij)}, \qquad \forall\, k \in \mathscr{G},\ w_{k(ij)} \in W_k, \tag{20}$$

$$z_{2k-1,s} = z_{2k,s} = g_k(s), \qquad \forall\, k \in \mathscr{G},\ s \in \mathscr{SNP} : g_k(s) \neq 2, \tag{21}$$

$$z_{2k-1,s} + z_{2k,s} = g_k(s) - 1, \qquad \forall\, k \in \mathscr{G},\ s \in \mathscr{SNP} : g_k(s) = 2, \tag{22}$$

$$z_{is} - z_{js} \leqslant y_{ij}, \qquad \forall\, i,j \in H,\ s \in \mathscr{SNP}, \tag{23}$$

$$z_{js} - z_{is} \leqslant y_{ij}, \qquad \forall\, i,j \in H,\ s \in \mathscr{SNP}, \tag{24}$$

$$z_{js} \leqslant y'_{ij}, \qquad \forall\, i \in H_e,\ j \in H,\ s \in \mathscr{SNP} : g_k(s) = 0, \tag{25}$$

$$1 - z_{js} \leqslant y'_{ij}, \qquad \forall\, i \in H_e,\ j \in H,\ s \in \mathscr{SNP} : g_k(s) = 1, \tag{26}$$

$$y'_{j,2k-1} \geqslant u_k, \qquad \forall\, k \in \mathscr{G},\ j \in H_e, \tag{27}$$

$$y'_{j,2k} \geqslant u_k, \qquad \forall\, k \in \mathscr{G},\ j \in H_e, \tag{28}$$

$$y'_{j,2k-1} \leqslant 1 - w_{k(ij)}, \qquad \forall\, k \in \mathscr{G},\ w_{k(ij)} \in W_k, \tag{29}$$

$$y'_{j,2k} \leqslant 1 - w_{k(ij)}, \qquad \forall\, k \in \mathscr{G},\ w_{k(ij)} \in W_k,\ j \neq *, \tag{30}$$

$$x_i \geqslant 2 - (q+i) + \sum_{j=1}^{q} y'_{ji} + \sum_{j=1}^{i-1} y_{ji}, \quad \forall\, i \in H, \tag{31}$$

$$x_i, x'_i, z_{ks}, y_{ij}, y'_{ij}, w_{k(ij)}, u_i \in \{0,1\}. \tag{32}$$

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Constraints (18) impose that each genotype be resolved by a pair of haplotypes such that both haplotypes belong to H, or to H_e, or one to H and another to H_e. Constraints (19)–(20) impose that variables $w_{k(ij)}$ can be used only if the corresponding haplotypes are chosen. Constraints (21)–(22) impose the sum operator among haplotypes. Constraints (23)–(24), respectively (25)–(26), are analogous to constraints (13)–(14) in Formulation 3. Constraints (27)–(28) impose that haplotypes belonging to H do not belongs to H_e. Similarly, constraints (29)–(30) impose that haplotypes belonging to H_e do not belongs to H. Finally, constraints (31) are analogous to constraints (15) in Formulation 3. Brown and Harrower (2006) compared HybridIP versus RTIP and Formulation 3 by using two sets of artificial and biological instances of PPH. The authors observed that HybridIP outperforms both RTIP and Formulation 3, proving to be on average faster and able to solve a larger number of instances.

At present, no direct comparison between the performances of HybridIP and Lancia and Serafini's (2008) model is possible, due to different biological instances and runtime environment used by both groups of authors.

Just as with Brown and Harrower, Bertolazzi et al. (2008) developed an alternative polynomial model able to combine both the practical size of Formulation 3 and the reasonable runtime of Gusfield's RTIP. The main differences from Brown and Harrower's model are the absence of the set H_e and the presence of TIP covering constraints (2). The model can be summarized as follows. Assume that the solution of PPH is represented by a binary matrix H just as with Halldórsson et al.'s model. For each haplotype in H, let x_i be a decision variable equal to 1 if haplotype i belongs to the solution, and 0 otherwise. For each genotype $k \in \mathscr{G}$ and pair of haplotypes $i,j \in H$, let y_{ij}^k be a decision variable equal to 1 if the pair of haplotypes (i,j) resolve k, and 0 otherwise. Finally, for each site $s \in \mathscr{SNP}$ and haplotype $i \in H$, let z_{is} be a decision variable equal to 1 if the sth of haplotype i is equal to "1", and 0 otherwise. Then, the following model is a valid formulation of PPH:

Formulation 5. *Bertolazzi et al.'s (2008) model-minimization problem*

$$\min \sum_{i=1}^{UB} x_i, \tag{33}$$

$$\text{s.t.} \sum_{i,j=1}^{UB} y_{ij}^k \geqslant 1, \qquad \forall\, k \in \mathscr{G}, \tag{34}$$

$$\sum_{j:i\neq j}^{UB} y_{ij}^k \leqslant x_i, \qquad \forall\, k \in \mathscr{G},\ i \in \{1,\ldots,UB\}, \tag{35}$$

$$z_{is} + \sum_{j:i\neq j}^{UB} y_{ij}^k \leqslant x_i, \qquad \forall\, k \in \mathscr{G},\ i \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 0, \tag{36}$$

$$z_{is} \geqslant \sum_{j:i\neq j}^{UB} y_{ij}^k, \qquad \forall\, k \in \mathscr{G},\ i \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 1, \tag{37}$$

$$z_{is} + z_{js} \geqslant y_{ij}^k, \qquad \forall\, k \in \mathscr{G},\ i,j \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 2,\ i \neq j, \tag{38}$$

