Generalizing the Splits Equivalence Theorem and Four Gamete Condition: Perfect Phylogeny on Three State Characters

Fumei Lam*, Dan Gusfield*, Srinath Sridhar**

Abstract. We study the three state perfect phylogeny problem and show that there is a three state perfect phylogeny for a set of input sequences if and only if there is a perfect phylogeny for every subset of three characters. In establishing these results, we prove fundamental structural features of the perfect phylogeny problem on three state characters and completely characterize the obstruction sets that must occur in input sequences that do not have a perfect phylogeny. We also give a proof for a stated lower bound involved in the conjectured generalization of our main result to any number of states.

1 Introduction

One of the fundamental problems in biology is the construction of phylogenies, or evolutionary trees, to describe ancestral relationships between a set of observed taxa. Each taxon is represented by a sequence and the evolutionary tree provides an explanation of branching patterns of mutation events transforming one sequence into another. There have been many elegant theoretical and algorithmic results on the problem of reconstructing a plausible history of mutations that generate a given set of observed sequences and to determine the minimum number of such events needed to explain the sequences. A widely used assumption in phylogeny construction is based on the *infinite sites model*, in which each state arises exactly once in the entire evolutionary history. In this case, the underlying phylogeny is called *perfect*.

The construction of perfect phylogenies has been extensively studied and in the case of binary sequences, the well-known *Splits Equivalence Theorem* (also known as the *four gamete condition*) gives a necessary and sufficient condition for the existence of a perfect phylogeny.

Theorem 1 (Splits Equivalence Theorem, Four Gamete Condition [11,15,27]). *A perfect phylogeny exists for binary input sequences if and only if no pair of characters contains all four possible binary pairs 00, 01, 10, 11.*

It follows from this theorem that for binary input, it is possible to either construct a perfect phylogeny, or output a pair of characters containing all four

* Department of Computer Science, University of California, Davis
** Department of Computer Science, Carnegie Mellon University

gametes as an obstruction set witnessing the nonexistence of a perfect phylogeny. This test is the building block for many theoretical results and practical algorithms. Among the many applications of this theorem, Gusfield et al. [18,19] and Huson et al. [24] apply the theorem to achieve decomposition theorems for phylogenies, Gusfield, Hickerson, and Eddhu [21] Bafna and Bansal [1,2], and Hudson and Kaplan [23] use it to obtain lower bounds for recombination events, Gusfield et al. [17,20] use it to obtain algorithms for constructing networks with constrained recombination, Sridhar et al. [5,32,33] and Satya et al. [29] use it to achieve a faster near-perfect phylogeny reconstruction algorithm, Gusfield [16] uses it to infer phase inference (with subsequent papers by Gusfield et al. [3,4,9,18], Eskin, Halperin, and Karp [10,22], Satya and Mukherjee [28] and Bonizzoni [7]), and Sridhar [31] et al. use it to obtain phylogenies from genotypes.

This work focuses on extending results for the binary perfect phylogeny problem to the multiple state character case, addressing the following natural questions arising from the Splits Equivalence Theorem. Given a set of sequences on r states ($r \geq 3$), is there a necessary and sufficient condition for the existence of a perfect phylogeny analogous to the Splits Equivalence Theorem? If no perfect phylogeny exists, what is the size of the smallest witnessing obstruction set?

In 1975, Fitch gave an example of input S over three states such that every *pair* of characters in S allows a perfect phylogeny while the entire set of characters S does not [12,13,14,30]. In 1983, Meacham generalized these results to characters over r states ($r \geq 3$)[27], constructing a class of sequences called *Fitch-Mecham examples*, which we examine in detail in Section 6. Mecham writes:

> "The Fitch examples show that any algorithm to determine whether a set of characters is compatible must consider the set as a whole and cannot take the shortcut of only checking pairs of characters." [27]

However, while the Fitch-Meacham construction does show that checking pairs of characters is not sufficient for the existence of a perfect phylogeny, our main result will show that for three state input, there is a sufficient condition which does *not* need to consider the entire set of characters simultaneously. In particular, we give a complete answer to the questions posed above for the three state case, by

1. showing the existence of a necessary and sufficient condition analogous to the Splits Equivalence Theorem (Sections 3, 4),
2. in the case no perfect phylogeny exists, proving the existence of a small obstruction set as a witness (Section 4),
3. giving a complete characterization of all minimal obstruction sets (Section 5), and
4. giving the proof for a stated lower bound involved in the conjectured generalization of our main result to any number of states (Section 6).

In establishing these results, we prove fundamental structural features of the perfect phylogeny problem on three state characters.

2 Perfect Phylogenies and Partition Intersection Graphs

The input to our problem is a set of n sequences (representing taxa), where each sequence is a string of length m over r states. Throughout this paper, the states under consideration will be the set $\{0, 1, 2, \ldots r-1\}$ (in particular, in the case $r = 2$, the input are sequences over $\{0, 1\}$). The input can be considered as a matrix of size $n \times m$, where each row corresponds to a sequence and each column corresponds to a character (or site). We denote characters by $\mathcal{C} = \{\chi^1, \chi^2, \chi^3, \ldots\ \chi^m\}$ and the states of character χ^i by χ^i_j for $0 \leq j \leq r-1$. A *species* is a sequence $s_1, s_2, \ldots s_m \in \chi^1_{j_1} \times \chi^2_{j_2} \times \cdots \chi^m_{j_m}$, where s_i is the *state* of character χ^i for s.

The *perfect phylogeny problem* is to determine whether an input set S can be displayed on a tree such that

1. each sequence in input set S labels exactly one leaf in T
2. each vertex of T is labeled by a species
3. for every character χ^i and for every state χ^i_j of character χ^i, the set of all vertices in T such that the state of character χ^i is χ^i_j forms a connected subtree of T.

Definition 1 ([8,30]). *For a set of input sequences S, the* partition intersection graph *$G(S)$ is obtained by associating a vertex for each character state and an edge between two vertices χ^i_j and χ^k_l if there exists a sequence s with state j in character $\chi^i \in \mathcal{C}$ and state l in character $\chi^k \in \mathcal{C}$. We say s is a row that* witnesses *edge (χ^i_j, χ^k_l). For a subset of characters $\Phi = \{\chi^{i_1}, \chi^{i_2}, \ldots \chi^{i_k}\}$, let $G(\Phi)$ denote the partition intersection graph $G(S)$ restricted to the characters in Φ.*

Note that by definition, there are no edges in the partition intersection graph between states of the same character.

Definition 2. *A graph H is* chordal*, or* triangulated*, if there are no induced chordless cycles of length four or greater in H.*

Consider coloring the vertices of the partition intersection graph $G(S)$ in the following way. For each character χ^i, assign a single color to the vertices $\chi^i_0, \chi^i_1, \ldots \chi^i_{r-1}$. A *proper triangulation* of the partition intersection graph $G(S)$ is a chordal supergraph of $G(S)$ such that every edge has endpoints with different colors. In [8], Buneman established the following fundamental connection between the perfect phylogeny problem and triangulations of the partition intersection graph.

Theorem 2. *[8,30] A set of taxa S admits a perfect phylogeny if and only if the corresponding partition intersection graph $G(S)$ has a proper triangulation.*

We will use Theorem 2 to extend the Splits Equivalence Theorem to a test for the existence of a perfect phylogeny on trinary state characters. To outline

our approach, suppose a perfect phylogeny exists for S and consider every subset of three characters. Then each of these $\binom{m}{3}$ subsets also has a perfect phylogeny. We show that this necessary condition is also sufficient and moreover, we can systematically piece together the proper triangulations for each triple of characters to obtain a triangulation for the entire set of characters. On the other hand, if no perfect phylogeny exists, then we show there exists a witness set of three characters for which no perfect phylogeny exists. This extends the Splits Equivalence Theorem to show that for binary and trinary state input, the number of characters needed for a witness obstruction set is equal to the number of character states. The following is the main theorem of the paper.

Theorem 3. *Given an input set S on m characters with at most three states per character ($r \leq 3$), S admits a perfect phylogeny if and only if every subset of three characters of S admits a perfect phylogeny.*

This theorem demonstrates that to verify that a trinary state input matrix S has a perfect phylogeny, it suffices to verify that partition intersection graphs $G[\chi^i, \chi^j, \chi^k]$ have proper triangulations for all triples $\chi^i, \chi^j, \chi^k \in \mathcal{C}$. In Section 6, we will show that the Fitch-Meacham examples [13,27] demonstrate that the size of the witness set in Theorem 3 is best possible.

