A Concise Necessary and Sufficient Condition for the Existence of a Galled-Tree

Yun S. Song

Abstract— Galled-trees are a special class of graphical representation of evolutionary history that has proved amenable to efficient, polynomial-time algorithms. The goal of this paper is to construct a concise necessary and sufficient condition for the existence of a galled-tree for M, a set of binary sequences that purportedly have evolved in the presence of recombination. Both root-known and root-unknown cases are considered here.

Index Terms—Galled-trees, recombination, quadpartition, incompatibility

1 Introduction

MEIOTIC recombination is a biological process that produces a new sequence as a mosaic of two sequences of equal length. Two possible outcomes of this process are illustrated in Figure 1, where x and y denote "parental" sequences of equal length, say m, while z denotes a "recombinant" sequence, also of length m. For each $1 \leq i \leq m$, the ith character in z comes from the ith character in either x or y. The switching of the parent in the recombinant sequence z is called *crossing over*. Figure 1a depicts a case of multiple crossovers, whereas Figure 1b illustrates a case with a single crossover. So far, most mathematical work on recombination has focused on single-crossover meiotic recombination, and our present work, too, is on that particular topic. In what follows, when we say recombination without any qualification, we mean single-crossover meiotic recombination.

A common assumption, which we adopt in our work as well, is that point (or site) mutations are governed by the infinite-sites model, which permits at most one mutation per site. This assumption implies that the sequences one considers are binary, i.e., have at most two distinct states at each site; SNP (single nucleotide polymorphism) sequences satisfy this condition and are now of considerable interest (e.g., see [9]). In the presence of recombination, the evolution of sampled sequences is represented not by a tree but by a directed graph, called an ancestral recombination graph (ARG) in the population genetics literature [2]. See [3], [5] for a formal definition of an ARG. For a given set M of binary sequences, constructing an ARG (allowing only one mutation per site) with the minimum number of recombinations is a well-known problem, first posed by Hein [7], [8] about 15 years ago. No general efficient algorithm for solving that problem is known, and the case with a known root sequence has been shown to be NP-hard [13]. As a way of simplifying the problem, Wang *et al.* [13] proposed studying a special class of ARGs with constrained structures; these constrained ARGs later became to be called "galled-trees" [5].

An accessible introduction to galled-trees is provided in [3]. See Figure 2 for an example of a galled-tree. A *recombination cycle* is an undirected cycle in an ARG consisting of two directed paths originating at a "coalescent" vertex and terminating at a "recombination" vertex. A *gall* is a recombination cycle in an ARG that does not share a vertex with any other recombination cycle. Finally, a *galled-tree* is an ARG in which every recombination cycle is a gall. If there exists a galled-tree for M, then the minimum number $R_{\min}(M)$ of recombination events needed in any ARG for M is equal to the number of non-trivial connected components in the associated incompatibility graph $G(M)$, which we define in the next section. (See [3], [6] for details. A related work is [1].)

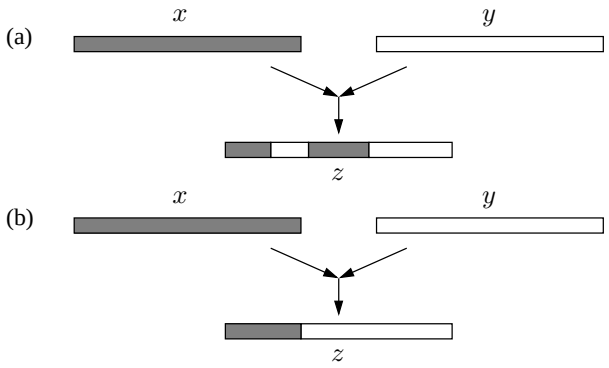

Fig. 1. Possible outcomes of meiotic recombination. The ith character in z comes from the ith character in either x or y. (a) Multiple crossovers. (b) A single crossover.

Wang *et al.* [13] assumed that the ancestral sequence at the root is known and suggested an inaccurate polynomial-time algorithm for determining whether a given set M of sequences can be derived on a galled-tree. This line of work was subsequently made more complete by Gusfield *et al.* [5], who provided a correct polynomial-time algorithm. Furthermore, Gusfield [3] later generalized the results to the root-unknown case and showed that it also admits a polynomial-time algorithm. Other works on constructing galled-trees include [10], [11], which take trees as input.

