On intractability of haplotype inferring via galled-tree networks

Arvind Gupta [*] Ján Maňuch [†]
Ladislav Stacho [‡] Xiaohong Zhao

School of Computing Science and Department of Mathematics
Simon Fraser University, Canada
{arvind|jmanuch|lstacho|xzhao2}@sfu.ca

Abstract

The problem of determining haplotypes from genotypes has gained considerable prominence in the research community. Here the focus is on determining the actual DNA sequence of individual chromosomes since such information captures the genetic causes of disease. Present algorithmic tools for haplotyping make effective use of phylogenetic tree construction and, in particular, perfect phylogeny [7]. Here the underlying assumption is that recombinations are not present, an assumption with some basis in experimentation. However these experiments do not exclude recombinations and models are needed that incorporate this extra degree of complication.

First attempt in haplotyping via models which allow a limited number of biological events that violate the perfect phylogeny model was taken in [15]. In this paper a polynomial algorithm for haplotyping via imperfect phylogenies with a single homoplasy was presented, as well as a practical algorithm for haplotyping via galled-tree networks with one gall. In earlier work we characterized the existence of the galled-tree networks [6]. Building on this work, we are able to reduce the problem of haplotype inferring via galled-tree networks to a hypergraph covering problem, although we require that the genotype matrices satisfy a combinatorial condition. Note that our experiments on real and simulated data show that this condition is almost always satisfied. In [5], we have presented a polynomial algorithm based on the above reduction for haplotype inference via galled-tree networks with galls having exactly two mutations. It is very natural to ask whether the assumption on the size of galls can be dropped and still hope for a polynomial algorithm. In this paper we show that without this assumption, we are unable to solve the problem in polynomial time. Indeed, we show that the hypergraph covering problem we obtain in the general case is NP-complete by reduction from 3-SAT. Hence, we have a strong evidence that the haplotype inferring via galled-tree networks is intractable.

1 Introduction

With the sequencing of the Human Genome, research has focused on the problem of determining the DNA sequence of individuals on each of their chromosomes. Such information captures genetic variation and is already playing a central role in helping to determine the genetic causes of disease and in designing effective pharmaceutical responses to these diseases. Experimental methods allow

[*] Research supported in part by NSERC (Natural Science and Engineering Research Council of Canada) grant.

[†] Research supported in part by PIMS (Pacific Institute for Mathematical Sciences).

[‡] Research supported in part by NSERC (Natural Science and Engineering Research Council of Canada) grant.

for cost-effective determination of genotype information (the combined information for an individual across both matching chromosomes) and so the problem can be reduced to computationally determining haplotypes from genotypes.

Helmuth [10] was among the first to note the importance of haplotypes and it's role in understanding diseases. Other researchers (see [3, 14] for example) also noted that haplotypes have important implications for identifying disease associations and isolating environmental and genetic affects on disease. Gabirel et al [4] noted that haplotyping contributed to the identification of genes for Mendelian diseases amongst others. This body of work is now encompassed in the International HapMap project (cf. http://www.hapmap.org/abouthapmap.html).

While in-vitro techniques can be used for haplotyping as well, the cost is prohibitive, and therefore is a strong interest in the development of algorithmic tools [12]. The first heuristic algorithm for computational haplotype inference was designed by Clark [2]. Gusfield [7] developed the first exact algorithms based on the assumption of no recombinations which allowed him to make effective use of phylogenetic trees. This assumption was justified by experimental results that show chromosomes are partitioned into blocks with a strong correlation between sites on the same block ([3, 14]). As such these experiments do not exclude recombinations within a block and models were needed that allow for a small rate of recombinations.

First attempt in haplotyping via models which allow a limited number of biological events that violate the perfect phylogeny model was taken in [15]. In this paper a polynomial algorithm for haplotyping via imperfect phylogenies with a single homoplasy was presented, as well as a practical algorithm for haplotyping via galled-tree networks with one gall. In earlier work[6] we considered the problem of characterizing the existence of galled-tree networks. Later in [5], we have presented a polynomial algorithm for haplotype inference via galled-tree networks with galls having two mutations based on reduction of haplotyping problem to a hypergraph covering problem. It is very natural to ask whether the assumption on the size of galls can be dropped and still hope for a polynomial algorithm. In this paper we show that without this assumption, we are unable to solve the problem using the same technique as in [5] in polynomial time. Indeed, we show that the modified hypergraph covering problem we obtain in the general case is NP-complete by reduction from 3-SAT. Hence, we have a strong evidence that the haplotype inferring via galled-tree networks is intractable.

1.1 Definitions of phylogenetic and galled-tree networks

We will assume that a genotype matrix has values in $\{0, 1, 2\}$. Similarly, a haplotype matrix has values in $\{0, 1\}$.

Definition 1. A *phylogenetic network* N on m characters is a directed acyclic graph containing exactly one vertex (the root) with no incoming edges, a set of internal vertices that have both incoming and outgoing edges, and exactly n vertices (the leaves) with no outgoing edges. Each vertex other than the root has either one or two incoming edges. If it has one incoming edge, the edge is called a *mutation edge*, otherwise it is called a *recombination edge.* A vertex x with two incoming edges is called a *recombination vertex.*

Each integer (character) from 1 to m is assigned to exactly one mutation edge in N and each mutation edge is assigned one character. Each vertex in N is labeled by a binary sequence of length m, starting with the root vertex which is labeled with the all-0 sequence. Since N is acyclic, the vertices in N can be topologically sorted into a list, where every vertex occurs in the list only after its parent(s). Using that list, we can define the labels of the non-root vertices, in order of their appearance in the list, as follows:

- For a non-recombination vertex v, let e be the mutation edge labeled c coming into v. The label of v is obtained from the label of v's parent by changing the value at position c from 0 to 1.
- Each recombination vertex x is associated with an integer $\mathrm{r}_x \in \{2, \ldots, m\}$, called the *recombination point* for x. Label the two recombination edges coming to x P and S, respectively. Let $P(x)$ ($S(x)$) be the sequence of the parent of x on the edge labeled P (S). Then the label of x consists of the first $\mathrm{r}_x - 1$ characters of $P(x)$, followed by the last $m - \mathrm{r}_x + 1$ characters of $S(x)$. Hence $P(x)$ contributes a prefix and $S(x)$ contributes a suffix to x's sequence.

In this paper, the sequence at the root of the phylogenetic network is always the all-0 sequence, and all results are relative to that assumption. More general phylogenetic networks with unknown root were studied in a recent paper by Gusfield[8]. Note also that there are slight differences in the definition of phylogenetic networks from the original definition of Wang et al.[16]. We assume that each mutation edge has exactly one label. Every phylogenetic network without this assumption can be easily transformed to our model by replacing every mutation edge with multiple labels by a sequence of edges each having one of these labels, and contracting all mutation edge without a label, and vice versa. Our definition results in a more uniform phylogenetic networks, however we cannot require that all sequences of an input matrix appear at the leaves of the network.

Definition 2. Given an $n \times m$ matrix A with values in $\{0, 1\}$, we say that a phylogenetic network N with m characters *explains* A if each sequence of A is a label of some vertex in N.

By the definition of the galled-tree network, the following observation follows:

Observation 1. *Given a haplotype matrix M, let M' be a matrix obtained from M by duplicating one row or adding a row with all-0 sequence. Matrix M' can be explained by a galled-tree network N if and only if M can be explained by N. Let M'' be a matrix obtained from M by removing one row. If M can be explained by a galled-tree network N then also M'' can be explained by N.*

Definition 3. (Galled-tree network) In a phylogenetic network N, let v be a vertex that has two paths out of it that meet at a recombination vertex x (v is the least common ancestor of the parents of x). The two paths together form a *recombination cycle C*. The vertex v is called the *coalescent vertex*. We say that C contains a character i, if i labels one of the mutation edges of C.

A phylogenetic network is called a *galled-tree network* if no two recombination cycles share an edge. A recombination cycle of a galled-tree network is sometimes referred to as a *gall*.

Note that in the original definition of galled-tree network[16] it is required that recombination cycles do not share vertices. It is easy to see that our modification is only a minor difference (one can be transformed to the other easily) introduced for technical reasons.

In the following definition we describe two basic operations on the matrices which we will use frequently.

Definition 4. Given a haplotype matrix M, let S be a subset of characters of M. The matrix $M[S]$ is the sub-matrix of M restricted to the columns in S. We will assume that the names of columns in $M[S]$ are the same as in the original matrix M. Let x be a binary sequence of length $|S|$. By $M[S] - x$, we denote the sub-matrix of $M[S]$ from which we remove all rows whose strings are identical to x.

Definition 5. We say that the characters c_1 and c_2 conflict in a haplotype matrix M if $M[c_1, c_2]$ contains all three pairs $[0, 1]$, $[1, 0]$ and $[1, 1]$. The *conflict graph* G_M has the vertex set $\{1, \ldots, m\}$ and for every two characters c_1 and c_2, (c_1, c_2) is an (undirected) edge of G_M if they conflict.