3 Structure of Partition Intersection Graphs for Three Characters

We begin by studying the structure of partition intersection graphs on three characters with at most three states per character ($m \leq 3$, $r \leq 3$). For convenience, we will denote the three characters by the letters a, b, c (interchangeably referring to them as characters and colors) and denote the states of these characters by a_i, b_i, c_i ($i \in \{0, 1, 2\}$).

The problem of finding proper triangulations for graphs on at most three colors and arbitrary number of states ($m = 3$, r arbitrary) has been studied in a series of papers [6,25,26]. However, it will be unnecessary in our problem to employ these triangulation algorithms, as our instances will be restricted to those arising from character data on at most three states ($m = 3, r \leq 3$). In such instances, we will show that if a proper triangulation exists, then the structure of the triangulation is very simple. We begin by proving a sequence of lemmas characterizing the possible cycles contained in the partition intersection graph.

Lemma 1. *Let S be a set of input species on three characters a, b, and c with at most three states per character. Suppose every pair of characters induces a properly triangulatable character partition intersection graph (i.e., $G[a, b]$, $G[b, c]$ and $G[a, c]$ are properly triangulatable) and let C be a chordless cycle in $G[a, b, c]$. Then C cannot contain all three states of any character.*

Proof. Suppose there is a color, say a, such that all three states a_0, a_1 and a_2 appear in C. Note that C must contain all three colors a, b, and c (since any pair of colors induces a properly triangulatable graph). We have the following cases.

Case I. Suppose there is an edge e in C neither of whose endpoints have color a (without loss of generality, let $e = (b_0, c_0)$). The row that witnesses this edge must contain some state in a, say a_0. This implies that the vertices a_0, b_0, and c_0 form a triangle in $G[a, b, c]$, a contradiction since C is assumed to be chordless (see Figure 1).

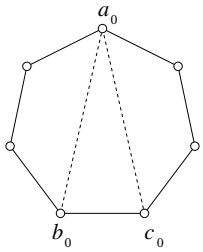


Fig. 1. The row witnessing edge (b_0, c_0) must contain a state in character a .

Case II. Otherwise, every edge has an endpoint of color a and implying each edge has color pattern either (a, b) or (a, c). Since all three states of a appear, the color pattern up to relabeling must be as shown in Figure 2(a) (in the figure, color b appears twice and color c appears once).

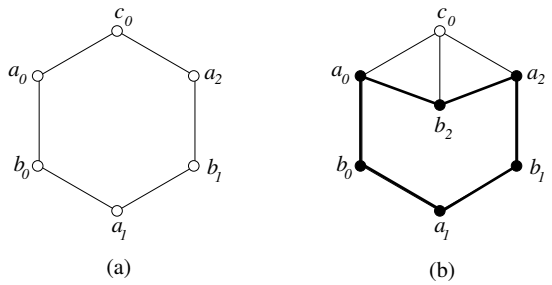


Fig. 2. The row witnesses for edges (a_0, c_0) and (c_0, a_2) must share the same state of b.

In this case, the row witness for edge (a_0, c_0) must contain the final state b_2 of b (otherwise there would be an edge between c_0 and either b_0 or b_1, a contradiction since C is chordless). Similarly, the row witness for edge (c_0, a_2) must also be state b_2. As shown in Figure 2(b), this gives a cycle (a_0, b_2), (b_2, a_2), (a_2, b_1), (b_1, a_1), (a_1, b_0), (b_0, a_0) on two colors. Such a cycle is not properly triangulatable, and therefore $G[a, b]$ is not properly triangulatable, a contradiction.

Since Case I and Case II cannot occur, it follows that a_0, a_1 and a_2 cannot all appear in C, proving the lemma. □

Before stating the next lemma, we give the following definition.

Definition 3. *Suppose the endpoints of edge e have colors χ^i and χ^j. Then any other edge whose endpoints also have colors χ^i and χ^j is called* color equivalent *to e. Two edges are called* nonadjacent *if they do not share a common endpoint.*

For example, the edges (c_1, a_2) and (c_0, a_1) in Figure 1 are nonadjacent and color equivalent.

Lemma 2. *Let S be a set of input species on three characters a, b, and c with at most three states per character. If the partition intersection graph $G[a, b, c]$ is properly triangulatable, then for every chordless cycle C in $G[a, b, c]$, there exists a color (a, b, or c) that appears exactly once in C.*

Proof. Consider any chordless cycle C of $G[a, b, c]$. By Lemma 1, no color appears in all three states in C. To obtain a contradiction, suppose each color a, b, and c appears exactly twice in C and relabel the states so that the vertices appearing on the cycle are a_0, a_1, b_0, b_1, c_0, and c_1. We first show that C has a pair of nonadjacent edges that are color equivalent. Up to symmetry and relabelling of colors, there are two cases for the color pattern of C as follows.

Case 1. There is a vertex in the cycle whose neighbors in the cycle have the same color. Up to relabeling, we can assume this vertex has color a (say in state a_0) and the two adjacent vertices have color b. The states for the remaining vertices of the cycle are a_1, c_0, and c_1. Now, consider the vertices adjacent to b_0 and b_1 other than a_0. These vertices must be c_0 and c_1 (otherwise, the two states of c would be adjacent in the cycle). This color pattern is shown in Figure 3(a).

Case 2. No vertex in the cycle is adjacent to two vertices of the same color. Then the two neighbors of a vertex with color a must have colors b and c. Then the vertex following b in the cycle must have color c (otherwise vertex b is adjacent to two vertices of the same color). By working this way around the cycle, the only color pattern possible is as shown in Figure 3(b).

Note that both color patterns contain a pair of nonadjacent and color equivalent edges (edges e and e' in Figure 3).

Consider this pair of nonadjacent and color equivalent edges e and e'. Without loss of generality, assume that the endpoints of these edges have colors b and c. Let s be the row witness for e and s' be the row witness for e'. Since cycle C is chordless, the state in character a of row s cannot be a_0 or a_1. Similarly, the state in character a of row s' cannot be a_0 or a_1. Since a_2 is the only remaining state of character a, both s and s' must contain a_2. This implies that the partition intersection graph $G[a, b, c]$ must induce one of the two color patterns in Figure 4.

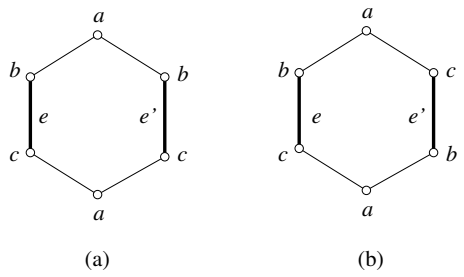


Fig. 3. Color Patterns and nonadjacent color equivalent edges e and e'.

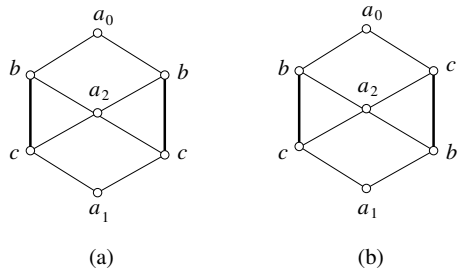


Fig. 4. Induced Color Patterns

In the case illustrated in Figure 4(a), there is a cycle on four vertices induced by the two characters a and b (see Figure 5(a)), implying $G[a, b, c]$ is not properly triangulatable. In the case illustrated in Figure 4(b), there are two edge-disjoint cycles of length four with color pattern a, b, a, c. Since edges in a proper triangulation cannot connect vertices of the same color, any proper triangulation of G must contain the two edges f and f' connecting vertices of color b and c (see Figure 5(b)). However, this induces a cycle of length four on the states of b and c, which does not have a proper triangulation. This again shows that $G[a, b, c]$ is not properly triangulatable.

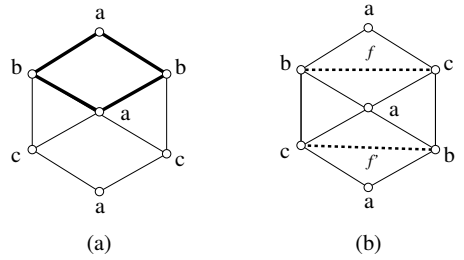


Fig. 5. (a) Induced cycle of length four on two colors; (b) Forced Edges f and f'

Since all of these cases result in contradictions, it follows that there exists a color that appears exactly once in C. □

Lemmas 1 and 2 show that if C is a chordless cycle in a properly triangulatable graph $G[a, b, c]$, then no color can appear in all three states and one color appears uniquely. This leaves two possibilities for chordless cycles in $G[a, b, c]$ (see Figure 6):

- a chordless four cycle, with two colors appearing uniquely and the remaining color appearing twice
- a chordless five cycle, with one color appearing uniquely and the other two colors each appearing twice

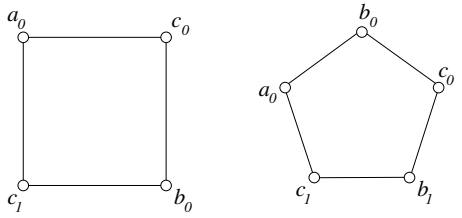


Fig. 6. The only possible chordless cycles in $G[a, b, c]$: (a) characters a and b appear uniquely while character c appears twice; (b) character a appears uniquely while characters b and c each appear twice.