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

$$z_{is} + z_{js} \leqslant x_i + x_j - y_{ij}^k, \quad \forall\, k \in \mathscr{G},\ i,j \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 2,\ i \neq j, \tag{39}$$

$$x_i, y_{ij}^k, z_{is} \in \{0,1\}. \tag{40}$$

where UB is an upper bound on the overall number of haplotypes needed to resolve the instance analyzed. Constraints (34) impose that each genotype must be resolved by at least a pair of haplotypes. Constraints (35) impose that a genotype can be explained by haplotype i only if i belongs to the solution. Finally, constraints (36)–(39) impose the sum operator between haplotypes. Bertolazzi et al. (2008) set the initial value of UB by means of a heuristic called Collaps (see Section 5.2). However, the authors observed that PPH can be solved by replacing Formulation 5 by a sequence of maximization problems. Specifically, the authors proposed an iterative procedure in which at each step the value of UB is decremented by one and the following maximization problem is solved:

Formulation 6. *Bertolazzi et al.'s (2008) model-maximization problem*

$$\max \sum_{k=1}^{m} \sum_{i,j=1}^{UB} y_{ij}^k, \tag{41}$$

$$\text{s.t.} \sum_{i,j=1}^{UB} y_{ij}^k \leqslant 1, \qquad \forall\, k \in \mathscr{G}, \tag{42}$$

$$\sum_{k \in \mathscr{G}} y_{ij}^k \leqslant 1, \qquad \forall\, i,j \in \{1,\ldots,UB\}, \tag{43}$$

$$z_{is} + \sum_{j:i \neq j}^{UB} y_{ij}^k \leqslant 1, \quad \forall\, k \in \mathscr{G},\ i \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 0, \tag{44}$$

$$z_{is} \geqslant \sum_{j:i \neq j}^{UB} y_{ij}^k, \qquad \forall\, k \in \mathscr{G},\ i \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 1, \tag{45}$$

$$z_{is} + z_{js} \geqslant y_{ij}^k, \qquad \forall\, k \in \mathscr{G},\ i,j \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 2,\ i \neq j, \tag{46}$$

$$z_{is} + z_{js} \leqslant 2 - y_{ij}^k, \quad \forall\, k \in \mathscr{G},\ i,j \in \{1,\ldots,UB\},\ s \in \mathscr{SNP} : g_k(s) = 2,\ i \neq j, \tag{47}$$

$$y_{ij}^k, z_{is} \in \{0,1\}. \tag{48}$$

Formulation 6 provides the maximum number of genotypes that can be resolved using at most UB haplotypes. If its optimal value is m then UB is decreased by one unit. The iterative procedure is repeated until no further decrement of UB is possible. Hence, the final value of UB is the optimal value of PPH. The authors showed that Formulation 6 can be further strengthened by adding clique inequalities, symmetry-breaking inequalities, and dominance relations. Computational results evidenced that the resulting formulation has performances comparable with Brown and Harrower's model.

The most recent polynomial model for PPH has been proposed by Catanzaro et al. (2007) and is based on the class representatives with smallest index similar to the one proposed by Campelo et al. (2008) for the node coloring problem (Garey and Johnson, 2003). The authors observed that

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

any feasible solution to PPH is constituted by (i) a set $\mathscr{H}$ of (at most $2m$) haplotypes and (ii) for each genotype $g_k \in \mathscr{G}$, a specification of a pair of haplotypes, say $\{h_i, h_j\}$, resolving g_k, i.e., such that $g_k = h_i \oplus h_j$. For example, the alternative (minimum) solutions relative to the instance of PPH shown in Fig. 3 satisfy both conditions (i) and (ii). However, note that, although having the same set of haplotypes, Solutions 1 and 2 of Fig. 3 are different, as they satisfy condition (ii) in a different way.

The authors also observed that a feasible solution to PPH can be represented by means of a bipartite graph in which each vertex $g_k \in \mathscr{G}$ is of degree 2 and the two other vertices, say h_i and h_j, adjacent to g_k satisfy $g_k = h_i \oplus h_j$. As an example, the bipartite graphs corresponding to Solutions 1 and 2 of Fig. 3 are depicted in Fig. 4a and b, respectively. The bipartite graph representation of a solution suggests that in a feasible solution to PPH the haplotypes induce a family of subsets of genotypes satisfying the following three properties: (i) each subset of genotypes shares one haplotype, (ii) each genotype belongs to exactly two subsets, and (iii) every pair of subsets intersects in at most one genotype. As an example, the haplotypes of Solution 1 in Fig. 3 induce the family of subsets of Fig. 5a satisfying properties (i)–(iii). Specifically, the subsets are induced

Instance of PPH

Genotypes	SNP			
Genotype 1	2	2	2	2
Genotype 2	1	2	0	2
Genotype 3	0	2	1	2

Minimum Solution 1

Haplotypes	SNP			
Haplotype 1	1	1	0	0
Haplotype 2	0	0	1	1
Haplotype 3	1	0	0	1
Haplotype 4	0	1	1	0

Conflation				
Genotype 1	=	**Haplotype 1**	⊕	**Haplotype 2**
Genotype 2	=	**Haplotype 1**	⊕	**Haplotype 3**
Genotype 3	=	**Haplotype 2**	⊕	**Haplotype 4**

Minimum Solution 2

Haplotypes	SNP			
Haplotype 1	1	1	0	0
Haplotype 2	0	0	1	1
Haplotype 3	1	0	0	1
Haplotype 4	0	1	1	0