The following characterization of the existence of galled-tree network was obtained in [6].

Theorem 1. *([6]) Given a haplotype matrix M, matrix M can be explained by a galled-tree network if and only if every nontrivial component (having at least two vertices) K of the conflict graph G_M satisfies the following conditions:*

*(1) K is bipartite with partitions L and R such that all characters in L are smaller than all characters in R (*the ordered component property*); and*

(2) there exists a sequence $x \neq 0^{|K|}$ such that $M[K] - x$ has no conflicting characters.

Note that the necessity part of Theorem 1 has been proved in[9].

2 Inferring haplotypes via galled-tree network

In this section we introduce both the perfect phylogeny haplotype (PPH) problem and our galled-tree network haplotype (GTNH) problem.

Definition 6. Given a genotype $n \times m$ matrix A, find a $2n \times m$ haplotype matrix B with values in $\{0, 1\}$, where rows $2i - 1$ and $2i$ of B represent haplotypes for the genotype in row i of A. We say that B is *inferred* from A if and only if for every character $c \in \{1, \dots, m\}$,

- if $A(i, c) \in \{0, 1\}$, then $B(2i - 1, c) = B(2i, c) = A(i, c)$; and
- if $A(i, c) = 2$, $B(2i - 1, c) \neq B(2i, c)$.

Let $\mathcal{X}_A = \{x_{rc_1c_2};\ A(r, c_1) = A(r, c_2) = 2,\ c_1 < c_2\}$ be the set of Boolean variables. The value of the variable $x_{rc_1c_2}$ determines the way how the pair of 2's in the row r and columns c_1 and c_2 is resolved. Define assignment $I_B : \mathcal{X}_A \to \{0, 1\}$ as follows. Let $I_B(x_{rc_1c_2}) = 0$ if the pair is resolved equally in B, i.e, $\begin{pmatrix} 2 & 2 \end{pmatrix} \to \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$; and $I_B(x_{rc_1c_2}) = 1$ if the pair is resolved unequally in B, i.e, $\begin{pmatrix} 2 & 2 \end{pmatrix} \to \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Note that specifying the assignment I_B is equivalent to specifying the matrix B (up to swapping rows $2i - 1$ and $2i$, for any $i \in \{1, \dots, n\}$).

Definition 7. Given a genotype matrix A, we say that A can be *explained* by phylogenetic tree/galled-tree network if there exists a haplotype matrix B inferred from A such that B can be explained by a phylogenetic tree/galled-tree network.

Problem 1. (Perfect phylogeny haplotype (PPH) problem/Galled-tree network haplotype (GTNH) problem) Given a genotype matrix A, decide if A can be explained by a phylogenetic tree/galled-tree network.

In what follows we will need a characterization of the existence of the solution to the PPH problem. We will use the characterization given by Bafna et al. in [1].

Definition 8. Given a genotype matrix A, for every $x, y \in \{0, 1\}$, we say that a pair of columns c_1, c_2 *induces* $[x, y]$ *in* A, if $A[c_1, c_2]$ contains at least one of the pairs $[x, y]$, $[2, y]$ and $[x, 2]$.

Theorem 2. *(Bafna et al.[1]) Given a genotype matrix A, there is a solution to PPH problem on A if and only if no two columns $c_1 < c_2$ induce all three pairs $[1, 1]$, $[0, 1]$ and $[1, 0]$ in A, and there exists an assignment I_B such that*

(1) for every $x_{rc_1c_2}, x_{r'c_1c_2} \in \mathcal{X}_A$*,* $I_B(x_{rc_1c_2}) = I_B(x_{r'c_1c_2})$*;*

(2) for every $x_{rc_1c_2} \in \mathcal{X}_A$*,* $I_B(x_{rc_1c_2}) = 0$ *if* c_1, c_2 *induce* $[1,1]$ *in* A*;*

(3) for every $x_{rc_1c_2} \in \mathcal{X}_A$*,* $I_B(x_{rc_1c_2}) = 1$ *if* c_1, c_2 *induce both* $[0,1]$ *and* $[1,0]$ *in* A*; and*

(4) for every $x_{rc_1c_2}, x_{rc_1c_3}, x_{rc_2c_3} \in \mathcal{X}_A$*,* $I_B(x_{rc_1c_2}) + I_B(x_{rc_1c_3}) + I_B(x_{rc_2c_3}) = 0$*.*

When considering the GTNH problem, the above rules (except for (4)) do not apply since violating rules (1)–(3) just creates a conflict between corresponding columns/characters in B. We have the following observation about the relationship between the conflict graph of B and the assignment I_B.

Observation 2. *Given a genotype matrix* A*, infer a matrix* B*. Characters* $c_1 < c_2$ *conflict in* B *if and only if at least one of the following is true:*

(C0) c_1, c_2 *induce all three pairs:* $[1,1]$*,* $[0,1]$ *and* $[1,0]$ *in* A*;*

(C1) $I_B(x_{rc_1c_2}) \neq I_B(x_{r'c_1c_2})$ *for some* $x_{rc_1c_2}, x_{r'c_1c_2} \in \mathcal{X}_A$*;*

(C2) c_1, c_2 *induce* $[1,1]$ *in* A*, and there exists* $x_{rc_1c_2} \in \mathcal{X}_A$ *such that* $I_B(x_{rc_1c_2}) = 1$*;*

(C3) c_1, c_2 *induce both* $[0,1]$ *and* $[1,0]$ *in* A*, and there exists* $x_{rc_1c_2} \in \mathcal{X}_A$ *such that* $I_B(x_{rc_1c_2}) = 0$*.*

3 Special instance of GTNH problem

The GTNH problem for general matrices seems to be very hard, therefore we study special instances of this problem. In[5], we studied the instance when the input genotype matrix has the property that every row that contains 2, contains 1 as well and we required that the solution (haplotype matrix) can be explained by a galled-tree network with galls having exactly two mutation edges. In the instance studied in this paper, we put no restriction on the galled-tree networks but assume a little bit stronger assumption about the input genotype matrix.

Definition 9. Given a genotype matrix A, we say that a pair of characters is *active* if it contains $[2,2]$, or it induces all three pairs $[1,1]$, $[0,1]$ and $[1,0]$. Further, we say that a pair c_1, c_2 is *weakly active* if either it is active, or if there is a character c_3 such that c_1, c_3 and c_2, c_3 are both active pairs. Now, we say that A has the *weak diagonal property* if every weakly active pair of characters induces both $[0,1]$ and $[1,0]$.

In fact, many data sets, including real data and simulated data sets, satisfy the WD property. We tested Daly's genotype data ([3]), which has 103 columns and 387 rows. The data contains 11 blocks, where the evolutionary history for haplotypes on each block is claimed to have very few recombinations or in other words satisfy the assumption of galled-tree network. The genotype matrices for 9 out of the 11 blocks satisfy WD property. The remaining two blocks related genotype matrices have the WD property after removing one column from each of them respectively. We also tested the WD property on hundreds of simulated matrices based on Hudson's simulation program ([11]). In particular, for each binary matrix generated using Hudson's program, we randomly repeat each row 2 to 4 times and randomly pair two rows to get a genotype. In average, 1/2 of the simulated genotype matrices satisfy the WD property.

Definition 10. For every $x, y, z \in \{0,1\}$, we say that a triple of columns c_1, c_2, c_3 *induces* $[x,y,z]$ *in* A, if $A[c_1, c_2, c_3]$ contains at least one of the triples $[x,y,z]$, $[2,y,z]$, $[x,2,z]$ and $[x,y,2]$.

Note that if columns c_1, c_2, c_3 induce $[x, y, z]$ in A then every matrix B inferred from A contains $[x, y, z]$ in $B[c_1, c_2, c_3]$.

If A has the weak diagonal property, then for every $x_{rc_1c_2} \in \mathcal{X}_A$ such that $I_B(x_{rc_1c_2}) = 0$, c_1 and c_2 conflict in B. By Observation 2, we have the following observation.

Observation 3. *Given a genotype matrix A with the weak diagonal property, let B be a haplotype matrix inferred from A. Characters $c_1 < c_2$ conflict in G_B if and only if they induce $[0, 1]$, $[1, 0]$ and $[1, 1]$ in A, or there is $x_{rc_1c_2} \in \mathcal{X_A}$ such that $I_B(x_{rc_1c_2}) = 0$. Consequently, if a pair $c_1 < c_2$ conflicts in B, then it is active in A.*

First, we will study some basic characteristics of matrices with the weak diagonal property. In the following subsections, we will use these results to reduce the problem to a hypergraph covering problem similar to the one introduced in [5].

3.1 Characteristics of matrices with the weak diagonal property

In this subsection we observe some properties of matrices with the weak diagonal property which can be explained by a galled-tree network.