In the next lemma, we show that if $G[a, b, c]$ is properly triangulatable, the second case cannot occur, i.e., $G[a, b, c]$ cannot contain a chordless five cycle.

Lemma 3. *Let S be a set of input species on three characters a, b, and c with at most three states per character. If the partition intersection graph $G[a, b, c]$ is properly triangulatable, then $G[a, b, c]$ cannot contain chordless cycles of length five or greater.*

Proof. Lemmas 1 and 2 together show that $G[a, b, c]$ cannot contain chordless cycles of length six or greater, so it remains to show that $G[a, b, c]$ cannot contain chordless cycles of length equal to five.

Suppose C is a chordless cycle in $G[a, b, c]$ of length five; without loss of generality, let a be the color appearing exactly once in C (say in state a_0), let b_0, b_1 be the two states of b in C, and let c_0, c_1 be the two states of c in C. Up to relabeling of the states, the cycle is as shown in Figure 6(b).

Now, any proper triangulation of $G[a, b, c]$ must triangulate cycle C by edges (a_0, c_0) and (a_0, b_1) shown in Figure 7 (since the only other edge between nonadjacent vertices of different colors is (b_0, c_1), which would create a non-triangulatable four cycle on the two colors b and c.

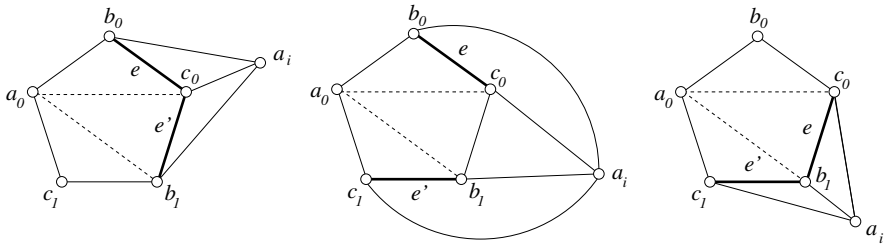


Fig. 7. Edges e and e' are both witnessed by state a_i.

The row witnesses for edges (b_0, c_0), (c_0, b_1), and (c_1, b_1) must contain a state in color a that is one of a_1 or a_2 (otherwise, a_0 would have an edge to a non-adjacent vertex in cycle C, implying C is not chordless). Since there are three edges and two possible witness states in color a, there are two edges among (b_0, c_0), (c_0, b_1), (c_1, b_1) that share a witness a_i. We denote these two edges by e and e'; as shown in Figure 7, there are three ways to choose e and e'.

Figure 8 shows that all three cases induce a four cycle on two colors, a contradiction since $G[a, b, c]$ is properly triangulatable. Therefore, $G[a, b, c]$ cannot contain a chordless 5-cycle. □

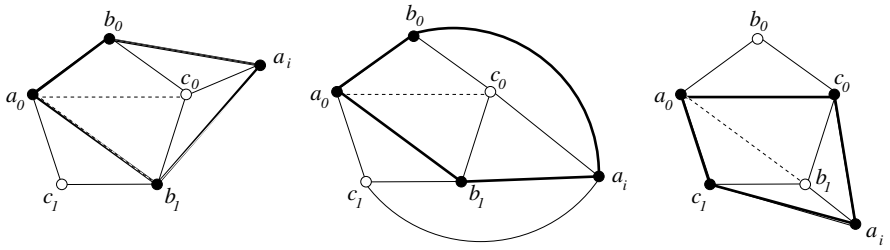


Fig. 8. Forced cycles of length four on two colors.

Lemma 4. *Let S be a set of input species on three characters a, b, and c with at most three states per character. If the partition intersection graph $G[a, b, c]$ is properly triangulatable, then every chordless cycle in $G[a, b, c]$ is* uniquely *triangulatable.*

Proof. By Lemma 3, if C is a chordless cycle in $G[a, b, c]$, then C must be a four cycle with the color pattern shown in Figure 9 (up to relabeling of the colors). Then C is uniquely triangulatable by adding the edge between the two colors appearing uniquely (in Figure 9, these are colors a and b).

□

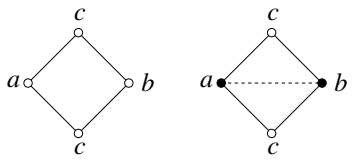


Fig. 9. Color pattern for chordless cycle C.

For any three colors a, b, c, Lemma 4 gives a simple algorithm to properly triangulate $G[a, b, c]$: for each chordless cycle C in $G[a, b, c]$, check that C is a four cycle with two nonadjacent vertices having colors that appear exactly once in C and add an edge between these two vertices.

4 The 3-SNP test

4.1 Triangulating Triples of Characters

We now consider the case of trinary input sequences S on m characters (for m greater or equal to 4). Our goal is to prove that the existence of proper triangulations for all subsets of three characters at a time is a sufficient condition to guarantee existence of a proper triangulation for *all* m characters.

By Lemma 2, if a set of three characters χ^i, χ^j, χ^k is properly triangulatable, then there is a unique set of edges $F(\chi^i, \chi^j, \chi^k)$ that must be added to triangulate the chordless cycles in $G[\chi^i, \chi^j, \chi^k]$. Construct a new graph $G'(S)$ on the same vertices as $G(S)$ with edge set $E(G(S)) \cup \{\cup_{1 \leq i < j < k \leq m} F(\chi^i, \chi^j, \chi^k)\}$. $G'(S)$ is the partition intersection graph $G(S)$ together with all of the additional edges used to properly triangulate chordless cycles in $G[\chi^i, \chi^j, \chi^k]$ $(1 \leq i < j < k \leq m)$. In $G'(S)$, edges from the partition intersection graph $G(S)$ are called E-edges and edges that have been added as triangulation edges for some triple of columns are called F-edges. We call a cycle consisting only of E-edges an *E-cycle*.

Example 1. Input set S and partition intersection graph $G(S)$ are shown in Figure 10. Each triple of characters in S induces a chordal graph while the entire partition intersection graph $G(S)$ contains a chordless cycle of length four. Since each triple of characters induces a chordal graph, no F-edges are added and $G(S) = G'(S)$.

As Example 1 illustrates, the addition of F-edges alone may not be sufficient to triangulate the entire partition intersection graph. We now turn to the problem of triangulating the remaining chordless E-cycles in $G'(S)$.

Consider any E-cycle C that is chordless in $G'(S)$ satisfying the properties

1. C has length equal to four
2. all colors of C are distinct

S

a	b	c	d
0	0	0	0
1	0	1	2
2	1	1	1
0	2	2	1

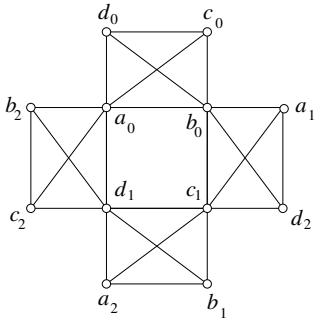


Fig. 10. Example 1. Partition intersection graph $G'(S)$ contains a chordless four cycle.

For every such chordless cycle, add the chords between the two pairs of nonadjacent vertices in C (note that these are legal edges). Call this set of edges F'-edges and let $G''(S)$ denote the graph $G'(S)$ with the addition of F'-edges. Note that the sets of E-edges, F-edges, and F'-edges are pairwise disjoint; we call the set of F and F'-edges *non-E edges*.

We begin by investigating structural properties of cycles in $G'(S)$ and $G''(S)$ containing at least one F-edge or F'-edge. Let C be a cycle in $G'(S)$ or $G''(S)$ containing an edge f that is an F-edge or F'-edge (without loss of generality, let $f = (a_0, b_0)$). This edge must be added due to an E-cycle D containing a_0, b_0 and two other vertices w and z as shown in Figure 11(a) (note that w and z cannot have color a or b). If f is an F-edge, then w and z have the same color and therefore cannot be adjacent in $G'(S)$. If f is an F'-edge, then since D is a chordless E-cycle in $G'(S)$, w and z are nonadjacent in $G'(S)$. The cycle C created by edge f is shown in Figure 11(b).

