Extended Islands of Tractability for Parsimony Haplotyping

Rudolf Fleischer¹, Jiong Guo², Rolf Niedermeier³, Johannes Uhlmann³, Yihui Wang¹, Mathias Weller³, and Xi Wu¹

¹ School of Computer Science, IIPL, Fudan University, Shanghai, China
² Universität des Saarlandes, Campus E 1.4, D-66123 Saarbrücken, Germany
³ Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

Abstract. Parsimony haplotyping is the problem of finding a smallest size set of haplotypes that can explain a given set of genotypes. The problem is NP-hard, and many heuristic and approximation algorithms as well as polynomial-time solvable special cases have been discovered. We propose improved fixed-parameter tractability results with respect to the parameter “size of the target haplotype set” \( k \) by presenting an \( O^*(k^4k) \)-time algorithm. This also applies to the practically important constrained case, where we can only use haplotypes from a given set. Furthermore, we show that the problem becomes polynomial-time solvable if the given set of genotypes is complete, i.e., contains all possible genotypes that can be explained by the set of haplotypes.

1 Introduction

Over the last few years, haplotype inference has become one of the central problems in algorithmic bioinformatics [10,2]. Its applications include drug design, pharmacogenetics, mapping of disease genes, and inference of population histories. One of the major approaches to haplotype inference is parsimony haplotyping: Given a set of genotypes, the task is to find a minimum-cardinality set of haplotypes that explains the input set of genotypes. The task to select as few haplotypes as possible (parsimony criterion) is motivated by the observation that in natural populations the number of haplotypes is much smaller than the number of genotypes [2]. Referring for the background in molecular biology to the rich literature (see, e.g., the surveys by Catanzaro and Labbé [2] and Gusfield and Orzack [10]), we focus on the underlying combinatorial problem. In an abstract way, a genotype can be seen as a length-\( m \) string over the alphabet \( \{0, 1, 2\} \).

* Supported by the DFG, research projects PABI, NI 369/7, and DARE, GU 1023/1, NI 369/11, NSF China (No. 60973026), Shanghai Leading Academic Discipline Project (project number B114), Shanghai Committee of Science and Technology of China (nos. 08DZ2271800 and 09DZ2272800), the Excellence Cluster on Multimodal Computing and Interaction (MMCI), and Robert Bosch Foundation (Science Bridge China 32.5.8003.0040.0).
while a haplotype can be seen as a length-$m$ string over the alphabet $\{0, 1\}$. A set $H$ of haplotypes explains, or resolves, a set $G$ of genotypes if for every $g \in G$ there is either an $h \in H$ with $g = h$ (trivial case), or there are two haplotypes $h_1$ and $h_2$ in $H$ such that, for all $i \in \{1, \ldots, m\}$,

- if $g$ has letter 0 or 1 at position $i$, then both $h_1$ and $h_2$ have this letter at position $i$, and
- if $g$ has letter 2 at position $i$, then one of $h_1$ or $h_2$ has letter 0 at position $i$ while the other one has letter 1.

For example, $H = \{00100, 01110, 10110\}$ resolves $G = \{02120, 20120, 22110\}$. Parsimony haplotyping is NP-hard, and numerous algorithmic approaches based on heuristics and integer linear programming methods are applied in practice [2]. There is also a growing list of combinatorial approaches (with provable performance guarantees) including the identification of polynomial-time solvable special cases, approximation algorithms, and fixed-parameter algorithms [5,13,14,16,11].

In this work, we contribute new combinatorial algorithms for parsimony haplotyping, based on new insights into the combinatorial structure of a haplotype solution. Lancia and Rizzi [14] showed that parsimony haplotyping can be solved in polynomial time if every genotype string contains at most two letters 2, while the problem becomes NP-hard if genotypes may contain three letters 2 [13]. Sharan et al. [16] proved that parsimony haplotyping is APX-hard in even very restricted cases and identified instances with a specific structure that allow for polynomial-time exact solutions or constant-factor approximations. Moreover, they showed that the problem is fixed-parameter tractable with respect to the parameter $k = \text{"number of haplotypes in the solution set"}$. The corresponding exact algorithm has running time $O(k^{k^2 + km})$. These results were further extended by van Iersel et al. [11] to cases where the genotype matrix (the rows are the genotypes and the columns are the $m$ positions in the genotype strings) has restrictions on the number of 2’s in the rows and/or columns. They identified various special cases of haplotyping with polynomial-time exact or approximation algorithms with approximation factors depending on the numbers of 2’s per column and/or row, leaving open the complexity of the case with at most two 2’s per column (and an unbounded number of 2’s per row). Further results in this direction have been recently provided by Cicalaese and Milanić [3]. Finally, Fellows et al. [5] introduced the constrained parsimony haplotyping problem where the set of haplotypes may not be chosen arbitrarily from $\{0, 1\}^m$ but only from a pool $\tilde{H}$ of plausible haplotypes. Using an intricate dynamic programming algorithm, they extended the fixed-parameter tractability result of Sharan et al. [16] to the constrained case, proving a running time of $k^{O(k^2)} \cdot \text{poly}(m, |\tilde{H}|)$. Jäger et al. [12] recently presented an experimental study of algorithms for computing all possible haplotype solutions for a given set of genotypes, where the integer linear programming and branch-and-bound algorithms were sped up using some insights into the combinatorial structure of the haplotype solution, as for example eliminating equal columns from the genotype matrix and recursively decomposing a large problem into smaller ones.

Our contributions are as follows. We simplify and improve the fixed-parameter tractability results of Sharan et al. [16] and Fellows et al. [5] by proposing