Galled-trees possess numerous combinatorial constraints, and many necessary conditions for the existence of a galled-tree are known [3], [4], [5]. The algorithms in [3], [5] provide an efficient test for the existence of a galled-tree, but no concise *necessary and sufficient* condition is currently known. The goal of this paper is to devise a concise necessary and

Y.S. is with the Department of Computer Science, University of California at Davis, Davis, CA 95616, USA. E-mail: yssong@cs.ucdavis.edu

$$M = \begin{bmatrix} 0001101 \\ 0001111 \\ 1001111 \\ 1011111 \\ 1011101 \\ 1111111 \\ 1110111 \\ 1110010 \\ 1011010 \\ 1011110 \end{bmatrix}$$

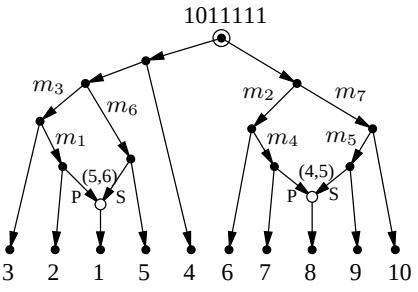

Fig. 2. A galled-tree with two galls. A mutation event at site i occurs on the edge labeled m_i. The root sequence for this galled-tree is 1011111. Recombination vertices are denoted by open circles, and the notation (i, j) indicates that a crossover occurs between sites i and j, with the incoming edge labeled "P" ("S") contributing a prefix (suffix).

sufficient condition (NASC) for the existence of a galled-tree for M. We consider both root-known and root-unknown cases. We draw a link to an earlier work [12], based on studying a system of combinatorial constraints induced by pairs of columns in M, and formulate our NASC in that framework.

This paper is organized as follows. In Section 2, we provide some necessary definitions and describe our notational convention. The notion of *quadpartition*, the key to our approach, is described there. In Section 3, we construct a NASC for the existence of a galled tree for M with a single non-trivial connected component in $G(M)$. The results from that section are then used in Section 4, where we describe a NASC for the existence of a galled-tree for M with multiple non-trivial connected components in $G(M)$.

2 PRELIMINARIES

We use M to denote a set of n binary sequences, each of length m. In what follows, we view M as an n by m 0-1 matrix. The ith column is sometimes called the ith *site*, and we use these terms interchangeably. Greek indices $\alpha, \beta, \gamma, etc.$ denote row labels, while Roman indices $i, j, k, etc.$ denote column labels. We use $M[\alpha, i]$ to denote the (α, i) entry in M and $M[\alpha, (i, j)]$ to denote the ordered pair $(M[\alpha, i], M[\alpha, j])$. If $M[\alpha, (i, j)] = (a, b)$, we simply write $M[\alpha, (i, j)] = ab$. Given a row α in M, we use $M - \alpha$ to denote an $(n-1)$ by m matrix obtained by removing row α from M. Similarly, we use $M + \rho$ to denote an $(n+1)$ by m matrix created from M by adding an additional row ρ, which may or may not already be present in M.

Let $X(M)$ be the set of row labels in M. The row label set for $M + \rho$ is $X(M) \cup \{\rho\}$. To each column i in M, there corresponds a bipartition $A_i|B_i$, also known as a split, of $X(M)$ into disjoint subsets A_i and B_i, such that α and β belong to the same subset if and only if $M[\alpha, i] = M[\beta, i]$. For example, if $n = 5$ and the ith column of M is $(0, 1, 0, 0, 1)$, then the corresponding bipartition is $\{1, 3, 4\}|\{2, 5\}$.

Central to our work are the notions of incompatibility and of a quadpartition, which we now define.

Definition 1 (Incompatibility & Quadpartition). *A pair of sites i, j is called compatible if at least one of the following intersections is empty:*

$$A_i \cap A_j \ , \ A_i \cap B_j \ , \ B_i \cap A_j \ , \ B_i \cap B_j \ ,$$

$$M = \begin{bmatrix} 1010 \\ 1001 \\ 0101 \\ 0100 \\ 0010 \end{bmatrix}$$

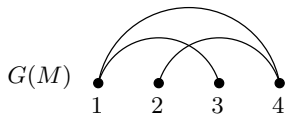

Fig. 3. The data set M and its incompatibility graph $G(M)$ considered in Example 1.

where $A_i|B_i$ and $A_j|B_j$ are bipartitions of $X(M)$ corresponding to sites i and j, respectively. If none of the above intersections is empty, then the pair i, j is called incompatible, and we say that i conflicts with j, and vice versa. A site not in conflict with any other site is said to be conflict-free. To each incompatible pair i, j, there corresponds a quadpartition

$$\begin{array}{c|c} A_i \cap A_j & A_i \cap B_j \\ \hline B_i \cap A_j & B_i \cap B_j \end{array} \tag{1}$$

of $X(M)$ into four pair-wise disjoint proper subsets. We denote this quadpartition by $\pi^{ij} := \{Q^{ij}(00), Q^{ij}(01), Q^{ij}(10), Q^{ij}(11)\}$, where $Q^{ij}(ab)$ contains $\alpha \in X(M)$ if and only if $M[\alpha, (i, j)] = ab$.

One way to test whether two sites i, j are incompatible is to check whether all state pairs $00, 01, 10, 11$ appear in the two columns. This test is called the *four-gamete* test in the population genetics literature.