Conflation				
Genotype 1	=	**Haplotype 3**	⊕	**Haplotype 4**
Genotype 2	=	**Haplotype 1**	⊕	**Haplotype 3**
Genotype 3	=	**Haplotype 2**	⊕	**Haplotype 4**

Fig. 3. Graphical representation of an instance of pure parsimony haplotyping (PPH) and two alternative solutions in Catanzaro et al.'s (2007) model.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

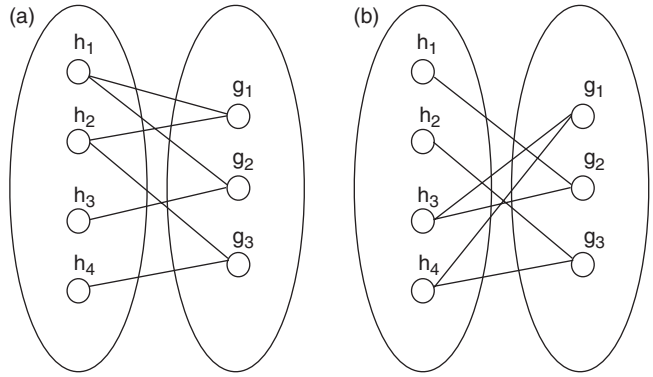

Fig. 4. Bipartite graph representation of Solutions 1 and 2 of Fig. 2.

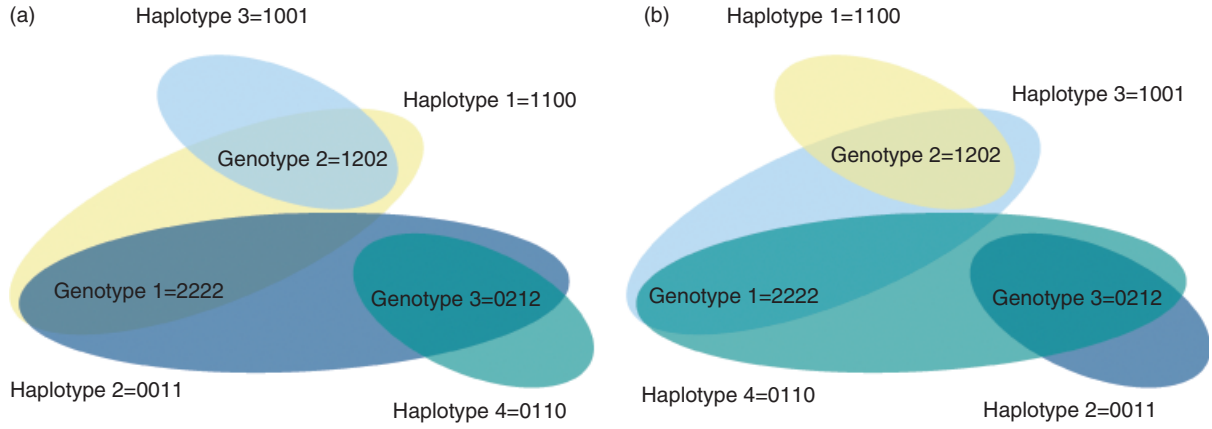

Fig. 5. Examples of subsets of genotypes induced by the two alternative solutions of Fig. 2.

by the following four haplotypes: $h_1 = \{1100\}$ inducing the subset $S_1 = \{2222, 1202\}$, $h_2 = \{0011\}$ inducing the subset $S_2 = \{2222, 0212\}$, $h_3 = \{1001\}$ inducing the subset $S_3 = \{1202\}$, and finally $h_4 = \{0110\}$ inducing the subset $S_4 = \{0212\}$. Similarly, the haplotypes of Solution 2 in Fig. 4 induce the family of subsets of Fig. 5b satisfying properties (i)–(iii). Specifically, the subsets are induced by the following four haplotypes: $h_1 = \{1100\}$ inducing the subset $S_1 = \{1202\}$, $h_2 = \{0011\}$ inducing the subset $S_2 = \{0212\}$, $h_3 = \{1001\}$ inducing the subset $S_3 = \{2222, 1202\}$, and finally $h_4 = \{0110\}$ inducing the subset $S_4 = \{2222, 0212\}$.

Catanzaro et al. (2007) exploited the bipartite graph representation of a solution to PPH to provide an ILP model that can be summarized as follows. Let associate an index with each subset S of genotypes induced by a haplotype h. Specifically, if i is the smallest index of a genotype belonging to S, then i is the index associated with S and the subset will be denoted as S_i. Since each genotype k belongs to exactly two subsets (as it must be explained by exactly two haplotypes) it may happen that k is itself the genotype with smallest index in both subsets. In this case a dummy genotype k' is added, and the subset $S_{k'}$ is created. As an example, one can imagine that

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

the haplotype h_1 induces the subset $S_i = \{g_i, g_j, g_k, \ldots\}$, h_2 induces the subset $S_{i'} = \{g_i, g_l, g_r, g_s, \ldots\}$, h_3 induces the subset $S_k = \{g_k, g_l, g_s, g_t, \ldots\}$, and so on. We remark that the index k' can be considered only if k was previously used, i.e., if the subset S_k already exists.