Claim 1. *Given an $n \times m$ genotype matrix A, assume that A has a row r which contains one 2. Let A' be a matrix obtained from A by replacing r with the $2 \times m$ matrix inferred from r. Then A can be explained by a galled-tree network N if and only if A' can be explained by N.*

Proof. First, note that $\mathcal{X}_A = \mathcal{X}_{A'}$. Hence, there is a one-to-one correspondence between matrices B and B' inferred from A and A': $I_B = I_{B'}$. The matrix B' can be obtained from B by duplicating two rows. The claim follows by Observation 1. □

Claim 2. *Given a genotype matrix A with the weak diagonal property, let B be a matrix inferred from A such that it can be explained by a galled-tree network. Then for every triple of 2's occurring in columns $c_1 < c_2 < c_3$ and in a same row r, exactly one of $I_B(x_{rc_1c_2}), I_B(x_{rc_1c_3}), I_B(x_{rc_2c_3})$ is equal to 0.*

Proof. The values have to satisfy condition (4) of Theorem 2, i.e., $I_B(x_{rc_1c_2}) + I_B(x_{rc_1c_3}) + I_B(x_{rc_2c_3}) = 0$. This implies that either all three variables are mapped to 0 or exactly one of them. In the latter case we are done. In the former case, due to the weak diagonal property, by Observation 3, we have that all three pairs c_1, c_2, c_1, c_3 and c_2, c_3 conflict in B. Hence, the conflict graph G_B is not bipartite, a contradiction with Theorem 1. □

The following two corollaries easily follow from the above claim.

Corollary 1. *Given a genotype matrix A with the weak diagonal property, let B be a matrix inferred from A such that it can be explained by a galled-tree network. Then for every four 2's occurring in columns $c_1 < c_2 < c_3 < c_4$ and in a same row r, there are pairs $d_1 < d_2$ and $d_3 < d_4$ such that $\{d_1, d_2, d_3, d_4\} = \{c_1, c_2, c_3, c_4\}$, $I_B(x_{rd_1d_2}) = I_B(x_{rd_3d_4}) = 0$ and for every pair $d < d'$, where $d, d' \in \{c_1, c_2, c_3, c_4\}$, different from $d_1 < d_2$ and $d_3 < d_4$, $I_B(x_{rdd'}) = 1$.*

Corollary 2. *Given a genotype matrix A with the weak diagonal property, if A has a row r with at least five 2's, then A cannot be explained by a galled-tree network.*

Proof. Let $C_r = \{c_1, \ldots, c_\ell\}$ be the ordered set of all columns containing 2 in the row r. Suppose $\ell \geq 5$. Assume that A can be explained by a galled-tree network, and let B be the corresponding haplotype matrix. By Claim 2, for every three character $c_i < c_j < c_k \in C_r$ exactly one of

$I_B(x_{rc_1c_2}), I_B(x_{rc_1c_3}), I_B(x_{rc_2c_3})$ is equal to 0. Without loss of generality we can assume that for triple c_1, c_2, c_3, we have $I_B(x_{rc_1c_2}) = 0$. It follows that values $I_B(x_{rc_1c_3})$, $I_B(x_{rc_2c_3})$, $I_B(x_{rc_1c_4})$, $I_B(x_{rc_2c_4})$, $I_B(x_{rc_1c_5})$, $I_B(x_{rc_2c_5})$ are all equal to 1, and hence the values $I_B(x_{rc_3c_4})$, $I_B(x_{rc_3c_5})$, $I_B(x_{rc_4c_5})$ are all equal to 0, a contradiction. □

From now on, we will assume that the input genotype matrix (with the weak diagonal property) has either no, two, three or four 2's in each row, since otherwise it either has no solution (Corollary 2) or can be converted to a matrix with this property (Claim 2). In the following definition we define a hypergraph assigned to a genotype matrix and its coverings.

Definition 11. Given an $n \times m$ genotype matrix A with the weak diagonal (WD) property, the *genotype hypergraph* H_A of A has the set of characters $\{1, \ldots, m\}$ as a vertex set, and for every row r of A containing at least two 2's, say in columns $c_1, \ldots, c_k$ there is a hyperedge $e_r = \{c_1, \ldots, c_k\}$. Furthermore, for every two columns c and c' inducing $[0,1]$, $[1,0]$ and $[1,1]$ in A, there is a *forced* hyperedge $[c, c']$ in H_A. The hypergraph H_A does not contain any other hyperedges. For distinction purpose, we call a 2-edge an *unforced* 2-edge.

We say that a graph G on the vertex set $V(H_A)$ *covers* a hypergraph H_A with hyperedges of cardinality at most 4, if G can be obtained as follows:

- for every forced 2-edge $[c_1, c_2]$ of H_A, add the edge (c_1, c_2) in G;
- for every unforced 2-edge $\{c_1, c_2\}$ of H_A, make a choice whether to add the edge (c_1, c_2) in G;
- for every 3-edge $\{c_1, c_2, c_3\}$ of H_A, add exactly one of the edges (c_1, c_2), (c_2, c_3) and (c_1, c_3) to G;
- for every 4-edge $\{c_1, c_2, c_3, c_4\}$ of H_A, add exactly two disjoint edges (d_1, d_2) and (d_3, d_4) such that $\{d_1, d_2, d_3, d_4\} = \{c_1, c_2, c_3, c_4\}$ to G.

Now, we can characterize all possible conflict graphs of matrices inferred from an input genotype matrix as follows.

Lemma 3. *Given a genotype matrix A with the weak diagonal property, let B be a haplotype matrix inferred from A which can be explained by a galled-tree network. Then the conflict graph G_B of B is a covering of the genotype hypergraph H_A of A.*

Conversely, let G be a covering for the graph H_A. Each way of finding the covering G from H_A (a collection of choices for every unforced hyperedge of H_A), defines a haplotype matrix B. This haplotype matrix B can be inferred from A and the conflict graph of B is G.

Proof. Let B be a haplotype matrix inferred from A which can be explained by a galled-tree network. By Corollary 2, A contains at most four 2's in every row. By Observation 3, characters $c_1 < c_2$ conflict if and only if they induce $[0,1]$, $[1,0]$ and $[1,1]$ in A, or if $I_B(x_{rc_1c_2}) = 0$ for some $x_{rc_1c_2} \in \mathcal{X}_A$. It follows by Definition 11, Claim 2 and Corollary 1 that G_B covers H_A.

Conversely, let G be a covering for H_A. Then A contains at most four 2's in every row, and there exists a choice for every (unforced) k-edge as described in Definition 11. By Claim 2 and Corollary 1, each such collection of choices for every unforced k-edge, defines values $I_B(x_{rcc'})$ for every $x_{rcc'} \in \mathcal{X}_A$ satisfying condition (4) of Theorem 2. Hence, also a haplotype matrix B can be inferred from A. By Observation 3, the conflict graph G_B is isomorphic to G. □

3.2 Characterization of conflict graphs of haplotype matrices explainable by a GTN inferred from a genotype matrix with the WD property

The following claim is crucial in restricting the possible cases we need to study.

Claim 3. *Given a genotype matrix A with the weak diagonal property, let B be a matrix inferred from A which can be explained by a galled-tree network. Let c_1, c_2, c_3 be three characters such that both pairs c_1, c_2 and c_2, c_3 conflict in B. Let K be the component containing c_1, c_2, c_3 and x a vector such that $B[K] - x$ has no conflicts. Then $x[c_1, c_2, c_3]$ is either $[0, 1, 1]$ or $[1, 1, 0]$.*

Proof. By Observation 3, both pairs c_1, c_2 and c_2, c_3 are active. Hence, the pair c_1, c_3 is weakly active and satisfies the weak diagonal condition, i.e., it induces $[0, 1]$ and $[1, 0]$ in A. Hence, there are two rows in B containing those two pairs in $B[c_1, c_3]$. Note that c_1 and c_3 cannot conflict in B, otherwise G_B is not a bipartite which would violate (1) of Theorem 1. Therefore, since x is a sequence in one of the rows of $B[K]$, we have $x[c_1, c_3] \neq [1, 1]$. On the other hand x has to remove conflicts between c_1, c_2 and between c_2, c_3, i.e, $x[c_1, c_2], x[c_2, c_3] \neq [0, 0]$. There are only three possibilities left for the value of $x[c_1, c_2, c_3]$: $[0, 1, 0]$, $[0, 1, 1]$ and $[1, 1, 0]$.

To finish the proof it is enough to consider (and exclude) the case $x[c_1, c_2, c_3] = [0, 1, 0]$. Since, c_1, c_2 conflict in B, there is a row r containing triple $[1, 1, y]$, where $y \in \{0, 1\}$, in $B[c_1, c_2, c_3]$. If $y = 1$, then c_1 and c_3 conflict in B, a contradiction. On the other hand, if $y = 0$, then $B[K] - x$ still contains the row r. Thus, character c_2 and c_3 still conflict in $B[K] - x$, a contradiction. □

Claim 3 is a powerful tool which helps to characterize all possible conflict graphs of haplotype matrices explainable by a galled-tree network inferred from a given genotype matrix with the WD property.