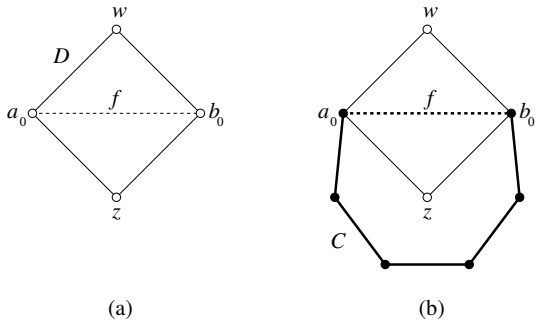


Fig. 11. (a) Chordless cycle D (b) edge $f = (a_0, b_0)$ creates cycle C (shown in bold).

Since D is an E-cycle, each edge in D has a row witness. Consider first the row witnesses for edges (a_0, w) and (a_0, z). These row witnesses must contain a state of b other than b_0 (since a_0 and b_0 are not connected by an E-edge). If both row witnesses share the same state b_i of b, then the cycle (b_i, w), (w, b_0), (b_0, z), (z, b_i) is a chordless E-cycle on at most three colors in $G'(S)$ as shown in Figure 12 (as argued above, w and z are nonadjacent in $G'(S)$). However, all chordless E-cycles on at most three colors have been triangulated in $G'(S)$, a contradiction.

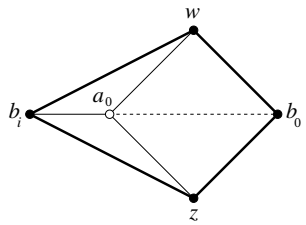


Fig. 12. If the row witnesses for (a_0, w) and (a_0, z) share a state of b, there is a chordless E-cycle of length four on at most three colors.

Therefore, the row witnesses for (a_0, w) and (a_0, z) cannot share the same state of b. Similarly, the row witnesses for (b_0, w) and (b_0, z) cannot share the same state of a. This implies the following situation, up to relabeling of the states, illustrated in Figure 13.

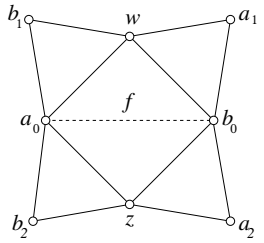


Fig. 13. Pattern of forced witnesses for edges in D.

In particular, the following two conditions must be satisfied.

$$\begin{array}{l}(1) a_0 \text{ is adjacent to both } b_1 \text{ and } b_2 \\ \\ (2) b_0 \text{ is adjacent to both } a_1 \text{ and } a_2\end{array} \qquad (*)$$

We use this structure to prove a sequence of lemmas eliminating the possibilities for chordless cycles in graph $G'(S)$. This sequence of lemmas will show $G'(S)$ cannot contain a chordless cycle with exactly one non-E edge (Lemmas 5 and 6), a chordless cycle with two or more non-E edges (Lemma 7), or a chordless E-cycle (Corollary 1).

Lemma 5. *$G'(S)$ cannot contain a chordless cycle with exactly one F-edge.*

Proof. Suppose that C is a chordless cycle in $G'(S)$ with exactly one F-edge, say $f = (a_0, b_0)$. Edge (a_0, b_0) must have been added due to a chordless E-cycle D on three colors as shown in Figure 13, where w and z are states of the same color. Note that edge (a_0, b_0) is a forced F-edge that creates cycle C (see Figure 14). If C contains only the two colors a and b, the partition intersection graph on the three colors a, b, and the shared color of w and z is not properly triangulatable, a contradiction.

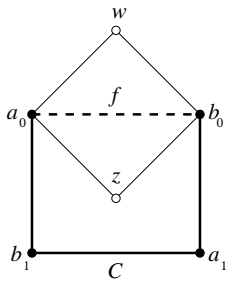


Fig. 14. Chordless cycle C on two colors with exactly one F-edge.

This implies any cycle C in $G'(S)$ with exactly one F-edge must contain three or more colors. As shown in Figure 15, if any of the edges (b_1, a_1), (b_1, a_2), (b_2, a_1), and (b_2, a_2) are present, there would be a chordless cycle on two colors with exactly one F-edge, which we have argued cannot occur. It follows that a_1 and a_2 are nonadjacent to b_1 and b_2 by E-edges.

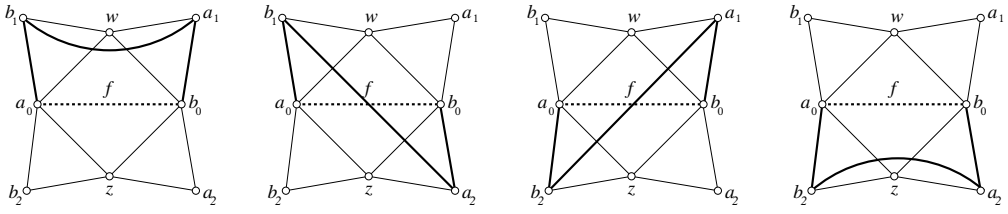


Fig. 15. If any of (b_1, a_1), (b_1, a_2), (b_2, a_1), or (b_2, a_2) are E-edges, there is a chordless four cycle in $G'(S)$ on two colors with exactly one F-edge.

Since a_1 is nonadjacent to b_1 or b_2 by E-edges, any row that contains a_1 must contain state b_0 in character b. We call this condition (A1). By a similar argument, the following conditions must be satisfied:

(A2) any row that contains a_2 must contain state b_0 in character b.
(B1) any row that contains b_1 must contain state a_0 in character a.
(B2) any row that contains b_2 must contain state a_0 in character a.

Now, let x be a vertex in $C\backslash\{a_0, b_0\}$ and consider the state of character a in any row that witnesses x (see Figure 16(a)). If this state is a_0, then x is adjacent to a_0 by an E-edge. Otherwise, if this state is either a_1 or a_2, then this row witness for x must contain state b_0 by (A1) and (A2). Since C is a chordless cycle, at most one vertex on $C\backslash\{a_0, b_0\}$ can be adjacent to each of a_0 and b_0. This shows there can be at most two such vertices x_1 and x_2 in $C\backslash\{a_0, b_0\}$, one of which is adjacent to a_0 and the other which is adjacent to b_0 (moreover, these are adjacencies by E-edges). Therefore, C has length equal to four formed by edges (a_0, x_1), (x_1, x_2), (x_2, b_0), and (b_0, a_0) (see Figure 16(b)).

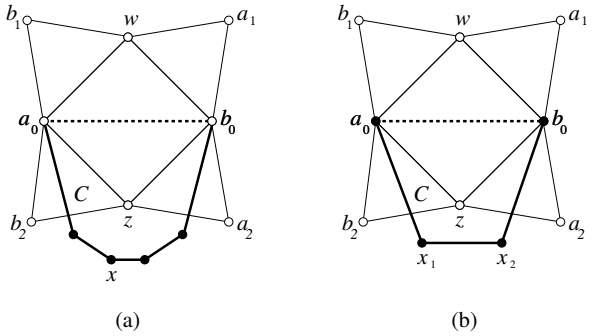


Fig. 16. Vertices on cycle C.

Edge (x_1, x_2) is an E-edge since $f = (a_0, b_0)$ is the unique F-edge in C by assumption. Furthermore, at least one of x_1 or x_2 has color different from a and b since C has three or more colors. Without loss of generality, assume vertex x_1 has color different from a and b. The color of x_2 is different from b since x_2 and b_0 are adjacent, implying edge (x_1, x_2) must have a witness in character b. If this witness is b_0, then x_1 and b_0 are adjacent by an E-edge, a contradiction to the chordlessness of cycle C. If this witness is either b_1 or b_2, then (B1) or (B2) imply that x_2 and a_0 are adjacent by an E-edge, again a contradiction to the chordlessness of cycle C.

This concludes the proof of Lemma 5. □

A similar proof shows that the lemma can be extended to the graph $G''(S)$.

Lemma 6. *$G''(S)$ cannot conatin a chordless cycle with exactly one non-E edge.*

Proof. Suppose that C is a chordless cycle in $G''(S)$ with exactly one non-E edge, say $f = (a_0, b_0)$. If f is an F-edge, then C would be a chordless cycle in $G'(S)$ with exactly one F-edge, contradicting Lemma 5. Therefore f is an F'-edge that is added due to chordless cycle D as shown in Figure 13 (with w and z different colors).

Case I. C contains only the two colors a and b. Since the graph in Figure 13 contains all the states of characters a and b, C must also contain one of the edges (b_1, a_1), (b_1, a_2), (b_2, a_1), or (b_2, a_2) as an E-edge and we have the following cases.