Example 1. Consider the 5 by 4 matrix M shown on the left hand side of Figure 3. The set of all incompatible pairs of sites in M is $\{(1, 3), (1, 4), (2, 4)\}$, and the corresponding quadpartitions are as follows:

(i, j)	$Q^{ij}(00)$	$Q^{ij}(01)$	$Q^{ij}(10)$	$Q^{ij}(11)$
$(1, 3)$	$\{3, 4\}$	$\{5\}$	$\{2\}$	$\{1\}$
$(1, 4)$	$\{4, 5\}$	$\{3\}$	$\{1\}$	$\{2\}$
$(2, 4)$	$\{1, 5\}$	$\{2\}$	$\{4\}$	$\{3\}$

Definition 2 (Incompatibility Graph). *Given an n by m binary matrix M, the incompatibility graph $G(M)$ of M is a graph with m vertices bijectively labeled by the sites in M, such that vertices i and j in $G(M)$ are joined by an edge if and only if sites i and j in M are incompatible. A connected component (CC) of $G(M)$ is called non-trivial if it contains at least two vertices.*

The incompatibility graph $G(M)$ for the data set considered in Example 1 is shown on the right hand side of Figure 3. It contains a single non-trivial CC.

We use $\mathcal{P}(M)$ to denote the set of all pairs (i, j), $i < j$, of incompatible sites in M. Similarly, $\mathcal{P}(C)$ denotes the set of all pairs (i, j), $i < j$, of incompatible sites in a non-trivial connected component C of $G(M)$. To each incompatible pair $(i, j) \in \mathcal{P}(M)$, we assign an *open* interval of the form $I(i, j) := \{x \in \mathbb{R} \,|\, i < x < j\}$. Given two pairs $(i, j), (k, l) \in \mathcal{P}(M)$, we say that the incompatibility between i and j *overlaps* with that between k and l if $I(i, j) \cap I(k, l) \neq \varnothing$.

3 GALLED-TREES FOR M WITH A SINGLE NON-TRIVIAL CC IN $G(M)$

In this section, we construct a concise necessary and sufficient condition for the existence of a galled-tree for M with exactly one non-trivial CC in $G(M)$. Results from this section will be used in the next section, where we consider galled-trees for M with multiple non-trivial CCs in $G(M)$.

3.1 Root-unknown case

We first examine the root-unknown case. The main result we establish is

Theorem 1. *Suppose that $\mathcal{P}(M) \neq \varnothing$. Then, there exists a galled-tree with exactly one gall that derives M if and only if M satisfies the following conditions:*

C-1. $\bigcap_{(i,j)\in\mathcal{P}(M)} I(i,j) \neq \varnothing$, *i.e., every incompatibility overlaps with all other incompatibilities, and*

C-2. *there exists a non-empty $\mathcal{Q} \in \bigcap_{(i,j)\in\mathcal{P}(M)} \pi^{ij}$. We call this $\mathcal{Q}$ a common Q-set.*

Remarks:

1) In Example 1, the only possible common Q-set is $\mathcal{Q} = \{2\}$.
2) It follows from Lemma 5 of [12] that if two incompatible pairs (i,j) and (k,l) belong to two different non-trivial CCs of $G(M)$, then there does not exist a common Q-set for them. Therefore, condition C-2 implies that $G(M)$ contains a single non-trivial CC.
3) In general, $\mathcal{Q}$ may not be unique.

We first establish some lemmas needed for proving the main theorem. Let M' denote the matrix obtained after removing all conflict-free columns from M and then removing all but one copy of identical rows. Since conflict-free columns do not influence the incompatibility pattern and identical rows always appear together in the same $Q^{ij}(kl) \in \pi^{ij}$ for all $(i,j) \in \mathcal{P}(M)$, if M satisfies C-1 and C-2, then so does the new matrix M'. In what follows, suppose that M' satisfies C-1 and C-2, and let $\mathcal{Q}'$ denote the corresponding common Q-set. Because M' contains distinct rows and every column in M' is in conflict with at least one other column in M', it follows that $\mathcal{Q}'$ contains only a single row label, which we denote by σ. In what follows, the following fact is used repeatedly:

Fact 1. *If σ is the all-zero sequence $00\ldots0$, and columns i and j are incompatible, then σ is the only row with 0s at both i and j.*

Since M' contains no conflict-free columns, if M' has m' columns, then C-1 implies that there exists a k such that

(a) columns $1, 2, \ldots, k$ in M' are pair-wise compatible,
(b) columns $k+1, k+2, \ldots, m'$ in M' are pair-wise compatible, and
(c) every column $i \in \{1, 2, \ldots, k\}$ is incompatible with at least one column in $\{k+1, \ldots, m'\}$, and vice versa.

Let $X(M')$ denote the set of rows in M'. Every column in M' defines a partition of $X(M')$ into two disjoint non-empty subsets. In the bipartition corresponding to column i, let B^i (respectively, A^i) denote the subset that contains (respectively, does not contain) σ. Then, we have the following lemma for compatible sites in M':

Lemma 1. *Suppose that M' satisfies conditions C-1 and C-2 in Theorem 1, and let k, m' be as described above. If $i_1, i_2 \in \{1, 2, \ldots, k\}$, then either*

$$B^{i_1} \subseteq B^{i_2} \text{ or } B^{i_2} \subseteq B^{i_1}.$$

Similarly, if $j_1, j_2 \in \{k+1, k+2, \ldots, m'\}$, then either

$$B^{j_1} \subseteq B^{j_2} \text{ or } B^{j_2} \subseteq B^{j_1}.$$

Proof. If $B^{i_1} = B^{i_2}$ the lemma holds trivially, so suppose $B^{i_1} \neq B^{i_2}$. Without loss of generality, let σ be the all-zero sequence. According to this convention, all rows in B^i have 0 at column i, whereas all rows in A^i have 1 at column i.