Corollary 3. *Given a genotype matrix A with the weak diagonal property, let B be a matrix inferred from A which can be explained by a galled-tree network. Every vertex in G_B has degree at most 2, i.e., G_B consists of cycles and paths.*

Proof. Assume for the contrary that there is a character c with degree at least 3 in the conflict graph G_B. Let c_1, c_2, c_3 be three neighbors of c in G_B. Let K be the component containing c, c_1, c_2, c_3 and x a sequence such that $B[K] - x$ does not contain any conflict. By Claim 3, $x[c_1, c, c_2]$ is either $[0, 1, 1]$ or $[1, 1, 0]$. Without loss of generality assume it is $[0, 1, 1]$. By Claim 3, $x[c_1, c, c_3] = [0, 1, 1]$. Now, $x[c_2, c, c_3] = [1, 1, 1]$, which contradicts Claim 3. □

Corollary 4. *Given a genotype matrix A with the weak diagonal property, let B be a matrix inferred from A which can be explained by a galled-tree network. Then each component of G_B is a path of length at most 3 (contains at most three edges).*

Proof. By Corollary 3, each component of G_B is either a cycle or a path. Assume to the contrary that a component K of G_B is a cycle or a path of length at least 4. Since by Theorem 1, K is a bipartite, if K is a cycle then its length is at least 4. Hence, K contains a path $(c_1, c_2, c_3, c_4, c_5)$, where characters c_1 and c_5 can be the same (in the case when K is a cycle of length 4). Let x be a sequence such that $B[K] - x$ does not contain any conflict. By Claim 3, each of $x[c_1, c_2, c_3]$, $x[c_2, c_3, c_4]$ and $x[c_3, c_4, c_5]$ is either $[0, 1, 1]$ or $[1, 1, 0]$. Obviously, this is not possible, a contradiction. □

3.3 Dealing with rows with two 2's

In this section we deal with the problem how to resolve rows containing only two 2's. The general strategy is to resolve such rows unequally if it helps to avoid conflict, and otherwise equally. The following two claims show the correctness of this strategy.

Claim 4. *Given a genotype matrix A with the weak diagonal property, consider any two columns $c_1 < c_2$. Let $r_1, \ldots, r_k$ be rows containing 2's in columns c_1, c_2 and no other 2's, and $r'_1, \ldots, r'_\ell$ rows containing 2's in columns c_1, c_2 and at least one 2 in some other column. Assume that $k \geq 1$ and that c_1 and c_2 do not induce $[1,1]$ in A. Let B be a haplotype matrix inferred from A such that*

- *B can be explained by a galled-tree network;*
- *for some $i = 1, \ldots, k$, $I_B(x_{r_i c_1 c_2}) = 0$; and*
- *for every $i = 1, \ldots, \ell$, $I_B(x_{r'_i c_1 c_2}) = 1$.*

Then the matrix B' such that for every $i = 1, \ldots, k$, $I_{B'}(x_{r_i c_1 c_2}) = 1$ and for every other $x_{rcc'} \in \mathcal{X}_A$, $I_{B'}(x_{rcc'}) = I_B(x_{rcc'})$, can be explained by a galled-tree network as well.

Proof. The conflict graph $G_{B'}$ differs from the conflict graph G_B only by not containing an edge (c_1, c_2). Hence, it satisfies condition (1) of Theorem 1. Let us verify the second condition. Consider component of G_B that contains c_1 and c_2. By Corollary 4, it is a path of length at most 3. Therefore, in $G_{B'}$ this component is split into two, each containing one of c_1, c_2. Consider a component K of $G_{B'}$. Since it contains at most one of c_1 and c_2 then $B[K]$ and $B'[K]$ contains the same set of sequences. Let K' be a component of G_B containing K. There is a sequence x such that $B[K'] - x$ has no conflict. Obviously, $B[K] - x[K]$ has no conflict as well, and hence also $B'[K] - x[K]$. By Theorem 1, B' can be explained by a galled-tree network.

□

Claim 5. *Given a genotype matrix A with the weak diagonal property, consider any two columns $c_1 < c_2$. Let $r_1, \ldots, r_k$ be rows containing 2's in columns c_1, c_2 and no other 2's, and $r'_1, \ldots, r'_\ell$ rows containing 2's in columns c_1, c_2 and at least one 2 in some other column. Assume that $k \geq 1$. Let B be a haplotype matrix inferred from A such that B can be explained by a galled-tree network. If either for some $i = 1, \ldots, \ell$, $I_B(x_{r'_i c_1 c_2}) = 0$ or c_1 and c_2 induce $[1,1]$ in A, then the matrix B' such that for every $i = 1, \ldots, k$, $I_{B'}(x_{r_i c_1 c_2}) = 0$ and for every other $x_{rcc'} \in \mathcal{X}_A$, $I_{B'}(x_{rcc'}) = I_B(x_{rcc'})$, can be explained by a galled-tree network as well.*

Proof. The conflict graph $G_{B'}$ is the same as the conflict graph G_B. Hence, it satisfies the first condition of Theorem 1. It is enough to verify the second condition. Consider a component K of $G_{B'} = G_B$. If it does not contain c_1 and c_2 then $B[K]$ and $B'[K]$ contains the same set of sequences, and condition (2) easily follows.

Since c_1 and c_2 conflict, we only need to consider the case when K contains both c_1 and c_2. If K does not contain any other character, condition (2) trivially holds. Without loss of generality, assume that K contains a character c_3 such that c_2, c_3 conflicts in B. Similarly, if for every $i = 1, \ldots, k$, $I_B(x_{r_i c_1 c_2}) = 0$, then $B' = B$, and the claim follows trivially. Hence, we can assume that there is $i \in \{1, \ldots, k\}$ such that $I_B(x_{r_i c_1 c_2}) = 1$. We will show that this is not possible. Since, B can be explained by a galled-tree network, there is a sequence x such that $B[K] - x$ has no conflict. By Claim 3, $x[c_1, c_2, c_3]$ is either $[0,1,1]$ or $[1,1,0]$. If $A[c_1, c_2, c_3]$ contains $[2,2,0]$ in row r_i then $B[c_1, c_2, c_3]$ contains $[0,1,0]$ in row $2r_i - 1$ or row $2r_i$. Hence, if $x[c_1, c_2, c_3] = [0,1,1]$, x cannot remove conflict between c_1 and c_2, and if $x[c_1, c_2, c_3] = [1,1,0]$, x cannot remove conflict between c_2 and c_3. On the other hand, if $A[c_1, c_2, c_3]$ contains $[2,2,1]$, then one of the rows $2r_i - 1$ and $2r_i$ contains $[1,1]$ in $B[c_1, c_3]$ and hence c_1 and c_3 conflict in B. This is a contradiction, since by Theorem 1, G_B is bipartite. □

The above two claims do not allow for resolving all rows with only two 2's before solving the problem in the case when the input genotype matrix has the weak diagonal property. However, in some certain cases such rows can be resolved before building the genotype hypergraph.

Corollary 5. *Given an $n \times m$ genotype matrix A with the weak diagonal property, for any row r containing only two 2's, say in columns $c_1 < c_2$,*

- *if c_1, c_2 induce $[1, 1]$ in A, replace r with two rows of the $2 \times m$ matrix R inferred from r such that $I_R(x_{rc_1c_2}) = 0$;*
- *if c_1, c_2 do not induce $[1, 1]$ in A and every row containing 2's in c_1 and c_2 contains only these two 2's, replace r with two rows of the $2 \times m$ matrix R inferred from r such that $I_R(x_{rc_1c_2}) = 1$.*

The new matrix A' obtained like that can be explained by a galled-tree network if and only if A does.

The above corollary follows by Observation 1 and Claims 4 and 5. Hence, we can assume that the input genotype matrix contains a row with two 2's, say in columns c_1 and c_2, only if there is another row with at least three 2's with two of them being in columns c_1 and c_2. We say that such a genotype matrix is *reduced*. For the remaining rows with two 2's, the resolution depends on the resolution of the other rows containing more than two 2's.

Lemma 4. *Given an $n \times m$ reduced genotype matrix A with the weak diagonal property, the matrix A can be explained by a galled-tree network if and only if there is a matrix B which can be explained by a galled-tree network and is inferred from A such that for every row r with only two 2's, say in columns c_1 and c_2,*

- *$I_B(x_{rc_1c_2}) = 0$, if there is a row r' with at least three 2's such that $x_{r'c_1c_2} \in \mathcal{X}_A$ and $I_B(x_{r'c_1c_2}) = 0$;*
- *$I_B(x_{rc_1c_2}) = 1$, otherwise.*

In other words, starting with a reduced genotype matrix A, a row with only two 2's is never resolved in a way which would introduce a conflict which is not introduced by another row with more than two 2's. This leads us to the following definition of finding a hypergraph covering of a given hypergraph.

Definition 12. We say that a graph G is a *canonical covering* of a hypergraph H if it is a covering for H such that no unforced 2-edge $\{c_1, c_2\}$ of H contributes the edge (c_1, c_2) to G, i.e., the edge (c_1, c_2) is formally added to G if and only if it is added to G by some other hyperedge than the unforced 2-edge $\{c_1, c_2\}$.

