Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Operations Research Letters 34 (2006) 289–295

Operations Research Letters
www.elsevier.com/locate/orl

A polynomial case of the parsimony haplotyping problem

Giuseppe Lancia[a,*], Romeo Rizzi[b]

[a]*Dipartimento di Matematica e Informatica, Università di Udine, Via delle Scienze 206, 33100 Udine, Italy*
[b]*Dipartimento di Informatica e Telecomunicazioni, Università di Trento, Italy*

Received 3 February 2005; accepted 7 May 2005
Available online 27 June 2005

Abstract

The *parsimony haplotyping* problem was shown to be NP-hard when each *genotype* had $k \leqslant 3$ *ambiguous positions*, while the case for $k \leqslant 2$ was open. In this paper, we show that the case for $k \leqslant 2$ is polynomial, and we give approximation and FPT algorithms for the general case of $k \geqslant 0$ ambiguous positions.

Keywords: Totally unimodular matrices; Approximation algorithms; Fixed parameter tractability; Parsimony haplotyping; Computational biology

1. Introduction

In this paper we focus on a combinatorial problem defined on ternary vectors and derived from a molecular genetics procedure known as *haplotyping*. We will briefly describe the biological motivations behind the problem definition in Section 1.1. The starting point of the investigations reported here is to settle the complexity for a special class of instances which had resisted prior complexity classification, but whose relevance to real-life applications appears minor. However, the deeper mathematical insight gained into these borderline instances will allow us to derive simple and practical combinatorial approximation algorithms for a wider and more significant class of instances. Moreover, FPT algorithms for this more relevant class of instances will also follow as a byproduct of our core structural results. Therefore, we prefer to focus on the mathematical aspects of the result more than on the implications that our algorithm can have for the practical solution of the biology problem.

The problem data consist in a set $\mathscr{G}$ of n *genotypes* $g_1, \ldots, g_n$, corresponding to n individuals in a population. Each genotype g is a vector with entries in $\{0, 1, 2\}$. Each position where a 2 appears is called an *ambiguous* position. For each genotype g we must determine a pair of *haplotypes* h_{P} and h_{M} (h_{P} stands for the *paternal* haplotype and h_{M} stands for the *maternal* haplotype), which are binary vectors such that $g = h_{\mathrm{P}} \oplus h_{\mathrm{M}}$ (where the definition of the $\oplus$ operation will be given later). A set $\mathscr{H}$ of haplotypes is said to

[*] Corresponding author.
E-mail address: lancia@dimi.uniud.it (G. Lancia).

0167-6377/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2005.05.007

Hapl. 1, paternal:	taggtcc**C**tattt**C**ccaggcgc**C**gtatacttcgacggg**T**ctata
Hapl. 1, maternal:	taggtcc**G**tattt**A**ccaggcgc**G**gtatacttcgacggg**T**ctata
Hapl. 2, paternal:	taggtcc**C**tattt**A**ccaggcgc**G**gtatacttcgacggg**T**ctata
Hapl. 2, maternal:	taggtcc**G**tattt**C**ccaggcgc**G**gtatacttcgacggg**C**ctata
Hapl. 3, paternal:	taggtcc**C**tattt**A**ccaggcgc**G**gtatacttcgacggg**T**ctata
Hapl. 3, maternal:	taggtcc**G**tattt**A**ccaggcgc**C**gtatacttcgacggg**C**ctata

Fig. 1. The haplotypes of 3 individuals, with 4 SNPs.

explain $\mathscr{G}$ if for each $g \in \mathscr{G}$ there is at least one pair of haplotypes $h', h'' \in \mathscr{H}$ such that $g = h' \oplus h''$.

Given a set $\mathscr{G}$ of genotypes, the *haplotyping* problem consists in finding a set $\mathscr{H}$ of haplotypes that explains $\mathscr{G}$. In this paper we will pursue the *parsimony* objective function, i.e., we will be interested in finding a set $\mathscr{H}$ of smallest possible cardinality.

1.1. Polymorphisms and biological motivations

A *single nucleotide polymorphism* (SNP) is a site of the human genome (i.e., the position of a specific nucleotide) showing a statistically significant variability within a population. Besides very rare exceptions, at each SNP site we observe only two (out of the possible four, i.e., A, T, C and G) nucleotides, called the SNP *alleles*. The recent completion of the sequencing phase of the Human Genome Project [15,12] has shown that the genome of any two individuals is the same in about 99% of the positions, and that most polymorphisms (i.e., differences at genomic level) are in fact SNPs [3].

Humans are *diploid* organisms, i.e., their DNA is organized in pairs of chromosomes. For each pair of chromosomes, one chromosome copy is inherited from the father and the other copy is inherited from the mother. For a given SNP, an individual can be either *homozygous* (i.e., possess the same allele on both chromosomes) or *heterozygous* (i.e., possess two different alleles). The values of a set of SNPs on a particular chromosome copy define a *haplotype*.

In Fig. 1, we illustrate a simplistic example of three individuals and four SNPs. The alleles for SNP 1, in this example, are C and G. Individual 1, in this example, is heterozygous for SNPs 1, 2 and 3, and homozygous for SNP 4. His haplotypes are `CCCT` and `GAGT`.

Haplotyping an individual consists of determining his two haplotypes, for a given chromosome. With the larger availability in SNP genomic data, recent years have witnessed the emergence of a set of new computational problems related to haplotyping. These problems are motivated by the fact that it is economically infeasible to determine the haplotypes experimentally. On the other hand, there is a reasonable experiment which can determine the (less informative and often ambiguous) genotypes, from which the haplotypes must then be retrieved computationally.

A *genotype* provides, for an individual, information about the multiplicity of each SNP allele, i.e., for each SNP site, the genotype specifies whether the individual is heterozygous or homozygous (in the latter case, it also specifies the allele).

The ambiguity comes from heterozygous sites, since, to retrieve the haplotypes, one has to decide how to distribute the two allele values on the two chromosome copies. *Resolving* (or *explaining*) a genotype g requires determining two haplotypes such that, if they are assumed to be the two chromosome copies, then, computing the multiplicity of each SNP allele, we obtain exactly the genotype g. Given a set $\mathscr{G}$ of genotypes, the *population haplotyping problem* requires to determine a set $\mathscr{H}$ of haplotypes such that each genotype $g \in \mathscr{G}$ is explained by two haplotypes $h', h'' \in \mathscr{H}$. For its importance (as we said, haplotyping from genotype data is nowadays the only viable way) the population haplotyping problem has been and is being extensively studied, under many objective functions, among which are: *Perfect phylogeny* [1,5], *Clark's rule* [6,7] and *Parsimony* [8,13,2].

Each model and objective function has specific biological motivations, which are discussed in the cited references. In this paper, we focus on the parsimony haplotyping problem. Under the parsimony objective, it is required that $\mathscr{H}$ has the minimum possible cardinality. The objective is based on the principle that, under many explanations of an observed phenomenon, one should choose the one that requires the fewest assumptions. This problem has been introduced by Gusfield [8], who adopted an integer-programming formulation for its practical solution. The problem is NP-hard, as first shown by Hubbel [11].

Among the several objective functions for haplotyping, pure parsimony is the most recent model, whose importance is now being recognized as crucial in the