Note that $B^{i_1} \cap B^{i_2} \neq \varnothing$, since by definition both B^{i_1} and B^{i_2} contain σ. Now, suppose that $B^{i_1} \not\subset B^{i_2}$ and $B^{i_2} \not\subset B^{i_1}$. Then, $B^{i_1} \cap A^{i_2} \neq \varnothing$ and $B^{i_2} \cap A^{i_1} \neq \varnothing$. Therefore, since i_1 and i_2 are compatible, $A^{i_1} \cap A^{i_2} = \varnothing$, which implies $A^{i_1} \subsetneq B^{i_2}$ and $A^{i_2} \subsetneq B^{i_1}$. Note that $A^{i_2} \neq \varnothing$, by property (c). As $A^{i_2} \subset B^{i_1}$, there exists a row α_1 such that $M'[\alpha_1, i_1] = 0$ and $M'[\alpha_1, i_2] = 1$. Also, by property (c), there exists $c_1 \in \{k+1, \ldots, m'\}$ such that i_1 and c_1 are incompatible. Hence, by property (c) there exist sequences α_2 and α_3 satisfying $M'[\alpha_2, (i_1, c_1)] = 11$ and $M'[\alpha_3, (i_1, c_1)] = 10$, respectively. What we just described is summarized in Figure 4a.

Now, since $\alpha_2, \alpha_3 \in A^{i_1} \subset B^{i_2}$, we conclude that $M'[\alpha_2, i_2] = 0$ and $M'[\alpha_3, i_2] = 0$. Furthermore, since i_1 and c_1 are incompatible, by Fact 1 σ is the only sequence with 0s at both i_1 and c_1, so we must have $M'[\alpha_1, c_1] = 1$. So far, we have established Figure 4b. As $M'[\alpha_3, (i_2, c_1)] = 00$, Fact 1 implies that i_2 and c_1 must be compatible. Hence, by property (c) there must exist another column, say c_2, in $\{k+1, \ldots, m'\}$ that is incompatible with i_2. Now, since $M'[\alpha_2, i_2] = 0$ and $M'[\alpha_3, i_2] = 0$, and i_2 and c_2 are incompatible, by Fact 1 $M'[\alpha_2, c_2] = 1$ and $M'[\alpha_3, c_2] = 1$. In summary, we have a situation as shown in Figure 4c.

Since c_1 and c_2 are compatible (by property (b)), we must have $M'[\alpha_1, c_2] = 1$. Let α_5 be a row with $M'[\alpha_5, (i_2, c_2)] = 10$. Such a sequence exists since i_2 and c_2 are incompatible, and 10 is not the state pair for any of the rows seen so far. This leads to Figure 4d. Now, since c_1 and c_2 are compatible, $M'[\alpha_5, c_1] = 0$. Lastly, since i_1 and c_1 are incompatible, by Fact 1 α_5 cannot be 0 at column i_1. This yields the situation shown in Figure 4e. But now, all 4 gametic types—namely $00, 01, 10, 11$—are present in columns i_1 and i_2. This contradicts the fact that i_1 and i_2 are compatible, and hence $B^{i_1} \subseteq B^{i_2}$ or $B^{i_2} \subseteq B^{i_1}$. The second part of the lemma follows from the same line of reasoning. ∎

Using Lemma 1, we can now obtain the following lemma:

Lemma 2. *Suppose that M' satisfies conditions C-1 and C-2 in Theorem 1, with the common Q-set $\mathcal{Q}' = \{\sigma\}$. Let k and m' be as described above. Then, in $M' - \sigma$, there exists a row that is identical to σ for columns $1, 2, \ldots k$. Similarly,*

(a)

	i_1	i_2	c_1
α_1	0	1	
α_2	1		1
α_3	1		0
σ	0	0	0

(b)

	i_1	i_2	c_1
α_1	0	1	1
α_2	1	0	1
α_3	1	0	0
σ	0	0	0

(c)

	i_1	i_2	c_1	c_2
α_1	0	1	1	
α_2	1	0	1	1
α_3	1	0	0	1
σ	0	0	0	0

(d)

	i_1	i_2	c_1	c_2
α_1	0	1	1	1
α_2	1	0	1	1
α_3	1	0	0	1
σ	0	0	0	0
α_5		1		0

(e)

	i_1	i_2	c_1	c_2
α_1	0	1	1	1
α_2	1	0	1	1
α_3	1	0	0	1
σ	0	0	0	0
α_5	1	1	0	0

Fig. 4. Sequential information constructed in the proof of Lemma 1. In (e), columns i_1 and i_2 are incompatible, in contradiction with the assumption.

there exists a row in $M' - \sigma$ that is identical to σ for columns $k+1, k+2, \ldots m'$.