Note that although adding edges to G for some unforced 2-edges of H does not affect the resulting graph covering, it affects the process of inferring which defines the matrix B.

As another corollary we have the following claim about finding a canonical covering.

Corollary 6. *Given a reduced genotype matrix A with the weak diagonal property, if A can be explained by a galled-tree network then there is a haplotype matrix B inferred from A which can be explained by a galled-tree network and its conflict graph is a canonical covering of H_A.*

There is an easy characterization how the rows with only two 2's are inferred for haplotype matrices defined during a process of finding a canonical covering.

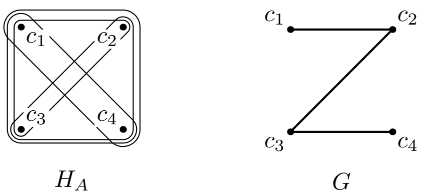
Observation 4. *Given a reduced genotype matrix A with the weak diagonal property, let G be a graph that canonically covers H_A and let B be a haplotype matrix defined during the process of finding a canonical covering G. Then for every row r with only two 2's, say in columns c_1 and c_2, $I_B(x_{rc_1c_2}) = 0$ if and only if (c_1, c_2) is an edge in G.*

Proof. By Lemma 3, $G_B = G$. Consider a row r with only two 2's, in columns c_1 and c_2. Obviously, the pair c_1, c_2 is active and hence, if $I_B(x_{rc_1c_2}) = 0$ then c_1, c_2 conflict in B, i.e., $(c_1, c_2) \in E(G)$. For the converse, assume that c_1, c_2 conflict in B, i.e., there is edge (c_1, c_2) in G. Since G canonically covers H_A and it defines B, there is either forced 2-edge $[c_1, c_2]$ in H_A, or a k-edge containing c_1 and c_2 with $k > 2$ and the edge (c_1, c_2) was added to G when processing this k-edge. Hence, when processing the unforced 2-edge $e_r = \{c_1, c_2\}$, the edge (c_1, c_2) is added to G and hence $I_B(x_{rc_1c_2}) = 0$. □

3.4 Reduction to the extended hypergraph covering problem

Now, we are ready to convert the GTNH problem in the case when the input genotype matrix has the weak diagonal property to a hypergraph covering problem. It easy to see that a conflict graph of a matrix inferred from a given genotype matrix A which can be explained by GTN, is also a covering of hypergraph H_A which satisfies all the observed conditions. The following examples show that these conditions are not sufficient.

A	c_1	c_2	c_3	c_4
	2	2	2	0
	0	2	2	2
	2	0	2	2
	2	2	0	2



B	c_1	c_2	c_3	c_4
	0	0	1	0
	1	1	0	0
	0	1	1	0
	0	0	0	1
	0	0	1	1
	1	0	0	0
	1	1	0	0
	0	0	0	1

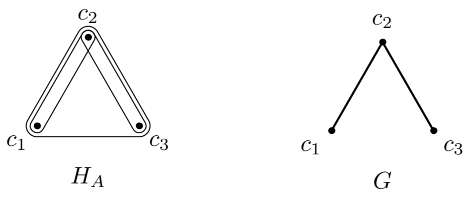
Figure 1: Example of a genotype matrix A with the weak diagonal property, the corresponding GI problem and its solution, and a haplotype matrix B inferred from A with the conflict graph equivalent to this solution. The matrix B does not satisfy condition (2) of Theorem 1.

Example 1. Figure 1 shows an example of a genotype matrix A with four characters. It is easy to see that A has the weak diagonal property since every pair of columns induces $[0, 1]$ and $[1, 0]$. The corresponding hypergraph H_A has four 3-edges. One possible solution is a covering G of H_A which can be constructed as follows: in 3-edge $\{c_1, c_2, c_3\}$ we select edge (c_1, c_2), in 3-edge $\{c_2, c_3, c_4\}$ we select edge (c_2, c_3), in 3-edge $\{c_1, c_3, c_4\}$ we select edge (c_3, c_4) and finally, in 3-edge $\{c_1, c_2, c_4\}$ we select edge (c_1, c_2). Note that this covering contains only paths of length at most 3 and is canonical since A does not contain any rows with only two 2's. A haplotype matrix B corresponds to this selection of edges in G, hence its conflict graph is G. However, removal of any row from matrix B is not able to eliminate all conflicts in B. Hence, the condition (2) of Theorem 1 is not satisfied and B cannot be explained by a galled-tree network.

Consider now another graph G' on the same set of characters with only two edges (c_1, c_4) and (c_2, c_3). This graph is a covering for H_A as well. Obviously, the haplotype matrix corresponding to this selection of edges for every hyperedge of H_A satisfies condition (2) of Theorem 1, i.e., can be explained by a galled-tree network. Hence, one of the additional constrains we will require from the new hypergraph covering problem is the minimality of the number of edges.

Example 2. Figure 2 shows an example of a reduced genotype matrix A with three characters which satisfies the weak diagonal property. It has a row with only two 2's, in columns c_1 and c_3. However,

$$\begin{array}{c|ccc} A & c_1 & c_2 & c_3 \\ \hline & 2 & 2 & 2 \\ & 2 & 0 & 2 \\ & 0 & 1 & 1 \\ & 1 & 1 & 0 \\ & 0 & 1 & 0 \end{array}$$



$$\begin{array}{c|ccc} B & c_1 & c_2 & c_3 \\ \hline & 1 & 1 & 0 \\ & 0 & 0 & 1 \\ & 1 & 0 & 0 \\ & 0 & 0 & 1 \\ & 0 & 1 & 1 \\ & 1 & 1 & 0 \\ & 0 & 1 & 0 \end{array}$$

Figure 2: Example of a genotype matrix A with the weak diagonal property, the corresponding GI problem and its unique solution, and a haplotype matrix B inferred from A with the conflict graph equivalent to this solution. The matrix B does not satisfy condition (2) of Theorem 1.

this row could not be resolved since c_1, c_3 do not induce $[1, 1]$ and there is another row with three 2's containing 2's in c_1 and c_3. The corresponding hypergraph H_A has one 3-edge, one unforced 2-edge and two forced 2-edges. There is only one graph covers H_A such that it contains only paths of length at most 3, the graph G. In particular, G can be constructed as follows: we have to select forced 2-edges, in 3-edge $\{c_1, c_2, c_3\}$ we select edge (c_1, c_2) (respectively, (c_2, c_3), it will not affect the hypergraph covering), and finally, we do not add unforced 2-edge $\{c_1, c_3\}$ to G. Note that this is a canonical covering. A haplotype matrix B corresponds to this selection of edges in G, hence its conflict graph is G. However, removal of any row from matrix B is not able to eliminate all conflicts in B. Hence, the condition (2) of Theorem 1 is not satisfied and B cannot be explained by a galled-tree network.

Since we have examined all possible ways how to infer a haplotype matrix from A, it follows by Corollary 6 that A cannot be explained by a galled-tree network.

The crucial reason why in the above example A cannot be explained by a galled-tree network even though we can easily find a valid hypergraph covering of the hypergraph of A is the fact that columns c_1, c_2, c_3 induce $[0, 1, 0]$ in A. The following claim captures this property.

Claim 6. *Given a genotype matrix A with the weak diagonal property, let B be a haplotype matrix inferred from A which can be explained by a galled-tree network. Let c_1, c_2, c_3 be three characters (not necessarily ordered in this way). If they induce $[0, 1, 0]$ in A then the conflict graph G_B cannot contain both edges (c_1, c_2) and (c_2, c_3).*

Proof. Assume for the contrary that both pairs c_1, c_2 and c_2, c_3 conflict in B. Hence, $B[c_1, c_2]$ (respectively, $B[c_2, c_3]$) contains all three pairs $[0, 1]$, $[1, 0]$ and $[1, 1]$. By Observation 3, the pair c_1, c_3 is weakly active in A, hence $B[c_1, c_3]$ contains $[0, 1]$ and $[1, 0]$. Since B can be explained by a galled-tree network, by Theorem 1, characters c_1 and c_3 cannot conflict, i.e., $B[c_1, c_3]$ cannot contain pair $[1, 1]$. Hence, $B[c_1, c_2, c_3]$ contains $[1, 0, 0]$, $[1, 1, 0]$, $[0, 0, 1]$ and $[0, 1, 1]$. Since the triple c_1, c_2, c_3 induces $[0, 1, 0]$, $B[c_1, c_2, c_3]$ contains also triple $[0, 1, 0]$. B does not satisfy condition (2) of Theorem 1, a contradiction. □

The following definition is a generalization of Definition 11 which incorporates constrains shown in Claim 6.