Since at most $2m$ haplotypes are necessary to resolve m genotypes (Lancia et al., 2004), then the indices i of the subsets S_i can vary inside $K \cup K'$, where $K = \{1, \ldots, m\}$ and $K' = \{1, \ldots, m'\}$. Assume that an order is defined on $K \cup K'$ in such a way that $1 < 1' < 2 < 2' < \ldots < m < m'$. Define x_i, $\forall i \in K \cup K'$, as a decision variable equal to 1 if, in the solution, there exists a haplotype inducing a subset S_i of genotypes whose smallest index genotype is g_i, and 0 otherwise. Denote y_{ij}^k, $\forall k \in K$, $\forall i, j \in K \cup K'$, $i < j$, as a decision variable equal to 1 if the genotype k belongs to the subsets S_i and S_j, and 0 otherwise. Finally, denote z_{is}, $\forall i \in K \cup K'$, $s \in \mathscr{SNP}$, as a decision variable equal to 1 if the haplotype inducing the subset S_i of genotypes has such a value at sth site, and 0 otherwise. Variables z_{is} shall describe explicitly the haplotypes of the solution just as with Bertolazzi et al.'s (2008) model. Then, the following model is a valid formulation of PPH:

Formulation 7. *Catanzaro et al.'s (2007) model*

$$\min \sum_{i \in K \cup K'} x_i \tag{49}$$

$$\text{s.t. } x_{i'} \leqslant x_i, \qquad \forall\, i \in K, \tag{50}$$

$$\sum_{i,j \in K \cup K'} y_{ij}^k \geqslant 1, \qquad \forall\, k \in K, \tag{51}$$

$$\sum_{j \in K \cup K' : j \geqslant i} y_{ij}^k + \sum_{j \in K \cup K' : j < i} y_{ji}^k \leqslant x_i, \qquad \forall\, k \in K,\ \forall\, i \in K \cup K', \tag{52}$$

$$y_{kk'}^k \leqslant x_{k'}, \qquad \forall\, k \in K, \tag{53}$$

$$z_{ks} = z_{k's} = 0, \qquad \forall\, k \in K,\ \forall\, s \in \mathscr{SNP} : g_k(s) = 0, \tag{54}$$

$$z_{ks} = z_{k's} = 1, \qquad \forall\, k \in K,\ \forall\, s \in \mathscr{SNP} : g_k(s) = 1, \tag{55}$$

$$z_{ks} + z_{k's} = 1, \qquad \forall\, k \in K,\ \forall\, s \in \mathscr{SNP} : g_k(s) = 2, \tag{56}$$

$$z_{is} \leqslant 1 - \sum_{j \in K \cup K' : j \geqslant i} y_{ij}^k - \sum_{j \in K \cup K' : j < i} y_{ji}^k, \qquad \forall\, s \in \mathscr{SNP},\ \forall\, k \in K : g_k(s) = 0,\ \forall\, i \in K \cup K',\ i \neq k, k', \tag{57}$$

$$z_{is} \geqslant \sum_{j \in K \cup K' : j \geqslant i} y_{ij}^k + \sum_{j \in K \cup K' : j < i} y_{ji}^k, \qquad \forall\, s \in \mathscr{SNP},\ \forall\, k \in K : g_k(s) = 1,\ \forall\, i \in K \cup K',\ i \neq k, k', \tag{58}$$

$$z_{is} + z_{js} \geqslant y_{ij}^k, \qquad \forall\, s \in \mathscr{SNP},\ \forall\, k \in K : g_k(s) = 2,\ \forall\, i, j \in K \cup K', \tag{59}$$

$$z_{is} + z_{js} \leqslant 2 - y_{ij}^k, \qquad \forall\, s \in \mathscr{SNP},\ \forall\, k \in K : g_k(s) = 2,\ \forall\, i, j \in K \cup K', \tag{60}$$

$$x_i,\ z_{is},\ y_{ij}^k \in \{0, 1\}. \tag{61}$$

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

The objective function (49) represents the number of distinct haplotypes or equivalently the cardinality of H. Since the index i' is considered only if i is already used, constraints (50) implies that if the haplotype h_i is not used, then $h_{i'}$ should not be used. Constraints (51) impose that each genotype g_k must belong to exactly two subsets S_i, S_j, and constraints (52) force x_i to be 1, i.e., to take haplotype h_i into account, if some genotype g_k is resolved by h_i. Constraints (53) are a consequence of the definition of the dummy genotype k'. Actually, they constitute a special version of constraints (52) when genotype k is resolved by haplotype k'. Constraints (54)–(56) translate the sum operation among haplotypes. Specifically, constraints (54) impose that sth site of the haplotype h_i ($h_{i'}$), inducing the subset S_i ($S_{i'}$), must be set to 0 when genotype g_i has its sth site equal to 0. By analogy, constraints (55) impose that sth site of the haplotype h_i ($h_{i'}$) must be set to 1 when genotype g_i has its sth site equal to 1. Constraints (56) impose that exactly one among the sth sites of the haplotypes h_i and $h_{i'}$ can be set to 1 when genotype g_i has its sth site equal to 2. Constraints (57) establish the relations between variables z_{is} and y_{ij}^{k}. Specifically, they force the sth site of the haplotype h_i to be equal to 0 when at least one genotype g_k, whose sth entry is equal to 0, belongs to the induced subset S_i. By analogy, constraints (58) force the sth site of the haplotype h_i to be equal to 1 when at least one genotype g_k, whose sth entry is equal to 1, belongs to the induced subset S_i. Finally, constraints (59)–(60) impose that the sum operation among haplotypes is respected for any pair of haplotypes h_i and h_j in the solution.

Catanzaro et al. (2007) discussed possible preprocessing techniques to reduce the overall number of variables and constraints, and provided valid inequalities to further strengthen the model. Computational experiments carried out on Brown and Harrower's benchmark instances (Brown and Harrower, 2006) evidenced that Formulation 7 outperforms Formulations 4 and 5, is characterized by a tight LP relaxation, and is able to handle instances containing hundreds of genotypes and SNPs. At present, Catanzaro et al.'s (2007) model is possibly the most powerful ILP approach available for PPH.