Case I(i). C contains edge (b_2, a_2). This results in an E-cycle of length five on at most three colors as shown in Figure 17(a). Such a cycle cannot be chordless in $G[a, b, w]$ by Lemma 3. Therefore, vertex w must be adjacent to one of b_2 or a_2 by an E-edge. This creates a chordless E-cycle in $G'(S)$ of length four on three colors; either cycle (b_2, w), (w, b_0), (b_0, z), (z, b_2) shown in Figure 17(b) or cycle (a_0, w), (w, a_2), (a_2, z), (z, a_0) shown in Figure 17(c) (note that w and z are nonadjacent in $G'(S)$ since cycle D is chordless). This is a contradiction since all cycles on at most three colors are triangulated in $G'(S)$.

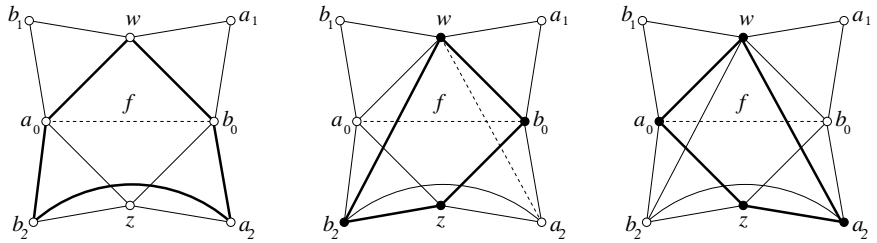


Fig. 17. (a) Edge (b_2, a_2) gives a five cycle C on at most three colors C (b),(c) Chordless cycle of length four containing three colors.

Case I(ii). C contains edge (b_1, a_1). This case is symmetric to Case I(i).
Case I(iii). C contains edge (b_2, a_1). This results in an E-cycle of length four on at most three colors as shown in Figure 18(a). Such a cycle is triangulated in $G'(S)$, implying there is either an E-edge or an F-edge between b_2 and w. Then the cycle (b_2, w), (w, b_0), (b_0, z), (z, b_2) is either a E-chordless cycle in $G'(S)$ (a contradiction since all E-cycles on at most three colors are triangulated in $G'(S)$) or a chordless cycle in $G'(S)$ with exactly one F-edge (contradicting Lemma 5).
Case I(iv). C contains edge (b_1, a_2). This case is symmetric to Case I(iii).

It follows that none of the vertex pairs (b_1, a_1), (b_1, a_2), (b_2, a_1), and (b_2, a_2) are adjacent by an E-edge and any cycle C in $G''(S)$ with exactly one non-E

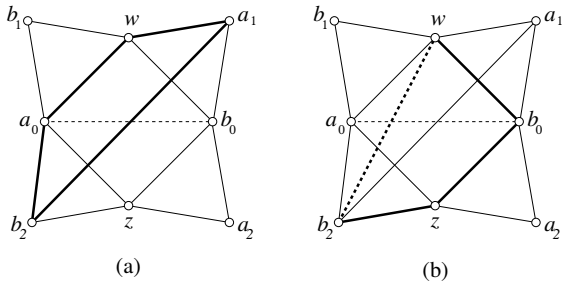


Fig. 18. (a) E-cycle of length four on three colors (b) Chordless cycle on three colors with exactly one F'-edge.

edge must contain three or more colors. Because of these nonadjacencies, the statements (A1), (A2), (B1), (B2) from Lemma 5 hold.

(A1) any row that contains a_1 must contain state b_0 in character b.
(A2) any row that contains a_2 must contain state b_0 in character b.
(B1) any row that contains b_1 must contain state a_0 in character a.
(B2) any row that contains b_2 must contain state a_0 in character a.

As in the proof of Lemma 5, it follows that cycle C has length equal to four formed by edges (a_0, x_1), (x_1, x_2), (x_2, b_0), and (b_0, a_0) (see Figure 19(b)).

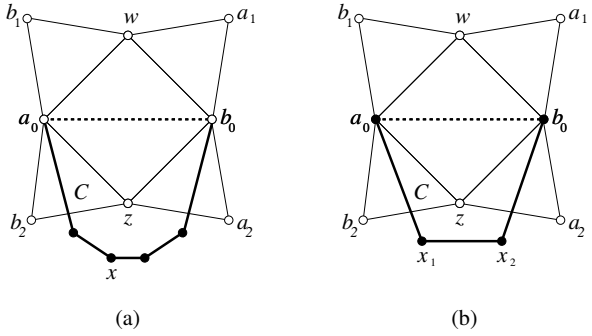


Fig. 19. Vertices on cycle C.

Now, edge (x_1, x_2) is an E-edge since $f = (a_0, b_0)$ is the unique non-E edge in C by assumption. The remainder of the proof follows exactly as in the proof of Lemma 5. □

We now consider chordless cycles in $G''(S)$ with two or more non-E edges.

Lemma 7. *$G''(S)$ cannot contain a chordless cycle with two or more non-E edges.*

Proof. Suppose otherwise and let C be a chordless cycle in $G''(S)$ with two or more non-E edges. Let f be one of the F or F'-edges in C and without loss of generality, let $f = (a_0, b_0)$. This edge must have been added due to an E-cycle D that is chordless in $G(S)$ on a_0, b_0 and two other vertices w and z (see Figure 11(a)). If f is an F-edge, then w and z have the same color and therefore are not adjacent in $G'(S)$. If f is an F'-edge, then w and z have different colors and are nonadjacent in $G'(S)$ (since they are nonadjacent vertices in chordless cycle D).

As argued previously, the situation up to relabeling of the states is illustrated in Figure 13. Furthermore, the proofs of Lemmas 5 and 6 establish conditions (A1), (A2), (B1), and (B2), implying C has length equal to four formed by edges (a_0, x_1), (x_1, x_2), (x_2, b_0), and (b_0, a_0) (see Figure 19(b)). Then since C has two or more non-E edges, the edge (x_1, x_2) in C is a non-E edge.

We have the following cases for the vertices of C.

Case I. One of x_1 and x_2 has color a and the other has color b. We can assume without loss of generality that $x_1 = b_2$ and $x_2 = a_2$ as illustrated in Figure 20(a). Since edge (x_1, x_2) is either an F-edge or an F'-edge, it was added because of a chordless E-cycle D' containing a_2, b_2 and two other vertices y_0 and y_1 (see Figure 20(a)). By (A2), both y_0 and y_1 are adjacent to a_0, giving an E-cycle of length four (a_0, y_0), (y_0, a_2), (a_2, y_1), (y_1, a_0) on at most three colors (see Figure 20(c)). This E-cycle must be triangulated in $G'(S)$. However, this cannot be the case since D' is a chordless cycle in $G'(S)$ and y_0 and y_1 are nonadjacent vertices in D'.

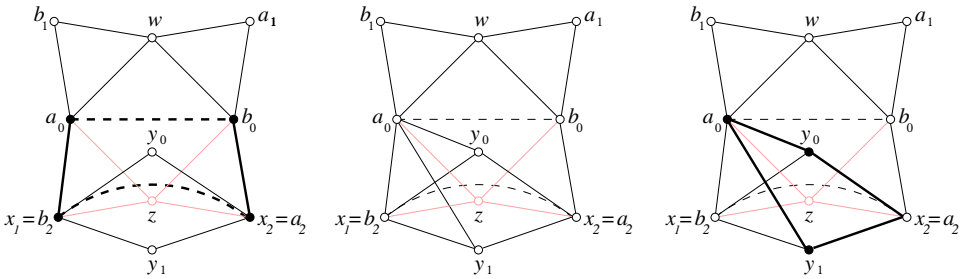


Fig. 20. Case I.

Case II. The color of x_1 is different from a and b and the color of x_2 is a or b. Without loss of generality, assume $x_2 = a_2$ as illustrated in Figure 21(a). Since edge $(x_1, x_2)(= (x_1, a_2))$ is an F-edge or F'-edge, it was added due to a chordless four cycle on $x_1, x_2(= a_2)$ and two other vertices y_0 and y_1. The row witnesses for edges (x_1, y_0) and (x_1, y_1) must contain state a_0 (otherwise, x_1 would be adjacent to b_0 by (A1) or (A2)). Then we have the E-cycle of length four (a_0, y_0), (y_0, a_2), (a_2, y_1), (y_1, a_0) on at most three colors. This E-cycle must be triangulated in

$G'(S)$. However, this cannot happen since y_0 and y_1 are nonadjacent vertices in cycle D'.