Definition 13. Given an $n \times m$ genotype matrix A with the weak diagonal property, the *extended genotype hypergraph* $\bar{H}_A$ of A has the set of characters $\{1, \ldots, m\}$ and contains all hyperedges of genotype hypergraph H_A. In addition, it contains ordered 3-edges, called *switches*. For every three characters c_1, c_2, c_3 such that $c_1 < c_3$ and c_1, c_2, c_3 induce $[0,1,0]$ in A, $\bar{H}_A$ contains the switch (c_1, c_2, c_3).

We say that a graph G is a *canonical covering* of extended hypergraph $\bar{H}_A$ if G is a canonical covering for underlying H_A and in addition for every switch (c_1, c_2, c_3), it contains at most one of the edges (c_1, c_2) and (c_2, c_3).

Using Claim 6 we can immediately extend the result in Corollary 6.

Corollary 7. *Given a genotype matrix A with the weak diagonal property, if A can be explained by a galled-tree network then there is a haplotype matrix B inferred from A which can be explained by a galled-tree network and its conflict graph is a canonical covering of the extended hypergraph $\bar{H}_A$.*

Problem 2 (Extended Hypergraph Covering (EHC) Problem)**.** Given an extended hypergraph $\bar{H}$, determine whether it is possible to find a canonical covering graph G for $\bar{H}$ such that each component of G is a path of length at most 3 satisfying the ordered component property. In case it is possible, find such a covering G with the minimum number of edges.

The following lemma shows that we can solve the GTNH problem in the case when the input genotype matrix has the weak diagonal property if we can solve the above problem.

Lemma 5. *The GTNH problem in the case when the input genotype matrix has the weak diagonal property can be solved using the EHC problem.*

Proof. Consider an input reduced genotype matrix A with the weak diagonal property. We will show that A can be explained by a galled-tree network if and only if it is possible to find a canonical covering graph G for the extended hypergraph $\bar{H}_A$ such that each component of G is a path of length at most 3. The forward implication follows easily by Corollaries 4 and 7. For the converse, we need the following claim.

Claim 7. *Let G be a canonical covering of extended hypergraph $\bar{H}_A$ with the minimum number of edges such that all its components are paths of length at most* 3 *and satisfy the ordered component property. Then there exists a matrix B inferred from A which can be explained by a galled-tree network and its conflict graph is G.*

Proof. There are several ways how to find the covering G from $\bar{H}_A$. It is sufficient to show that one of these coverings defines a haplotype matrix B which satisfies condition (2) of Theorem 1. Then since all components are paths of length at most 3 with the ordered component property, condition (1) holds and the claim follows.

Consider a component K of G. If it is a singleton or an edge, condition (2) is trivially satisfied. Second, assume that K is a path of length 2, say (c_1, c_2, c_3). Consider any B defined by any covering G from $\bar{H}_A$. As in the proof of Claim 6, the submatrix $B[K]$ must contain triples $[1,0,0]$, $[1,1,0]$, $[0,1,1]$ and $[0,0,1]$. It can also contain triples $[0,0,0]$ and $[0,1,0]$, all other triples would introduce a conflict between c_1 and c_3. It is easy to see that the component K satisfies condition (2) if and only if $B[K]$ does not contain $[0,1,0]$. We will show that $B[K]$ does not contain $[0,1,0]$.

Assume to the contrary that $B[K]$ contains $[0,1,0]$. It must be inferred from some sequence in $A[K]$. However, since G contains both edges (c_1, c_2) and (c_2, c_3), $\bar{H}_A$ does not contain switch (c_1, c_2, c_3), and hence c_1, c_2, c_3 do not induce $[0,1,0]$ in A. Therefore, there are only four possibilities: $[2,2,2]$, $[2,1,2]$, $[2,2,0]$ and $[0,2,2]$. In the first two cases, to infer $[0,1,0]$ in B, we would have to also infer $[1,0,1]$ (in the first case) or $[1,1,1]$ (in the second case). Since the pair c_1, c_3 is weakly active, it would conflict in B, a contradiction. Obviously, the third and fourth cases are symmetric, hence we will consider only the first of them. Let r be the row containing $[2,2,0]$ in $A[K]$. To infer $[0,1,0]$ from $[2,2,0]$, the pair must be resolved unequally, i.e., $I_B(x_{rc_1c_2}) = 1$. If r contains another 2, say in column c, then by Claim 2, one of the pairs c, c_1 and c, c_2 conflicts in B, which is a contradiction with the fact that K is a connected component of $G_B = G$. Hence, assume that r contains only two 2's. By Observation 4, $I_B(x_{rc_1c_2}) = 0$, i.e., the sequence $[0,1,0]$ cannot be inferred.

B	c_1	c_2	c_3	c_4
	0	0	0	0
$*$	0	0	0	1
	0	0	1	0
$**$	0	0	1	1
	0	1	0	0
$**$	0	1	1	0
$*$	1	0	0	0
?	1	0	0	1
$**$	1	1	0	0

Figure 3: All possible sequence (except for the one with the question mark) $B[K]$ can contain where a component K of G_B is a path (c_1, c_2, c_3, c_4). The rows with star(s) are necessarily contained in $B[K]$.

Finally, assume that K is a path of length 3, say (c_1, c_2, c_3, c_4). Consider any B defined by any covering G from $\bar{H}_A$. By Observation 3, the pairs c_1, c_2, c_2, c_3 and c_3, c_4 are active in A, hence the pairs c_1, c_3 and c_2, c_4 are weakly active. Since, the pairs c_1, c_3 and c_2, c_4 do not conflict in B, $B[K]$ can contain only following 9 quadruples: $[0,0,0,0]$, $[0,0,0,1]$, $[0,0,1,0]$, $[0,0,1,1]$, $[0,1,0,0]$, $[0,1,1,0]$, $[1,0,0,0]$, $[1,0,0,1]$ and $[1,1,0,0]$, cf. Figure 3. Since, the pairs c_1, c_2, c_2, c_3 and c_3, c_4 conflict in B, the rows with two stars "$**$" are necessarily in $B[K]$. Consequently, $B[c_1, c_4]$ contains pairs $[0,1]$ and $[1,0]$ and since c_1, c_4 do not conflict, the row with the question mark "?" cannot appear in $B[K]$. Without a row with "?" in $B[K]$, the rows with one star "$*$" become necessary to guarantee the conflicts between c_1, c_2 and between c_3, c_4. Hence, all rows with star(s) are necessarily in $B[K]$ and rows with no symbol are possibly in $B[K]$. The only candidate for x is 0110. It is easy to check that the condition (2) of Theorem 1 is satisfied if and only if $B[K]$ does not contain any of $[0,0,1,0]$ and $[0,1,0,0]$. Assume now that $B[K]$ contains $y = [0,0,1,0]$. By Definition 13, the extended hypergraph $\bar{H}_A$ does not contain the switch (c_2, c_3, c_4), and hence c_2, c_3, c_4 do not induce $[0,1,0]$ in A. As in the previous case, there are four possible sequences in $A[c_2, c_3, c_4]$ which could be inferred to produce $[0,1,0]$ in $B[c_2, c_3, c_4]$, and building on that 8 possible sequences in $A[K]$ which could be inferred to produce y: (a) $[0,2,2,2]$, (b) $[2,2,2,2]$, (c) $[0,2,1,2]$, (d) $[2,2,1,2]$, (e) $[0,2,2,0]$, (f) $[2,2,2,0]$, (g) $[0,0,2,2]$ and (h) $[2,0,2,2]$. Let us analyze these possibilities:

(a) The only way how to infer y from $[0,2,2,2]$ is to resolve c_2, c_4 equally. This would produce also quadruple $[0,1,0,1]$ in B, and hence a conflict between c_2 and c_4. Which contradicts the fact that K is a component.

(b) To infer sequences from $[2,2,2,2]$ avoiding any conflict between c_1, c_3, c_2, c_4 or c_1, c_4, c_1, c_2 and c_3, c_4 are resolved equally, and other pairs unequally. Hence, the sequence y is not produced.

(c) To infer sequences from $[0,2,1,2]$ avoiding a conflict between c_2, c_4, the pair is resolved unequally, i.e., y is not produced.

(d) The sequences $[2,2,1,2]$ cannot appear in $A[K]$ as otherwise the pair $[1,1]$ is induced by c_1, c_3 in A, i.e., c_1, c_3 would conflict.

(e) Let r be the row containing $[0,2,2,0]$ in A. To produce y, the pair c_2, c_3 has to be resolved unequally. If r contains another 2, say in column c, the one of the pairs c, c_2 and c, c_3 would have to be resolved equally, hence producing a conflict. This would contradict the fact that K is a component in G_B. On the other hand, if r contains only two 2's, by Observation 4, the pair c_2, c_3 is resolved equally in row r.