4.4. PBO algorithms

PBO algorithms are an alternative approach to combinatorial branch-and-bound algorithms and ILP models. At the core of a PBO is an iterative procedure that checks whether there exists a haplotype set H of fixed cardinality able to resolve the genotype set $\mathscr{G}$. Starting from a lower bound, the algorithm considers increasing values of the cardinality of H until an upper bound is reached. At each iteration, a Boolean satisfiability problem is solved, and in case of a positive answer the algorithm stops.

The first PBO algorithm for PPH was presented by Lynce and Marques-Silva (2006a, b) and can be summarized as follows. Denote $h_k(s)$ and $g_i(s)$ as the sth SNP of haplotype $h_k \in H$ and genotype $i \in \mathscr{G}$, respectively. Let r be the cardinality of H and let $s_{k_1 i}$ and $s_{k_2 i}$ be two Boolean variables such that if both of them are equal to true then genotype g_i is resolved by haplotypes h_{k1} and h_{k2}. Then, the following SAT model is a valid formulation of PPH:

Formulation 8. *Lynce and Marques-Silva's (2006a) PBO algorithm*

$$(\neg h_{k_1 s} \vee \neg s_{k_1 i}) \wedge (\neg h_{k_2 s} \vee \neg s_{k_2 i}), \qquad \forall\, k_1, k_2 \in H,\ i \in \mathscr{G},\ s \in \mathscr{SNP} : g_i(s) = 0, \tag{62}$$

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

$$(h_{k_1 s} \vee \neg s_{k_1 i}) \wedge (h_{k_2 s} \vee \neg s_{k_2 i}), \qquad \forall\, k_1, k_2 \in H,\ i \in \mathscr{G},\ s \in \mathscr{SNP} : g_i(s) = 1, \tag{63}$$

$$(g_i(s) \vee g_j(s)) \wedge (\neg g_i(s) \vee \neg g_j(s)), \qquad \forall\, i,j \in \mathscr{G},\ s \in \mathscr{SNP} : g_i(s),\ g_i(s) \in \{0,1\},\ g_i(s) \neq g_j(s), \tag{64}$$

$$(h_{k_1 s} \vee \neg g_i(s) \vee \neg s_{k_1 i}) \wedge (\neg h_{k_1 s} \vee g_i(s) \vee s_{k_1 i}) \wedge (h_{k_2 s} \vee \neg g_j(s) \vee \neg s_{k_2 j}) \wedge (\neg h_{k_2 s} \vee g_j(s) \vee s_{k_2 j}), \qquad \forall k_1, k_2 \in H,\ i,j \in \mathscr{G},\ s \in \mathscr{SNP} : g_i(s) = 2,\ i \neq j, \tag{65}$$

$$\left(\sum_{k_1=1}^{r} s_{k_1 i} = 1\right) \wedge \left(\sum_{k_2=1}^{r} s_{k_2 i} = 1\right), \qquad \forall i \in \mathscr{G}. \tag{66}$$

Constraints (62)–(65) impose the sum operator between haplotype. Constraints (66) impose that each genotype be resolved by exactly 2 haplotypes. The authors used a SAT solver to implement Formulation 8 and Brown and Harrower's (2006) instances to compare the performances versus HybridIP. Computational experiments showed that Formulation 8 outperforms HybridIP.

In a second article, Graça et al. (2007) developed techniques to break the symmetries in Formulation 8, and strategies to compute tight lower and upper bounds (Marques-Silva et al., 2007). The overall performances of the final formulation improves over Lynce and Silva's PBO; however, they are still not comparable with those of Catanzaro et al.'s (2007) model.

5. Non-exact approaches

5.1. Approximation algorithms

The first approximation algorithm for PPH was proposed by Lancia et al. (2004). The algorithm is mainly a primal heuristic for Formulation 3. Assume that each genotype $g \in \mathscr{G}$ contains at most k heterozygous sites. Let z_{LP} be the optimal value of the LP relaxation of Formulation 3, and let (x^*, y^*, z^*) be the corresponding optimal solution. Scale the value of each variable by a factor of 2^{k-1}. If the value so obtained is at least 1 then round the variable to 1, otherwise set it to 0. Lancia and colleagues proved that the so obtained solution is always feasible for PPH, and that its value is at most 2^{k-1} times from the optimum. In the same article, the authors also described an alternative approximation algorithm for PPH having, however, an approximation ratio of $2^{k+1}\lfloor \log p \rfloor (1 + \lceil \log p \rceil)$.

A second approximation algorithm was proposed more recently by Lancia and Rizzi (2006). The idea at its core exploits the mutual reduction between PPH and the NC problem (Garey and Johnson, 2003). Specifically, since NC can be approximated by means of a well-known 2^{k-1} primal–dual approximation algorithm (Vazirani, 2001) (where k is the number of nodes of the instance of NC), and since NC can be reduced polynomially to PPH, then also PPH can be approximated with the same approximation ratio.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

5.2. Greedy heuristics

Li et al. (2005) proposed the first greedy constructive heuristic for PPH. The algorithm, called Parsimonious Tree Grow (PTG), is based on the following rationale. If genotypes in $\mathcal{G}$ have only one site, then one can resolve $\mathcal{G}$ by means of no more than two distinct haplotypes. If genotypes in $\mathcal{G}$ have p sites and one is able to resolve a restricted instance of PPH in which genotypes have $(p-1)$ sites, then a solution for $\mathcal{G}$ can be obtained by adding opportunely a new site of value 0 or 1 to the haplotypes previously computed (Li et al., 2005). The haplotype construction process is iterative, and is performed by means of a binary tree in which the nodes of each level corresponds to the successive sites of genotypes that are explained. PTG is characterized by an overall computational complexity $O(mp^2)$ and can tackle, in about 500 s, instances of PPH containing up to 200 genotypes of 1000 SNPs each.