Proof. Recall that B^i denotes the subset containing σ in the bipartition corresponding to column i. Let i_M be a site with the minimum value of $|B^i|$. Then, by Lemma 1, $B^{i_M} \subseteq B^i$ for all $i \neq i_M$, so all sequences in B^{i_M} are identical for columns $1, 2, \ldots, k$. Furthermore, since i_M is in conflict with some other column $j \in \{k+1, k+2, \ldots m'\}$, $|B^{i_1}| \geq 2$. It therefore follows that M' contains a row other than σ that is identical to σ for columns $1, 2, \ldots, k$. The second part of the lemma can be shown in a similar vein. ■

We now provide a proof of our main result:

Proof of Theorem 1. If there exists a galled-tree N with a single gall, the set of sequences below the single recombination vertex in N is a common Q-set, and conditions C-1 and C-2 are clearly satisfied. We wish to show that if conditions C-1 and C-2 are satisfied, then there exists a galled-tree with a single gall. Let σ and M' be as described above. Remove row σ from M' and construct a perfect phylogeny for $M' - \sigma$. We know that there exists a perfect phylogeny for $M' - \sigma$ since removing σ from M' eliminates all incompatibilities (recall Fact 1 in the proof of Lemma 1). Further, Lemma 2 implies that a single recombination event is sufficient to derive σ starting from such a perfect phylogeny. Let N' denote this galled-tree for M'. A galled-tree for M, also with exactly one recombination vertex, can be obtained from N' by including mutation events for conflict-free columns and additional leaves for removed rows. ■

Example 2. The matrix M shown on the left hand side of Figure 5 contains 4 pairs of incompatible sites, given by $\mathcal{P}(M) = \{(1,3), (1,4), (2,3), (2,4)\}$. Its associated incompatibility graph $G(M)$ is shown in the middle of Figure 5, and the corresponding quadpartitions are as follows:

(i,j)	$Q^{ij}(00)$	$Q^{ij}(01)$	$Q^{ij}(10)$	$Q^{ij}(11)$
$(1,3)$	$\{2,3\}$	$\{4\}$	$\{1,6\}$	$\{5\}$
$(1,4)$	$\{4\}$	$\{2,3\}$	$\{5,6\}$	$\{1\}$
$(2,3)$	$\{1,2,6\}$	$\{5\}$	$\{3\}$	$\{4\}$
$(2,4)$	$\{5,6\}$	$\{1,2\}$	$\{4\}$	$\{3\}$

Condition C-1 in Theorem 1 is satisfied, and so is condition C-2, with the common Q-set being $\mathcal{Q} = \{4\}$. Therefore, we conclude that there exists a galled-tree for this example. Shown on the right hand side of Figure 5 is a galled-tree for M with

$$M = \begin{bmatrix} 1001 \\ 0001 \\ 0101 \\ 0110 \\ 1010 \\ 1000 \end{bmatrix}$$

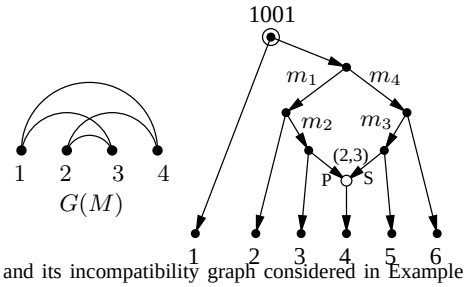

Fig. 5. The matrix and its incompatibility graph considered in Example 2. A galled-tree for the data set with a single gall is shown on the right.

1001 as the root sequence. There are several other possible galled-trees with a single gall that derives M. They all share the property that the sequence below the recombination vertex is 4, the single element in $\mathcal{Q}$.

3.2 Root-known case

Suppose that the ancestral sequence at the root is ρ. Then, the notion of incompatibility and quadpartition defined in Definition 1 can be applied to $M + \rho$. The following result can be obtained in a similar vein as in the root-unknown case:

Theorem 2. *Suppose that $\mathcal{P}(M+\rho) \neq \varnothing$. Then, there exists a galled-tree for M with a single gall and with ρ as the root sequence, if and only if $M + \rho$ satisfies the following conditions:*

D-1. $\bigcap_{(i,j)\in\mathcal{P}(M+\rho)} I(i,j) \neq \varnothing$,
D-2. *there exists a non-empty* $\mathcal{Q} \in \bigcap_{(i,j)\in\mathcal{P}(M+\rho)} \pi^{ij}$, *and*
D-3. $\rho \notin \mathcal{Q}$.