(f) Let r be a row containing $[2,2,2,0]$ in $A[K]$. First, note that r does not contain any other 2, say in column c. For otherwise there would be a 4-edge in $\bar{H}_A$ containing c_1, c_2, c_3 and c, which would introduce an edge between c and one of c_1, c_2, c_3 in G, a contradiction. There are two ways how to infer sequences from r avoiding conflicts not in K: (1) resolving c_1, c_2 equally; (2) resolving c_2, c_3 equally. In the case (2), y is not produced from $[2,2,2,0]$. In the case (1), we will consider another covering from $\bar{H}_A$, which differs from the current one by choosing edge (c_2, c_3) instead of (c_1, c_2) when processing a row containing $[2,2,2,0]$ in $A[K]$. This new covering might be not canonical. Indeed, it is not canonical if r was the only row containing 2's in c_1, c_2 and at least one other 2 such that $I_B(x_{rc_1c_2}) = 0$ and there is no forced 2-edge (c_1, c_2). In such a case, to make the new covering canonical, we also change the inferring of every row containing only two 2's and those in columns c_1 and c_2 from equal to unequal. Obviously, this new covering will not introduce any new conflict, although it will remove conflict between c_1 and c_2. In such a case, we have found another graph G' that canonically covers $\bar{H}_A$ with smaller number of edges, a contradiction with assumption that G had the minimum number of edges. Hence, we can assume that $G' = G$. If there is another row containing $[2,2,2,0]$ in $A[K]$, we repeat the whole process. Finally, we either get a contradiction or a canonical covering G of $\bar{H}_A$ such that y is not inferred from any row containing $[2,2,2,0]$ in $A[K]$.

(g) Similar to the case (e).

(h) To infer sequences from $[2,0,2,2]$ avoiding any conflict between c_1, c_3 or c_1, c_4, c_3, c_4 is resolved equally, and other pairs unequally. Hence, the sequence y is not produced.

Hence, we can assume that there is a canonical covering G of $\bar{H}_A$ such that the matrix B defined by this covering does not contain $[0,0,1,0]$ in $B[K]$. The similar proof applies to the sequence $[0,1,0,0]$. Hence, there is a haplotype matrix B inferred from A with conflict graph G which satisfied condition (2) of Theorem 1. □

Now, assume that it is possible to find a graph canonically covers $\bar{H}_A$. Let G be such a graph with the minimum number of edges. Then by Claim 7, there is a matrix B inferred from A which can be explained by a galled-tree network, i.e., A can be explained as well. □

3.5 The extended hypergraph covering problem is intractable

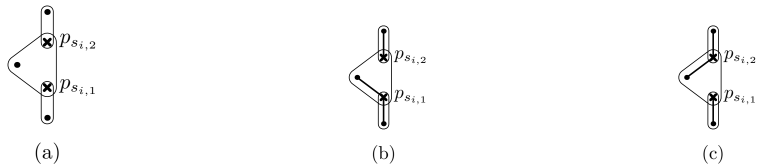
Unfortunately, the EHC problem is intractable as proved in the following theorem, and cannot be used to polynomially solve the weak diagonal instance of the GTNH problem.

Theorem 6. *The EHC problem is NP-hard.*

Proof. The proof is done by conversion from a special instance of 3-SAT problem. This problem is known to be NP-complete even when restricted to formulas where each clause contains 2 or 3 literals and every variable occurs in exactly 3 clauses — once positive and twice negated [13]. Let $f(x_1, x_2, \dots, x_m) = C_1 \wedge \dots \wedge C_k$ be such a formula in conjunctive normal form, where $C_1, \dots, C_k$ are clauses. Let $p_1, \dots, p_m$ be all occurrences of literals $\{x_1, \dots, x_m, \neg x_1, \dots, \neg x_m\}$ in f. For every $i = 1, \dots, k$, we have

$$C_i = p_{s_{i,1}} \vee p_{s_{i,2}} \qquad \text{or} \qquad C_i = p_{s_{i,1}} \vee p_{s_{i,2}} \vee p_{s_{i,3}}$$

depending on whether C_i contains 2 or 3 literals. Let $S_i = \{s_{i,1}, s_{i_2}\}$ or $S_i = \{s_{i,1}, s_{i,2}, s_{i,3}\}$, respectively. Sets $S_1, \dots, S_k$ forms a decomposition of set $\{1, \dots, m\}$. For every $i = 1, \dots, m$, let $p_{t_{i,1}}$ be the occurrence of positive literal x_i and $p_{t_{i,2}}, p_{t_{i,3}}$ two occurrences of negated literal $\neg x_i$. Let $T_i = \{t_{i,1}, t_{i,2}, t_{i,3}\}$, Again, sets $T_1, \dots, T_m$ form a decomposition of set $\{1, \dots, m\}$. We will construct an extended hypergraph $\bar{H}(f)$ such that it has a canonical covering with all components being paths of length at most 3 and satisfying the ordered-component property if and only if f is satisfiable. The hypergraph will only contain forced 2-edges and 3-edges. Hence, each covering is canonical, and we will not mention this property of the covering in the remaining of the proof. We will form the hypergraph $\bar{H}(f)$ as a union of several hypergraphs, called gadgets, one for each clause, and one for each variable. The numbering of vertices in the hypergraph is important as it influence the ordered-component property. All constructed gadgets will have to types of vertices (characters): depicted by a dot and depicted by a cross. For our construction it will be sufficient to number vertices so that all vertices with a cross have higher number than vertices with a dot, which can be easily achieved.



(a) (b) (c)

Figure 4: Part of hypergraph $\bar{H}(f)$ for clause $C_i = p_{s_{i,1}} \vee p_{s_{i,2}}$ and all possible graphs cover it, each representing one case how the clause can become satisfied. The vertices with cross have higher numbers than vertices with dots.

For every clause with two literals $C_i = p_{s_{i,1}} \vee p_{s_{i,2}}$, we construct a part of hypergraph (a gadget) consisting of two forced 2-edges and one 3-edge as depicted in Figure 4. The figure also shows all possible graphs cover the gadget satisfying the conditions of the EHC problem. In these figures, a variable p_j has value 1 if no other edge (from the other parts of $\bar{H}(f)$) can be joining the vertex p_j. For instance, in the first graph, $p_{s_{i,1}}$ has value 1, as any other edge joining $p_{s_{i,1}}$ would increase the degree of this vertex to 3, and $p_{s_{i,2}}$ has value 0. Note that in each hypergraph covering of the gadget at least one of $p_{s_{i,1}}$ and $p_{s_{i,2}}$ has value 1.

For every clause with three literals $C_i = p_{s_{i,1}} \vee p_{s_{i,2}} \vee p_{s_{i,3}}$, we construct a gadget consisting of four forced 2-edged and four 3-edges as depicted in Figure 5(a). Figure 5 also shows three possible graphs (b–d) that cover the part of hypergraph satisfying the conditions of the EHC problem. There are other graphs which can be covering for the part of $\bar{H}(f)$, however in each of them at least one of $p_{s_{i,1}}, p_{s_{i,2}}, p_{s_{i,3}}$ has value 1. Indeed, assume that all three values are set to 0. Then no edge from inside of gadget can be joining any of $p_{s_{i,1}}, p_{s_{i,2}}, p_{s_{i,3}}$. We have the situation depicted in Figure 5(e). Now, there is no edge to be selected from the 3-edge which vertices are connected with dashed edges without increasing degree of some vertex to 3.

The second part of the construction checks whether three occurrences of a variable x_i: $p_{t_{i,1}}, p_{t_{i,2}}, p_{t_{i,3}}$ do not have contradictory values. That is if $p_{t_{i,1}}$ (positive occurrence) has value 1 then both $p_{t_{i,2}}$ and $p_{t_{i,3}}$ (negated occurrences) should have values 0, and if at least one of $p_{t_{i,2}}$ and $p_{t_{i,3}}$ has value

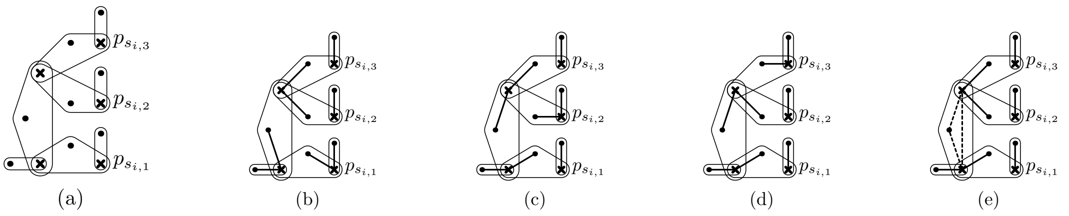


Figure 5: (a) The part of hypergraph $\bar{H}(f)$ for clause $C_i = p_{s_{i,1}} \vee p_{s_{i,2}} \vee p_{s_{i,3}}$. (b–d) Three possible graphs that cover from it, each representing one case how the clause can become satisfied. (e) A trial to search for a hypergraph covering with values of $p_{s_{i,1}}, p_{s_{i,2}}, p_{s_{i,3}}$ set to 0.