An alternative greedy heuristic, called Collaps, was proposed by Bertolazzi et al. (2008). The idea at its core is based on the fact that any solution of PPH may be represented by means of a matrix H having at most $2m$ distinct rows and p columns. Collaps assumes that the rows of H are sorted according to the order of $\mathcal{G}$ so that, considered the kth genotype $g_k \in \mathcal{G}$, h_{2k-1} and h_{2k} are the rows of H resolving g_k. Then, in correspondence with the sth heterozygous site of each genotype $g_k \in \mathcal{G}$, Collaps creates two logic variables, w_{ks} and $\bar{w}_{ks} = (1 - w_{ks})$, and sets $h_{2k-1,s} = w_{ks}$ and $h_{2k,s} = \bar{w}_{ks}$. Subsequently, the algorithm iteratively determines the values of the logic variables so that the number of distinct haplotypes be minimized. The value assignment is performed by enumerating a number of possible cases in which any two rows of H become equal. Experimental results, carried out on Brown and Harrower's benchmark instances, have evidenced that Collaps can tackle larger instances than PTG in a comparable computing time. For this reason, Collaps is possibly the best greedy heuristic currently available for PPH.

The most recent constructive heuristic was proposed by Lancia and Serafini (2008) and is based on a slight modification of Wang and Xu's greedy heuristic (Wang and Xu, 2003) (already described in Section 4). The only variant lies in the logic at the core of the selection of the haplotypes to resolve the genotypes. Specifically, the algorithm computes the smallest set of haplotypes that explain the largest number of genotypes, and then completes the haplotype set by adding those haplotypes necessary to resolve the remaining unexplained genotypes. The authors embodied the heuristic inside the exact algorithm for PPH and did not provide any specific experimental result for it. Hence, at present no evaluation of its performance is possible.

5.3. Semi-definite programming heuristics

Kalpakis and Namjoshi (2005) proposed a heuristic approach to the solution of PPH based on semi-definite programming. Specifically, the authors firstly provided a quadratic formulation of PPH, characterized by a polynomial number of variables and constraints, and hence, relaxed the integrality conditions, thus obtaining a semi-definite programming problem. Once the problem was solved, the authors proceeded either by rounding variables or by fixing them repeatedly to 0/1, until a feasible solution for PPH was obtained. Unfortunately, computational experiments showed that this kind of approach is extremely time consuming and can be applied to instances

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

containing no more than 20 genotypes. A very similar approach was also proposed by Huang et al. (2005). However, no information about its performance was given in that paper.

5.4. Metaheuristics

To the best of our knowledge, the only attempts at solving PPH by means of metaheuristics are restricted to the works of Wang et al. (2005) and Di Gaspero and Roli (2008). Wang et al. (2005) designed a genetic algorithm to tackle instances of PPH (Wang et al., 2005). Starting from a random initial haplotype set S, select two haplotype subsets S_1 and S_2 from S. By means of specific genetic operators (selection, tournament, crossover, and mutation; see Wang et al., 2005, for implementation details) generate a new subset S_3 from S_2 and set $S = S_1 \cup S_3$. Iterate the whole procedure until the stop conditions are met.

Wang et al. (2005) compared the performance of their genetic algorithm against Wang and Xu's exact algorithm (see Section 4) and observed that in general the former outperforms the latter on simulated instances. Unfortunately, at present no information about the performance of the genetic algorithm of Wang et al. (2005) on biological instances is available.

An alternate metaheuristic approach was proposed by Di Gaspero and Roli (2008) who designed a set of stochastic local searches for PPH, namely best improvement (BI), stochastic first improvement (SFI), simulated annealing (SA), and tabu search (TS). The local searches are instances of the general strategies described in Blum and Roli (2003). All of them start with a set of haplotypes of cardinality $2m$ and then explore the solution space iteratively, modifying pairs of resolving haplotypes in order to reduce the number of distinct ones. The authors designed specific *neighborhoods* of a haplotype (i.e., subsets of haplotypes that can be obtained from a given haplotype by modifying opportunely some sites), and hence defined some strategies to explore those neighborhoods. Specifically, BI and SFI explore a neighborhood by choosing the best, respectively the first, haplotype set that locally minimize the number of distinct haplotypes necessary to resolve g; SA explores a neighborhood by selecting a haplotype set according to probabilistic choice function. Finally, TS explores the neighborhood in the same way as BI but restricts the choice of the haplotypes with a set of forbidding rules. The authors tested their stochastic local searches on a number of instances including Brown and Harrower's (2006) benchmark instances. Computational results showed that such local searches are able to optimally solve almost all instances within 24 h.

6. Conclusion

The analysis of low-rate recombination genes of different molecular functions, e.g., chaperone, ligase, isomerase, kinase, and transferase (see The International HapMap Consortium, 2007), has shown that the number of haplotypes existing in a large population of individuals is generally much smaller than the overall number of distinct genotypes observed (Wang and Xu, 2003). This insight has suggested that, at least for low-rate recombination genes, the criterion of minimizing the overall number of haplotypes necessary to explain a set of genotypes may have good chances to recover the haplotype set (Wang and Xu, 2003). This is the fundamental consideration at the

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

core of the pure parsimony criterion of haplotype estimation and of the corresponding PPH problem.