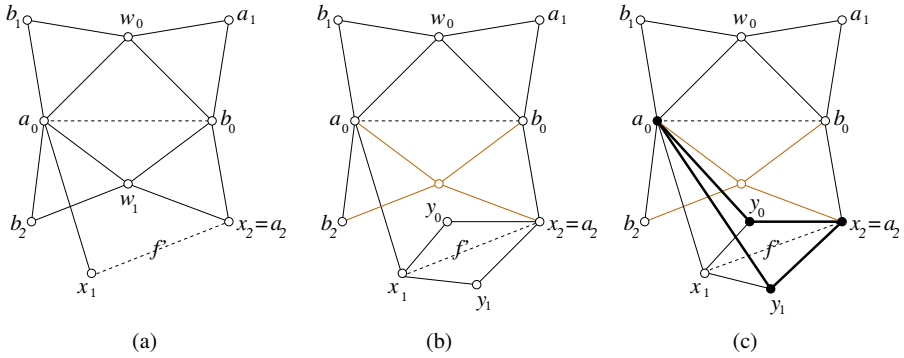


Fig. 21. Case II.

Case II'. The color of x_1 is a or b and the color of x_2 is different from a and b. This case is symmetric to that in Case II and is shown in Figure 22.

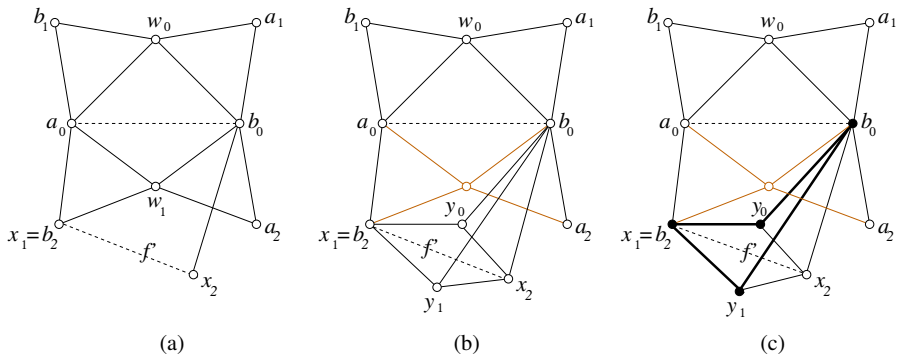


Fig. 22. Case II'.

Case III. Both x_1 and x_2 have colors different from a and b. Since edge (x_1, x_2) is an F-edge or F'-edge, it was added due to a chordless four cycle on x_1, x_2 and two other vertices y_0 and y_1. The row witnesses for edges (x_1, y_0) and (x_1, y_1) must contain state a_0 (otherwise, x_1 would be adjacent to b_0 by (A1) or (A2)). Then (a_0, y_0), (y_0, x_2), (x_2, y_1), (y_1, a_0) is an E-cycle of length four (see Figure 23(c)). Note that this cycle is chordless in $G'(S)$, since a_0 and x_2 are nonadjacent vertices in chordless cycle C and y_0 and y_1 are nonadjacent vertices in chordless cycle D'. If y_0 and y_1 have the same color, then C has only three colors and this

would force edge (a_0, x_2) to be an F-edge, a contradiction to the assumption that C is chordless in $G''(S)$. Therefore, the colors of a_0, x_2, y_0, y_1 are all distinct. This cycle would force edges (a_0, x_2) and (y_0, y_1) to be added as F'-edges, a contradiction since cycle C is chordless in $G''(S)$ and a_0 and x_2 are nonadjacent vertices in C.

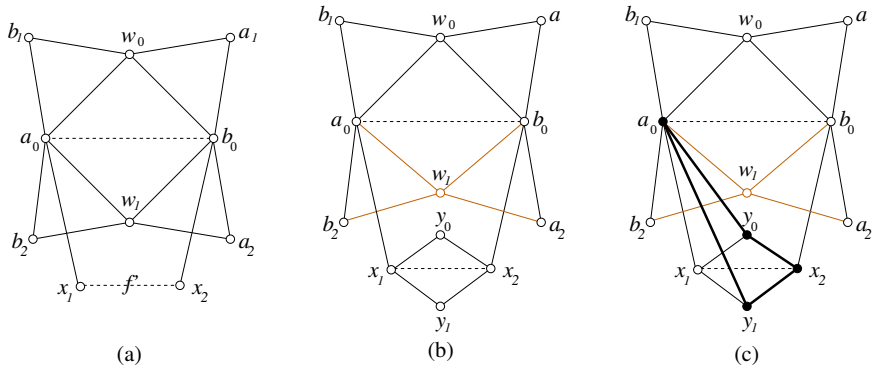


Fig. 23. Case III.

This proves the lemma. □

Lemmas 5, 6, and 7 eliminate the possibility of chordless cycles in $G''(S)$ containing non-E edges. To show that $G''(S)$ is properly triangulated, we proceed to show that $G''(S)$ does not contain chordless E-cycles. Suppose C is an E-cycle of length five or greater that is chordless in $G'(S)$ and suppose there is a character a that appears exactly once (say in state a_0) in C. Label the edges of the path $C \backslash a_0$ in order of appearance by $e_1, e_2, e_3, \ldots e_{k-1}$ with $e_i = (v_i, v_{i+1})$. Since C is chordless and all edges in C are E-edges, each edge e_i $(i = 1, 2, \ldots k-1)$ must be witnessed by a row s_i which contains either state a_1 or a_2 in color a. Without loss of generality, assume e_1 is witnessed by a_1 and let j be the largest index such that e_j is witnessed by a_1. If j is equal to $k-1$, then this creates a four cycle $(v_1, a_0), (a_0, v_k), (v_k, a_1), (a_1, v_1)$ on E-edges (see Figure 24(b)). Since v_1 and v_k are nonadjacent (by the chordlessness of C in $G'(S)$), this creates an E-cycle on at most three colors that is chordless in $G'(S)$, which cannot occur.

Therefore, j must be strictly less than $k-1$ and all of the remaining edges $e_{j+1}, \ldots e_{k-1}$ are witnessed by state a_2. Define the *a-complete cycle induced by cycle C and state a_0* as follows (see Figure 25):

$$I(C, a_0) = \begin{cases} (a_0, v_1), (v_1, v_2), (v_2, a_2), (a_2, v_k), (v_k, a_0) \text{ if } j = 1 \\ (a_0, v_1), (v_1, a_1), (a_1, v_{j+1}), (v_{j+1}, a_2), (a_2, v_k), (v_k, a_0) \text{ if } 1 < j < k-2 \\ (a_0, v_1), (v_1, a_1), (a_1, v_{k-1}), (v_{k-1}, v_k), (v_k, a_0) \text{ if } j = k-2 \end{cases}$$

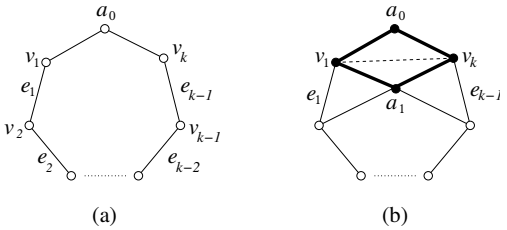


Fig. 24. Chordless Cycle C

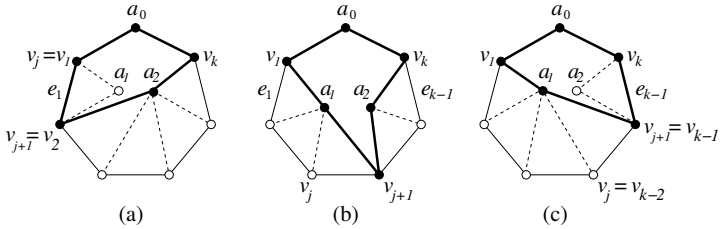


Fig. 25. The a-complete cycle induced by color C and state a_0 $I(C, a_0)$ (a) $j = 1$ (b) $1 < j < k - 1$ (c) $j = k - 2$.

Observation 4 *For an E-cycle C such that*

(i) C is chordless in $G'(S)$
(ii) C has length five or greater
(iii) C contains a character a appearing exactly once in state a_0,

the a-complete cycle $I(C, a_0)$ exists. Note that $I(C, a_0)$ contains at least two vertices of color a and has length five or greater.

We use this construction to prove the following lemma.