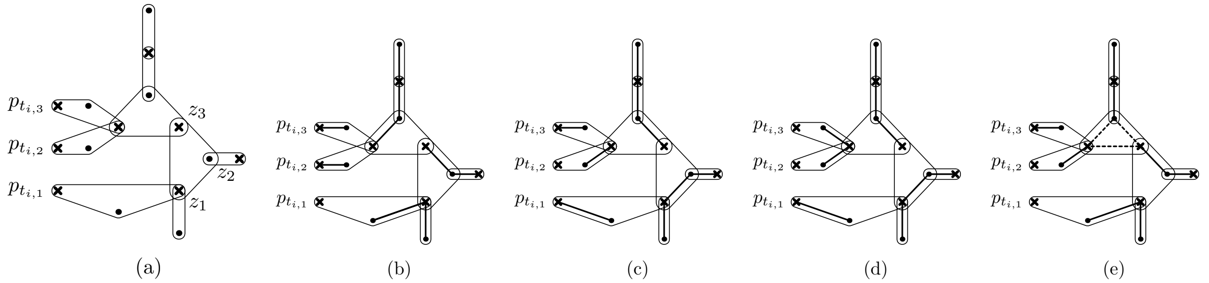


Figure 6: (a) The part of hypergraph $\bar{H}(f)$ verifying the values of three occurrences of a variable x_i. (b-d) Three possible hypergraph coverings. In (b), $p_{t_{i,1}}$ has value 1 and forces values of $p_{t_{i,2}}$ and $p_{t_{i,3}}$ to 0. In (c), $p_{t_{i,2}}$ has value 1 and in (d), both $p_{t_{i,2}}$ and $p_{t_{i,3}}$ have value 1. In both cases (c) and (d), $p_{t_{i,1}}$ is forced to have value 0. (e) A trial to search for a hypergraph covering with values of $p_{t_{i,1}}$ and $p_{t_{i,2}}$ set to 1.

1 then $p_{t_{i,1}}$ should have value 0. This is achieved by a gadget consisting of four forced 2-edges and four 3-edges depicted in Figure 6(a). Figures 6(b–d) show three possible graphs that cover the gadget. In these figures, a variable p_j has value 1 if no edge in the gadget joins p_j, which is in agreement with interpretation of values of p_i's in gadgets of the first part of construction.

Let us verify the claimed property of the gadget. Assume for instance that both $p_{t_{i,1}}$ and $p_{t_{i,2}}$ have values 1. Hence, no edge from inside of the gadget is joining these two vertices. Then to avoid a vertex of degree higher than 2, from the 3-edge $\{z_1, z_2, z_3\}$ we have to select edge (z_2, z_3), cf. Figure 6(e). Now, there is no edge to be selected from the 3-edge which vertices are connected with dashed edges without producing a path of length 4 or 5. The other cases can be proved using similar arguments.

Now, let us verify that it is possible to find a hypergraph covering of $\bar{H}(f)$ which satisfies conditions of the EHC problem if and only if f is satisfiable. First, consider a graph G that covers from $\bar{H}(f)$ such that each component is a path of length at most 3. For every clause C_i, at least one of $p_{s_{i,1}}, p_{s_{i,2}}$ (respectively, $p_{s_{i,1}}, p_{s_{i,2}}, p_{s_{i,3}}$) has value 1 in G. Let it be p_{q_i} (if there are several literals in C_i with value 1 in G, pick any of them). We will form a true assignment as follows. For every x_j, if there is $p_{q_i} = x_j$, set $x_j = 1$; if there is $p_{q_i} = \neg x_j$, set $x_j = 0$; otherwise set x_j to any value. As long as we guarantee that there are no i, i' such that $p_{q_i} = x_j$ and $p_{q_{i'}} = \neg x_j$, the above definition is correct and obviously is a true assignment for f. Assume for contrary that $p_{q_i} = x_j$ and $p_{q_{i'}} = \neg x_j$. Obviously, $q_i = t_{j,1}$ and $q_{i'}$ is either $t_{j,2}$ or $t_{j,3}$. Now, since we $p_{t_{j,1}}$ has value 1 and one of $p_{t_{j,2}}, p_{t_{j,3}}$ has value 1, we have a contradiction with the property of the gadget for x_j.

For the converse, consider a true assignment for f. For every clause $C_i = p_{s_{i,1}} \vee p_{s_{i,2}}$ with two literals, there is at least one literal p_{q_i} with value 1, where $q_i \in \{s_{i,1}, s_{i,2}\}$. If $q_i = s_{i,1}$, search for a hypergraph covering of the gadget for C_i as depicted in Figure 4(b). If $q_i = s_{i,2}$, as in Figure 4(c). Similarly, for every clause $C_i = p_{s_{i,1}} \vee p_{s_{i,2}} \vee p_{s_{i,3}}$ with three literals, there is at least one literal p_{q_i} with value 1, where $q_i \in \{p_{s_{i,1}}, p_{s_{i,2}}, p_{s_{i,3}}\}$. If $q_i = s_{i,1}$ (respectively, $q_i = s_{i,2}$, $q_i = s_{i,3}$), search for a hypergraph covering of the gadget for C_i as depicted in Figure 5(b) (respectively, Figure 5(c), Figure 5(d)). For the gadgets in the second part of construction we will search for coverings as follows. For every x_i, if value of x_i is 1, find a hypergraph covering of the gadget for x_i as depicted in Figure 6(b), and if value of x_i is 0, as in Figure 6(d). Let G be the union of graphs that cover all gadgets. We will show that G satisfies the conditions of the EHC problem.

First, it is easy to see that all possible edges of G are connecting a vertex with a cross with a vertex with a dot. The ordered-component property follows. It is also easy to see that the components of graphs that cover a single gadget are all paths of length at most 3. Hence, it is enough to verify that p_k-vertices which are shared between gadgets do not combine components to a component which is not a path of length at most 3. Observe that each p_k is shared by exactly two gadgets, one gadget for a clause C_i and one gadget for a variable x_j. In the gadget for C_i, the component containing p_k is either an edge, if p_k is not necessary to satisfy the clause, or a path of length with p_k as the middle vertex, if p_k is necessary to satisfy the clause. In the second case, by definition of G, p_k has value 1 in the considered true assignment for f. In the gadget for x_j, the component containing p_k is either a singleton, if p_k has value 1, and an edge, if p_k has value 0. The only way how to obtain an invalid component is to combine the middle vertex of path of length 2 with an edge, but this will never happen as the first one requires the value of p_k to be 1 and the other to be 0. □

Even though we have proved that the EHC problem is NP-complete, it does not imply the NP-completeness of the GTNH problem. Indeed, it is not possible to construct a genotype matrix resulting in the constructed gadgets for a given boolean formula. However, our recent research indicates that the weak diagonal instance of the GTNH problem might be the key for proving NP-completeness of the GTNH problem in general.

References

[1] V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. *Journal of Computational Biology*, 10(3-4):323–340, 2003.

[2] A. Clark. Inference of haplotypes from PCR-amplified samples of dipoid populations. *Molecular Biology and Evolution*, 7:111–122, 1990.

[3] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–232, 2001.

[4] S. Gabirel, S. Schaffner, H. Nguyen, J. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. Lander, M. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296, 2002.

[5] A. Gupta, J. Maňuch, L. Stacho, and X. Zhao. Algorithm for haplotype inferring via galled-tree networks with simple galls. (submitted).

[6] A. Gupta, J. Maňuch, L. Stacho, and X. Zhao. Characterization of the existence of galled-tree networks. In *Proceedings of the 4th Asia-Pacific Bioinformatics Conference*, pages 297–306, 2006.

[7] D. Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 166–175, New York, NY, USA, 2002. ACM Press.

[8] D. Gusfield. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *J. Comput. Syst. Sci.*, 70(3):381–398, 2005.

[9] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. of Bioinformatics and Computational Biology, Special issue on the 2003 IEEE CSB Bioinformatics Conference*, 2(1):173–213, 2004.

[10] L. Helmuth. Genome research: Map of the human genome 3.0. *Science*, 293(5530):583–585, 2001.

[11] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[12] R. D. Mitra, V. L. Butty, J. Shendure, B. R. Williams, D. E. Housman, and G. M. Church. Digital genotyping and haplotyping with polymerase colonies. In *Proceedings of the Nationlal Academy of Sciences of the United States of America*, volume 100, pages 5926–5931, 2003.

[13] C. H. Papadimitriou. *Computational Complexity.* Addison-Wisley Publishing Company, Inc., 1994.

[14] N. Patil, A. Berno, D. Hinds, W. Barrett, J. Doshi, C. Hacker, C. Kautzer, D. Lee, C. Marjoribanks, D. McDonough, B. Nguyen, M. Norris, J. Sheehan, N. Shen, D. Stern, R. Stokowski, D. Thomas, M. Trulson, K. Vyas, K. Frazer, S. Fodor, and D. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, 2001.

[15] Y. S. Song, Y. Wu, and D. Gusfield. Algorithms for imperfect phylogeny haplotyping (IPPH) with a single homoplasy or recombination event. In R. Casadio and G. Myers, editors, *WABI*, volume 3692 of *Lecture Notes in Computer Science*, pages 152–164. Springer, 2005.

[16] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8(1):69–78, 2001.