Here we have presented a general introduction and a review of the existing literature about PPH. Our purpose has been to introduce a classification scheme in order to provide a general framework for papers appearing in this area. We have described in detail the biological reasons at the core of the pure parsimony criterion, and formalized the corresponding optimization problem. Subsequently, we have classified the literature according to the different approaches to proposed solutions. This division has been further differentiated into distinct, approximately homogeneous sub-areas, and the basic aspects of each have been discussed. For each, also, the most relevant issues affecting their use in tackling real-world-sized problems have been outlined, as have the most interesting refinements deserving further research effort.

Acknowledgement

The first author acknowledges support from the Belgian National Fund for Scientific Research (FNRS), of which he is a Research Fellow. Both the first and the second authors also acknowledge support form Communauté Française de Belgique – Actions de Recherche Concertées (ARC).

References

Adkins, R.M., 2004. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genetics* 50, 22, 1–7.

Altshuler, D., et al., 2000. The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics* 26, 76–80.

Bell, G.I., Horita, S., Karam, J.H., 1984. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33, 176–183.

Bertolazzi, P., Godi, A., Labbé, M., Tininini, L., 2008. Solving haplotyping inference parsimony problem using a new basic polynomial formulation. *Computers and Mathematics with Applications* 55, 5, 900–911.

Blum, C., Roli, A., 2003. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys* 35, 3, 268–308.

Bonizzoni, P., Della Vedova, G., Dondi, R., Jing, L., 2003. The haplotyping problem: A view of computational models and solutions. *International Journal of Computers and Science Technology* 18, 6, 675–688.

Brown, D., Harrower, I.M., 2004. A new integer programming formulation for the pure parsimony problem in haplotype analysis. In Jonassen, I., Kim, J. (eds) *Proceedings of the Fourth Annual Workshop Algorithms in Bioinformatics*, Vol. 3240. Springer-Verlag, Berlin, pp. 254–265.

Brown, D., Harrower, I.M., 2005. A new formulation for haplotype inference by pure parsimony. Technical Report, Department of Computer Science, University of Waterloo, Canada.

Brown, D., Harrower, I.M., 2006. Integer programming approaches to haplotype inference by pure parsimony. *IEEE Transactions, Computational Biology and Bioinformatics* 3, 2, 141–154.

Campelo, M., Campos, V., Correa, R., 2008. On the asymmetric representatives formulation for the vertex coloring problem. *Discrete Applied Mathematics* 156, 7, 1097–1111.

Cargill, M., et al., 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* 22, 231–238.

Catanzaro, D., 2008. The minimum evolution problem: Overview and classification. *Networks* 53, 2, 89–90.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Catanzaro, D., Godi, A., Labbé, M., 2007. A class representative model for pure parsimony haplotyping. Technical Report, G.O.M. – Computer Science Department – Université Libre de Bruxelles (U.L.B.).

Cilibrasi, R., van Iersel, L., Kelk, S., Tromp, J., 2005. On the complexity of several haplotyping problems. In Casadio, R., Myers, G. (eds) *Algorithms in Bioinformatics, Vol. 3692 of Lecture Note in Computer Science*. pp. 128–139. Springer-Verlag, Berlin, Germany.

Clark, A.G., 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution* 7, 111–122.

Clark, V.J., Methey, N., Dean, M., Peterson, R.J., 2001. Statistical estimation and pedigree analysis of CCR2–CCR5 haplotypes. *Human Genetics* 108, 484–493.

Dahlbäck, B., 1997. Resistance to activated protein C caused by the factor V R506Q mutation is a common risk factor for venous thrombosis. *Journal of Thrombosis and Haemostasis* 78, 483–488.

Deeb, S.S., et al., 1998. A Pro12Ala substitution in PPAR γ 2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nature Genetics* 20, 284–287.

Di Gaspero, L., Roli, A., 2008. Stochastic local search for large-scale instances of the haplotype inference problem by pure parsimony. *Journal of Algorithms* 63, 3, 55–69.

Dorman, S.J., et al., 1990. Worldwide differences in the incidence of type I diabetes are associated with amino acid variation at position 57 of the HLA-DQ β chain. *Proceedings of the National Academy of Sciences of the USA* 87, 7370–7374.

Erixon, P., Svennblad, B., Britton, T., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology* 52, 665–673.

Excoffier, L., Slatkin, M., 1995. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* 12, 5, 921–927.

Fallin, D., Schork, N.J., 2000. Accuracy of haplotype frequency estimation for biallelic loci via the expectation maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics* 67, 947–959.

Garey, M.R., Johnson, D.S., 2003. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York.

Godi, A., Tininini, L., Bertolazzi, P., 2004. Haplotype inference by parsimony for large datasets. Technical Report 616, IASI, Istituto di Analisi dei Sistemi ed Informatica – CNR, Rome.

Graça, A., Marques-Silva, J., Lynce, I., Oliviera, A.L., 2007. Efficient haplotype inference with pseudo-Boolean optimization. *Lecture Notes in Computer Science* 4545, 125–139.

Gretarsdottir, S., et al., 2003. The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genetics* 35, 131–138.

Gusfield, D., 2001. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Journal of Computational Biology* 8, 305–324.

Gusfield, D., 2003. Haplotype inference by pure parsimony. In Baeza-Yates, R., Chávez, E., Crochemore, M. (eds) *Annual Symposium in Combinatorial Pattern Matching, Vol. 2676, Lecture Note in Computer Science*. Springer-Verlag, Berlin, pp. 144–155.