Lemma 8. *Suppose C is an E-cycle of length five or greater that is chordless in $G'(S)$ and suppose there is a character a appearing uniquely in C in state a_0. Then the two vertices adjacent to a_0 in C have different colors and $I(C, a_0)$ is an E-cycle that is chordless in $G'(S)$.*

Proof. Note that $I(C, a_0)$ exists by Observation 4 and all edges in $I(C, a_0)$ are E-edges. We show $I(C, a_0)$ is chordless in $G'(S)$. The vertex pairs (a_1, v_k) and (a_2, v_1) are not adjacent in $G'(S)$; otherwise we would obtain a four cycle on at most three colors with at most one F-edge that is chordless in $G'(S)$ (see Figure 26). This is a contradiction, since Lemma 5 implies $G'(S)$ cannot contain a chordless cycle with at most one F-edge. The remaining vertex pairs in $I(C, a_0)$ are in C and are nonadjacent in $G'(S)$ since C is chordless in $G'(S)$. It follows that $I(C, a_0)$ is chordless in $G'(S)$.

Now suppose for a contradiction that the vertices adjacent to a_0 (vertices v_1 and v_k in Figure 25) have the same color. Then $I(C, a_0)$ is a cycle on at most

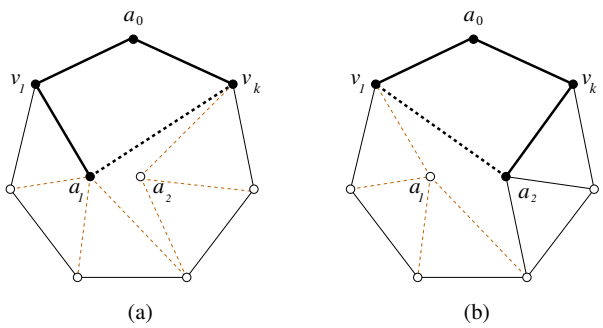


Fig. 26. If either (a_1, v_k) or (a_2, v_1) are adjacent, there is a cycle on at most three colors with at most one F-edge.

three colors (color a, the color of v_{j+1}, and the shared color of vertices v_1 and v_k). This is an E-cycle that has length five or greater and is chordless in the partition intersection graph on these three colors. This is forbidden by Lemma 3. Therefore, the two vertices adjacent to a_0 are states in two different colors.

This proves the lemma. □

We now use this construction to prove properties of chordless E-cycles in $G'(S)$.

Lemma 9. *If C is an E-cycle that is chordless in $G'(S)$, then C has length exactly four with four distinct colors.*

Proof. Suppose C is a chordless E-cycle in $G'(S)$. Note that C must contain four or more colors since any chordless E-cycle on at most three colors is triangulated in $G'(S)$. We first show every color in C appears uniquely. Suppose otherwise and let a be the color that appears the most often in C with f_a the number of times a appears. We consider the following cases.

Case I. $f_a = 3$, i.e., all three states a_0, a_1, and a_2 appear in C.

If there is an edge $e = (u, v)$ in C that does not have any of a_0, a_1, or a_2 as endpoints, then consider the row r that witnesses edge e; row r must contain some state of a, say a_i. This implies edges (u, a_i) and (v, a_i) are present in $G'(S)$ and C is not chordless, a contradiction. Therefore, in this case, every edge e in C must have exactly one endpoint of color a.

Since C contains four or more colors and every edge is adjacent to a state of a, by possibly renaming the character states, the color pattern must be as shown in Figure 27 (with distinct colors b, c, and d). Now, since C has length at least five and color b appears uniquely in cycle C, the b-complete graph $I(C, b_0)$ induced by C and b_0 exists. However, the vertices adjacent to b_0 in C are the same color (both having color a), which is forbidden by Lemma 8. It follows that $f_a < 3$.

Case II. $f_a = 2$: let a_0 and a_1 be the two states of a appearing in C. Since C contains four or more colors, it must be the case that one of the paths from a_0 to a_1 has three or more edges. We have the following cases.

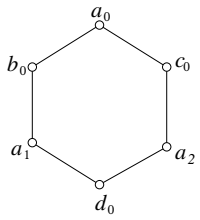


Fig. 27. Color pattern in Case I.

Case (IIa) both paths from a_0 to a_1 have at least three edges

Case (IIb) one path from a_0 to a_1 has two edges and the other path has three or more edges

Any edge that does not have color a as one of its endpoints must be witnessed by a row that contains the third state a_2, as illustrated in Figure 28.

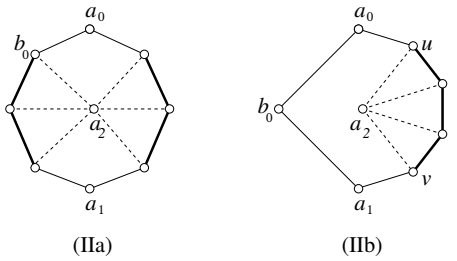


Fig. 28. Cases (IIa) and (IIb) in the proof of Lemma 9. The rows witnessing the edges shown in bold must contain state a_2 in character a.

In case (IIa), the second edge in both paths from a_0 to a_1 are witnessed by state a_2 and we obtain an E-cycle that is chordless in $G'(S)$ on at most three colors (shown in bold in Figure 29(a)). This is a contradiction since all E-cycles on at most three colors must be triangulated in $G'(S)$. In case (IIb), the second and second to last edge on the a_0 to a_1 path with three or more edges are witnessed by color a_2. Let D denote the E-cycle of edges (b_0, a_0), (a_0, u), (u, a_2), (a_2, v), (v, a_1), (a_1, b_0) (shown in Figure 29(b)). Then a_2 and b_0 are not adjacent in $G'(S)$ (otherwise, we would obtain a cycle of length four on at most three colors with at most one F-edge). This implies D is an E-cycle that is chordless in $G(S)$; in this cycle, b_0 has two adjacent vertices of color a and therefore cannot be the only state of b appearing in D, by Lemma 8. This implies one of u or v must also have color b and therefore D is a chordless cycle on at most three colors containing all three states of character a, contradicting Lemma 1.

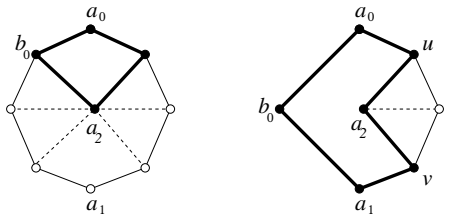


Fig. 29. Case II.

Case III. $f_a = 1$, i.e., every color in C appears uniquely. Suppose for a contradiction that C has length five or greater and let a_0 be a state appearing in C. Then $I(C, a_0)$ exists in $G'(S)$ by Observation 4. However, this gives a chordless cycle in $G'(S)$ with color a appearing two or more times, which cannot happen by Cases I and II.

It follows that C is a cycle of length four with all colors appearing uniquely in C, proving the lemma. □

Lemma 9 implies all chordless E-cycles in $G'(S)$ have length four containing four distinct colors. We have triangulated all such cycles by F'-edges in $G''(S)$, implying the following corollary.

Corollary 1. *$G''(S)$ cannot contain a chordless E-cycle.*

Lemmas 5, 6, 7, and Corollary 1 together imply that $G''(S)$ is properly triangulated, proving the main theorem.

Theorem 3 *Given an input set S on m characters with at most three states per character ($r = 3$), S admits a perfect phylogeny if and only if every subset of three characters of S admits a perfect phylogeny.*

5 Enumerating Obstruction Sets for Three State Characters

We now turn to the problem of enumerating all minimal obstruction sets to perfect phylogenies on three-state character input. By Theorem 3, it follows that the minimal obstruction sets are input sequences on at most three characters; we enumerate all instances S on three characters a, b, and c satisfying the following conditions:

(i) each character a, b and c has at most three states
(ii) every pair of characters allows a perfect phylogeny
(iii) the three characters a, b, and c together do not allow a perfect phylogeny.

Note that Condition (ii) implies the partition intersection graph $G(S)$ does not contain a cycle on exactly two colors and Condition (iii) implies $G(S)$ contains at least one chordless cycle. Let C be the largest chordless cycle in $G(S)$, i.e.,

$$C = \arg\max_{\text{chordless cycles } D \text{ in } G(S)} |D|$$

Condition (ii) and Lemma 1 together imply C cannot contain all three states of any character. Therefore, C has length at most six. If $G(S)$ contains a chordless six-cycle C, then each color appears exactly twice in C and C must have one of the color patterns shown in Figure 30.

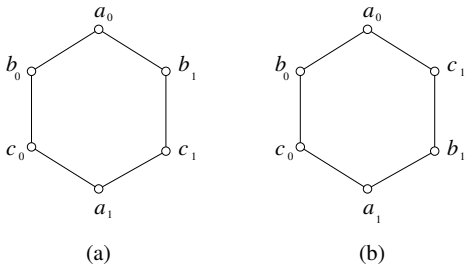


Fig. 30. Color patterns for chordless cycle of length six.