Proof. When there is a galled-tree N with a single gall that derives M, let $\mathcal{Q}$ be the set of sequences below the recombination vertex v. Now, if ρ is the root sequence of N and $\rho \in \mathcal{Q}$, then it implies that no recombination is needed, which is impossible since $\mathcal{P}(M + \rho) \neq \varnothing$. Hence conditions D-1, D-2 and D-3 are easily established. Conversely, if the conditions are satisfied, then we know that there exists a galled tree N with a single gall that derives $M + \rho$ (by Theorem 1) and that ρ is not below the recombination vertex in N (by D-3). A galled-tree for M, with a single gall and with ρ as the root sequence, can be obtained by redirecting some edges in N to make ρ the root sequence and removing the leaf corresponding to ρ. ■

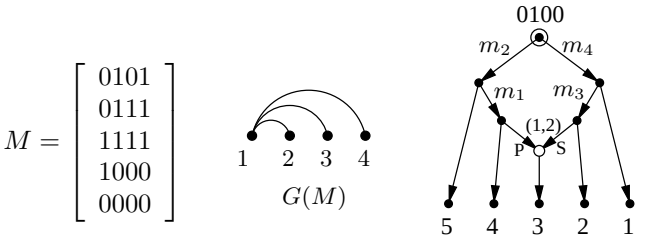

Fig. 6. The matrix M and its incompatibility graph $G(M)$ considered in Example 3. A galled-tree for the data set with a single gall is shown on the right.

Example 3. Consider the matrix M in Figure 6. If the root sequence is $\rho_1 = 0010$, condition D-1 in Theorem 2 is violated and hence there exists no galled-tree for M with ρ_1 as the root sequence. Suppose that the root sequence is $\rho_2 = 1001$. Then, the set of quadpartitions for $M + \rho_2$ is as follows:

(i,j)	$Q^{ij}(00)$	$Q^{ij}(01)$	$Q^{ij}(10)$	$Q^{ij}(11)$
$(1,2)$	$\{5\}$	$\{1,2\}$	$\{4,\rho_2\}$	$\{3\}$
$(1,3)$	$\{1,5\}$	$\{2\}$	$\{4,\rho_2\}$	$\{3\}$
$(1,4)$	$\{5\}$	$\{1,2\}$	$\{4\}$	$\{3,\rho_2\}$

This shows that condition D-2 is violated, and therefore we conclude that there exists no galled-tree for M with ρ_2 as the root sequence. Suppose that the root sequence is $\rho_3 = 0100$, in which case $M + \rho_3$ has the following quadpartitions:

(i,j)	$Q^{ij}(00)$	$Q^{ij}(01)$	$Q^{ij}(10)$	$Q^{ij}(11)$
$(1,2)$	$\{5\}$	$\{1,2,\rho_3\}$	$\{4\}$	$\{3\}$
$(1,3)$	$\{1,5,\rho_3\}$	$\{2\}$	$\{4\}$	$\{3\}$
$(1,4)$	$\{5,\rho_3\}$	$\{1,2\}$	$\{4\}$	$\{3\}$

All conditions in Theorem 2 can now be satisfied (e.g., with $\mathcal{Q} = \{3\}$). A galled-tree with ρ_3 as the root sequence is shown on the right hand side of Figure 6.

4 Galled-Trees for M with Multiple Non-Trivial CCs in $G(M)$

A NASC for the case of multiple galls can be obtained readily by combining the results from the previous section with earlier results on galled-trees existing in the literature.

4.1 Root-known case

Suppose that the root sequence is ρ. It follows from the work of Gusfield *et al.* [5] that a NASC for the existence of a galled-tree for M with ρ as the root sequence is that, for each non-trivial connected component C of $G(M+\rho)$, there exists a galled-tree whose root sequence is ρ restricted to C. Together with our result in Theorem 2, this immediately implies the following:

Corollary 1. *Suppose that there are g non-trivial connected components $C_1, \ldots, C_g$ in $G(M + \rho)$. Then, a NASC for the existence of a galled-tree for M with ρ as the root sequence is that M satisfies the following conditions:*

E-1. *For every $a = 1, \ldots, g$, $\bigcap_{(i,j)\in\mathcal{P}(C_a)} I(i,j) \neq \varnothing$; i.e., every incompatibility in C_a overlaps with all other incompatibilities in C_a.*

E-2. *For every $a = 1, \ldots, g$, there exists a non-empty $\mathcal{Q}_a \in \bigcap_{(i,j)\in\mathcal{P}(C_a)} \pi^{ij}$ (i.e., there exists a common Q-set for each non-trivial connected component C_a), and*

E-3. *$\rho \notin \mathcal{Q}_a$ for all $a = 1, \ldots g$.*

A polynomial-time algorithm for checking these conditions is as follows.