Gusfield, D., Orzack, S.H., 2005. Haplotype inference. In Aluru, S. (ed) *Handbook on Bioinformatics*. CRC Press, Boca Raton, FL, pp. 1–28.

Halldórsson, B.V., Bafna, V., Edwards, N., Lippert, R., 2003. Combinatorial problems arising in SNP and haplotype analysis. In Calude, C.S. (ed) *Discrete Mathematics and Theoretical Computer Science, Vol. 2731 of Lecture Note in Computer Science*. Springer-Verlag, Berlin, pp. 26–47.

Halldórsson, B.V., Bafna, V., Edwards, N., Lippert, R., Yooseph, S., 2004. A survey of computational methods for determining haplotypes. *Lecture Notes in Computer Science* 2983, 26–47.

Haluska, M.K., et al., 1999. Patterns of single nucleotide polymorphisms in candidate genes of blood pressure homeostasis. *Nature Genetics* 22, 239–247.

Huang, Y.T., Chao, K.M., Chen, T., 2005. An approximation algorithm for haplotype inference by maximum parsimony. In Haddad, H., Liebrock, L.M., Omicini, A., Wainwright, R.L. (eds) *ACM Symposium on Applied Computing*. Springer-Verlag, Berlin, pp. 146–150.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–2314.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Hugot, J.P., et al., 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411, 599–603.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426, 18, 789–796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437, 27, 1299–1314.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 18, 851–861.

Kalpakis, K., Namjoshi, P., 2005. Haplotype phasing using semidefinite programming. In Kapyris, G., Bourbakis, N., Tsai, J.J. (eds) *Bioinformatics and Bioengineering (BIBE) 2005*. IEEE Computer Society, Minneapolis, MN, pp. 145–152.

Lancia, G., 2008. The phasing of heterozygous traits: algorithms and complexity. *Computer and Mathematics with Applications* 55, 960–969.

Lancia, G., Rizzi, R., 2006. A polynomial case of the parsimony haplotyping problem. *Operations Research Letters* 34, 3, 289–295.

Lancia, G., Serafini, P., 2008. A set covering approach with column generation for parsimony haplotyping. *INFORMS Journal on Computing* 21, 151–166.

Lancia, G., Pinotti, M.C., Rizzi, R., 2004. Haplotyping populations by pure parsimony: complexity of exact and approximate algorithms. *INFORMS Journal on Computing* 16, 4, 348–359.

Larget, B., Simon, D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16, 750–759.

Li, J., Jiang, T., 2003. Efficient inference of haplotype from genotype on a pedigree. *Journal of Bioinformatics and Computational Biology* 10, 1, 41–69.

Li, W.H., Sadler, L.A., 1991. Low nucleotide diversity in man. *Genetics* 129, 513–523.

Li, Z.P., Zhou, W.F., Zhang, X.S., Chen, L., 2005. A parsimonious tree-grow method for haplotype inference. *Bioinformatics* 21, 17, 3475–3481.

Lu, X., Niu, T., Liu, J.S., 2003. Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. *Genome Research* 13, 2112–2117.

Lynce, I., Marques-Silva, J., 2006a. Efficient haplotype inference with Boolean satisfiability. In National Conference on Artificial Intelligence, Boston, MA.

Lynce, I., Marques-Silva, J., 2006b. SAT in bioinformatics: making the case with haplotype inference. *Lecture Notes in Computer Science* 4121, 136–141.

Marques-Silva, J., Lynce, I., Oliviera, A.L., 2007. Efficient and tight upper bounds for haplotype inference by pure parsimony using delayed haplotype selection. *Lecture Notes in Artificial Intelligence* 4874, 621–632.

Nisticó, L., et al., 1996. The ctla-4 gene region of chromosome 2q33 is linked to, and associated with, type I diabetes. *Human Molecular Genetics* 5, 1075–1080.

Niu, T., Qin, Z.S., Liu, J.S., 2002a. Partition–ligation–expectation–maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics* 71, 1242–1247.

Niu, T., Qin, Z.S., Xu, X., Liu, J.S., 2002b. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics* 70, 157–169.

Ogura, Y., et al., 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 411, 603–606.

Ozaki, K., et al., 2002. Functional SNPs in the lymphotoxin-a gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* 32, 650–654.

Papadimitriou, C., Steiglitz, K., 1998. *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, Mineola.

Pennacchio, L.A., et al., 2001. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* 294, 169–173.

Rioux, J.D., et al., 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genetics* 29, 223–228.

Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.

Stefansson, H., et al., 2002. Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics* 71, 877–892.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies

Stephens, M., Donnelly, P., 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* 73, 1162–1169.

Stephens, M., Smith, N.J., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 68, 978–989.

Strittmatter, W.J., Roses, A.D., 1996. Apolipoprotein E and Alzheimer's disease. *Annual Reviews – Neuroscience* 19, 53–77.

Van Eerdewegh, P., et al., 2002. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 418, 426–430.

Vazirani, V., 2001. *Approximation Algorithms*. Springer-Verlag, Berlin.

Wang, D.G., et al., 1998. Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.

Wang, L., Xu, Y., 2003. Haplotype inference by maximum parsimony. *Bioinformatics* 19, 14, 1773–1780.

Wang, R.S., Zhang, X.S., Sheng, L., 2005. Haplotype inference by pure parsimony via genetic algorithm. *Lecture Notes in Operations Research* 5, 308–318.

Zhang, X.S., Wang, R.S., Wu, L.Y., Chen, L., 2006. Models and algorithms for haplotyping problem. *Current Bioinformatics* 1, 105–114.

© 2009 The Authors.
Journal compilation © 2009 International Federation of Operational Research Societies