Time complexity. Viewing each column to be a number in base 2, radix sort $M+\rho$ and group identical columns in $O(nm)$ time. If columns $i_1, i_2 \ldots, i_k$, where $i_1 < i_2 \ldots < i_k$, in $M+\rho$ are identical, keep only i_1 and let $L(i_1) = (i_1, i_2, \ldots, i_k)$ be the sorted list of columns identical to i_1. If the resulting matrix $\hat{M}$ contains more than $2n$ columns, then $M+\rho$ has no galled-tree [13]. Otherwise, find $O(n^2)$ incompatible pairs in $\hat{M}$ and create the incompatibility graph $G(\hat{M})$, containing g non-trivial CCs $C'_1, \ldots, C'_g$. This takes $O(n^3)$ time. Condition E-1 is satisfied by $M + \rho$ if and only if $\hat{M}$ satisfies that condition and every edge (x, y) in $G(\hat{M})$ satisfies either $\max(L(x)) < \min(L(y))$ or $\max(L(y)) < \min(L(x))$. This test takes $O(n^2)$ time. If E-1 is satisfied, then check E-2 and E-3 as follows in $O(n^3)$ time. Create a quadpartition π^{ij} for every $(i,j) \in \mathcal{P}(\hat{M})$ such that every subset $Q^{ij}(kl) \in \pi^{ij}$ is sorted, and let $\mathcal{R}^{ij} \in \pi^{ij}$ denote the subset that contains ρ. Conditions E-2 and E-3 are satisfied by $M + \rho$ if and only if $\bigcap_{(i,j)\in\mathcal{P}(C'_a)}(\pi^{ij} - \mathcal{R}^{ij}) \neq \varnothing$, for all $a = 1, \ldots, g$. In summary, the algorithm just described runs in $O(nm + n^3)$ time, which, in fact, is equal to the time bound of the root-known galled-tree construction algorithm devised by Gusfield *et al.* [5].

4.2 Root-unknown case

Gusfield [3] established that if the root sequence is unknown, a NASC for the existence of a galled-tree for M is that there is at least one row α in M such that there exists a galled-tree for M with α as the root sequence. That result and Corollary 1 together imply the following:

Corollary 2. *Suppose that there are g non-trivial connected components $C_1, \ldots, C_g$ in $G(M)$. Then, a NASC for the existence of a galled-tree for M is that M satisfies the following conditions:*

F-1. *For every $a = 1, \ldots, g$, $\bigcap_{(i,j)\in\mathcal{P}(C_a)} I(i,j) \neq \varnothing$.*

F-2. *For every $a = 1, \ldots, g$, there exists a non-empty $\mathcal{Q}_a \in \bigcap_{(i,j)\in\mathcal{P}(C_a)} \pi^{ij}$, and*

F-3. *$\bigcup_{a=1}^{g} \mathcal{Q}_a \neq X(M)$.*

The last condition ensures that there exists a row α not contained in any of $\mathcal{Q}_1, \ldots, \mathcal{Q}_g$, such that it may serve as the root sequence of a galled-tree. As described below, there exists a polynomial-time algorithm for checking these conditions.

Time complexity. An $O(nm + n^3)$ time algorithm similar to that (described above) for the root-known case can be used to determine whether there are less than or equal to $2n$ distinct columns in M, and whether conditions F-1 and F-2 are satisfied, finding all possible common $\mathcal{Q}$-sets for

each connected component C_a, $a = 1, \ldots, g$. (For the root-unknown case, we need to use M and π^{ij} instead of $M + \rho$ and $\pi^{ij} - \mathcal{R}^{ij}$, respectively.) Now, condition F-3 is satisfied if and only if the following holds: For some row $\alpha \in M$, every C_a, $a = 1, \ldots, g$, has at least one common $\mathcal{Q}$-set *not* containing α. Since there are $O(m) = O(n)$ CCs, this test can be done in $O(n^3)$ time. Hence, the algorithm just described runs in $O(nm+n^3)$ time, which is equal to the time bound of Gusfield's root-unknown galled-tree construction algorithm [3].

Example 4. Let M^* be the matrix $M^* = \begin{bmatrix} M & K \\ K & M \end{bmatrix}$, where M is the 6 by 4 matrix shown in Figure 5 and K is a matrix containing 6 identical rows of 0110. There are two isomorphic connected components C_1, C_2 in $G(M^*)$, with C_1 containing sites 1 through 4 and C_2 sites 5 through 8. From Example 2, we know that there exists a galled-tree for M with a single gall. To determine whether there exists a galled-tree for the entire matrix M^*, we need to check condition F-3 in Corollary 2. With $X_1 = \{1, \ldots, 6\}$ and $X_2 = \{7, \ldots, 12\}$, quadpartitions for $(i, j) \in \mathcal{P}(M^*)$ are as follows:

(i,j)	$Q^{ij}(00)$	$Q^{ij}(01)$	$Q^{ij}(10)$	$Q^{ij}(11)$
$(1,3)$	$\{2,3\}$	$\{4\} \cup X_2$	$\{1,6\}$	$\{5\}$
$(1,4)$	$\{4\} \cup X_2$	$\{2,3\}$	$\{5,6\}$	$\{1\}$
$(2,3)$	$\{1,2,6\}$	$\{5\}$	$\{3\}$	$\{4\} \cup X_2$
$(2,4)$	$\{5,6\}$	$\{1,2\}$	$\{4\} \cup X_2$	$\{3\}$
$(5,7)$	$\{8,9\}$	$\{10\} \cup X_1$	$\{7,12\}$	$\{11\}$
$(5,8)$	$\{10\} \cup X_1$	$\{8,9\}$	$\{11,12\}$	$\{7\}$
$(6,7)$	$\{7,8,12\}$	$\{11\}$	$\{9\}$	$\{10\} \cup X_1$
$(6,8)$	$\{11,12\}$	$\{7,8\}$	$\{10\} \cup X_1$	$\{9\}$

As this table shows, the common Q-set for C_1 is $\mathcal{Q}_1 = \{4\} \cup X_2$, whereas the common Q-set for C_2 is $\mathcal{Q}_2 = \{10\} \cup X_1$. Therefore, $\mathcal{Q}_1 \cup \mathcal{Q}_2 = \{1, \ldots, 12\} = X(M^*)$, which violates condition F-3. Hence, we conclude that there does not exist a galled-tree for M^*. In fact, it can be shown that $R_{\min}(M^*) = 3$.

Example 5. Consider the matrix M shown in Figure 2. There are two non-trivial CCs in the incompatibility graph $G(M)$, shown in Figure 7. It is easy to see from $G(M)$ that condition F-1 in Corollary 2 is satisfied. To check conditions F-2 and F-3, examine the following quadpartitions:

(i,j)	$Q^{ij}(00)$	$Q^{ij}(01)$	$Q^{ij}(10)$	$Q^{ij}(11)$
$(1,6)$	$\{1\}$	$\{2\}$	$\{5\}$	$\{3,4,6,7,8,9,10\}$
$(3,6)$	$\{1\}$	$\{2,3\}$	$\{5\}$	$\{4,6,7,8,9,10\}$
$(2,5)$	$\{9\}$	$\{1, \ldots, 5, 10\}$	$\{8\}$	$\{6,7\}$
$(4,5)$	$\{8\}$	$\{7\}$	$\{9\}$	$\{1, \ldots, 6, 10\}$
$(4,7)$	$\{8\}$	$\{7\}$	$\{9,10\}$	$\{1, \ldots, 6\}$
$(2,7)$	$\{9,10\}$	$\{1, \ldots, 5\}$	$\{8\}$	$\{6,7\}$

For connected component C_2, the common Q-set is $\mathcal{Q}_2 = \{8\}$. For connected component C_1, either $\{1\}$ or $\{5\}$ may be used as $\mathcal{Q}_1$ to satisfy condition F-2. For either choice, $\mathcal{Q}_1 \cup \mathcal{Q}_2 \neq X(M) = \{1, \ldots, 10\}$, and therefore condition F-3 is also satisfied. We thus conclude that there exists a galled-tree for M with exactly 2 galls. The one shown in Figure 2 is a galled-tree for this example.

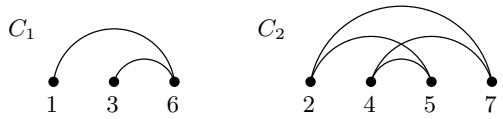

Fig. 7. Incompatibility graph, containing two non-trivial connected components, for the matrix shown in Figure 2.

ACKNOWLEDGMENT

The author gratefully acknowledges Dan Gusfield and Yufeng Wu for useful discussion. Dan Gusfield is also thanked for helpful comments on a preliminary version of this manuscript. This research is supported by NSF grants EIA-0220154 and IIS-0513910.

REFERENCES

[1] V. Bafna and V. Bansal. The number of recombination events in a sample history: conflict graph and lower bounds. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1:78–90, 2004.

[2] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3:479–502, 1996.

[3] D. Gusfield. Optimal, efficient reconstruction of Root-Unknown phylogenetic networks with constrained recombination. *J. Comput. Sys. Sci.*, 70:381–398, 2005.

[4] D. Gusfield, S. Eddhu, and C. Langley. The fine structure of galls in phylogenetic networks. *INFORMS J. on Computing, special issue on Computational Biology*, 16:459–469, 2004.

[5] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinf. Comput. Biol.*, 2:173–213, 2004.

[6] D. Gusfield, D. Hickerson, and S. Eddhu. An efficiently computed lower bound on the number of recombinations in phylogenetic networks: Theory and empirical study. *Discrete Applied Math*, in press.

[7] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, 98:185–200, 1990.

[8] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405, 1993.

[9] D. Hinds, L. Stuve, G. Nilsen, E. Halperin, E. Eskin, D. Gallinger, K. Frazer, and D. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307:1072–1079, 2005.

[10] T. N. D., Huynh, J. Jansson, N. B. Nguyen, and W.-K. Sung. Constructing a smallest refining galled phylogenetic network. *Proc. of RECOMB*, 265–280, 2005.

[11] L. Nakhleh, T. Warnow, and C.R. Linder. Reconstructing reticulate evolution in species – Theory and practice. *Proc. of RECOMB*, 337–346, 2004.

[12] Y. S. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *J. Math. Biol.*, 48:160–186, 2004.

[13] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *J. Comput. Biol.*, 8:69–78, 2001.