In Figures 30(a) and 30(b), there is one state in each character that does not appear in C (states a_2, b_2, and c_2). Since C is chordless, the witness for each edge is forced to contain the missing state in the third character. This implies Figure 30(a) must be completed by the edges in Figure 31(a) and Figure 30(b) must be completed by the edges in Figure 32(a). In both cases, there is a cycle on two characters a and b (see Figures 31(b) and 32(b)). This implies the pair of characters a and b is not properly triangulatable, a contradiction to condition (ii). Therefore, $G(S)$ cannot contain chordless cycles of length six.

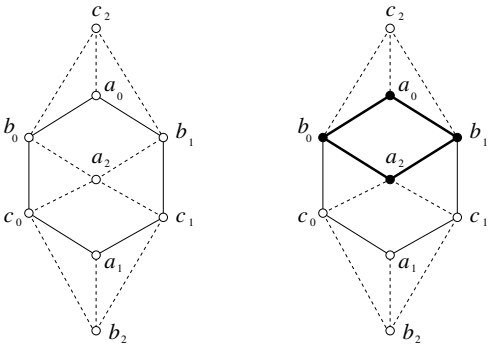


Fig. 31. Forced patterns for row witnesses of Figure 30(a).

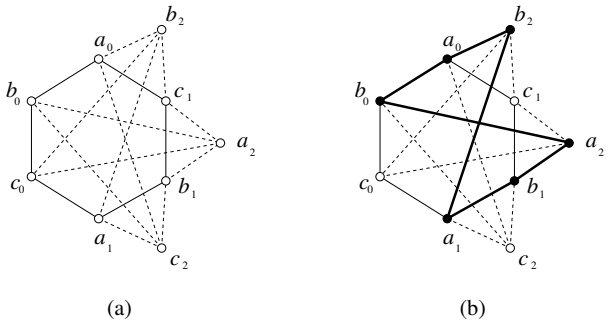


Fig. 32. Forced patterns for row witnesses of Figure 30(b).

If C is a chordless cycle in $G(S)$ of length five, then $G(S)$ is not properly triangulatable by Lemma 3, implying Condition (iii) is satisfied. In this case, there must be two characters (say b and c) appearing in two different states and one character appearing once in C, as shown in Figure 33 (up to relabeling of the states). Cycle C contains three edges that are not adjacent to character a (edges (b_0, c_0), (c_0, b_1), (b_1, c_1) in Figure 33). The row witnesses for these edges must contain either state a_1 or a_2 in character a.

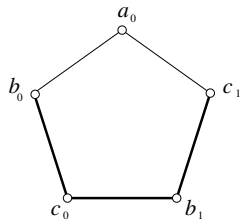
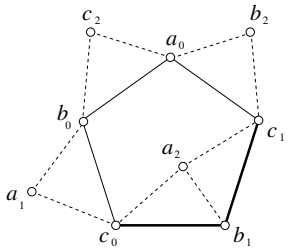


Fig. 33. Color pattern for cycle C of length five.

Case I. The row witnesses for two adjacent edges share the same state of a and the row witness for the third edge contains the final state in a. Without loss of generality, assume (c_0, b_1) and (b_1, c_1) are the two adjacent edges sharing the same state of a. In this case, $G(S)$ and the corresponding input sequences S are shown in Figure 34 (up to relabeling of the states).

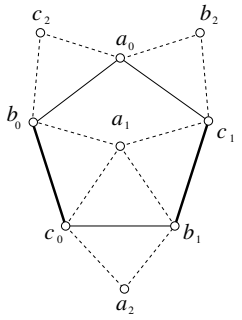


a	b	c
0	0	2
0	2	1
1	0	0
2	1	0
2	1	1

Fig. 34. Case I. Row witnesses for two adjacent edges share the same state of a.

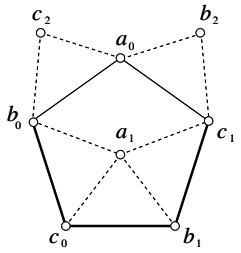
Case II. The row witnesses for the two nonadjacent edges share the same state of a and the row witness for the third edge contains the final state in a. In this case, $G(S)$ and the corresponding input sequences S are shown in Figure 35 (up to relabeling of the states).

Case III. The row witnesses for all three edges share the same state of a. In this case, $G(S)$ and the corresponding input sequences S are shown in Figure 36 (up to relabeling of the states).



a	b	c
0	0	2
0	2	1
1	0	0
2	1	0
1	1	1

Fig. 35. Case II. Row witnesses for two nonadjacent edges share the same state of a.



a	b	c
0	0	2
0	2	1
1	0	0
1	1	0
1	1	1

Fig. 36. Case III. Row witnesses for all three edges share the same state of a.

If C is a chordless cycle of length four, then without loss of generality it must have the color pattern shown in Figure 37.

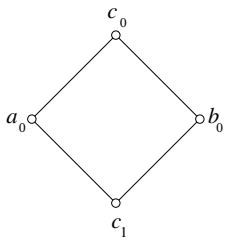


Fig. 37. Color Pattern for chordless cycle of length four.

Consider the row witnesses for edges (a_0, c_0) and (a_0, c_1). These row witnesses cannot share the same state of b (otherwise, there would be a cycle on two colors b and c, a contradiction). Similarly, row witnesses for edges (b_0, c_0) and (b_0, c_1) cannot share the same state of a. Therefore, up to relabeling of the states, the row witnesses are forced to have the pattern shown in Figure 39.

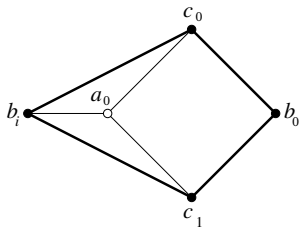


Fig. 38. Row witnesses for edges (a_0, c_0) and (a_0, c_1) cannot share the same state of b.

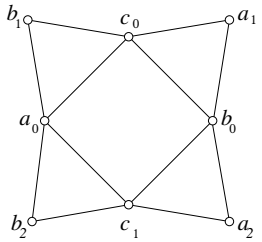


Fig. 39. Forced pattern of row witnesses.

Note that b_2 and c_0 cannot be adjacent in $G(S)$; otherwise, there is a cycle on two colors b and c (see Figure 40(a)). By symmetry, we can argue the following pairs are also nonadjacent (see Figure 40(b)).

The pairs (b_2, c_0), (a_2, c_0), (b_1, c_1), and (a_1, c_1) are nonadjacent in $G(S)$. (*)

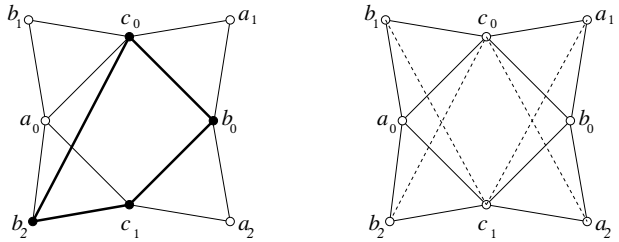
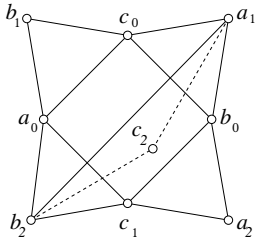


Fig. 40. Cycle on two colors b and c.

Now, suppose b_2 and a_1 are adjacent in $G(S)$. Then the row witness for (b_2, a_1) cannot be c_0 and cannot be c_1 by (*). Therefore, the row witness for this edge must be the third state c_2 of character c (see Figure 41). The partition intersection graph $G(S)$ and corresponding input sequences S are shown in Figure 41. Note that $G(S)$ is not properly triangulatable and condition (iii) is satisfied, since the edge (a_0, b_0) is a forced edge to triangulate cycle C, creating a cycle on two colors (a_0, b_0), (b_0, a_1), (a_1, b_2), (b_2, a_0) which cannot be properly triangulated.



a	*b*	*c*
0	1	0
1	0	0
2	0	1
0	2	1
1	2	2

Fig. 41. Input sequences S and partition intersection graph $G(S)$ with a chordless cycle of length four.

If b_2 and a_2 are adjacent in $G(S)$, then this induces a chordless cycle D of length five (b_2, a_0), (a_0, c_0), (c_0, b_0), (b_0, a_2), (a_2, b_2) (the pairs (b_2, c_0) and (a_2, c_0)

are nonadjacent by (*) and (a_0, b_0) are nonadjacent since they are nonadjacent vertices in chordless cycle C). This is a contradiction since C is chosen to be the largest chordless cycle in $G(S)$.

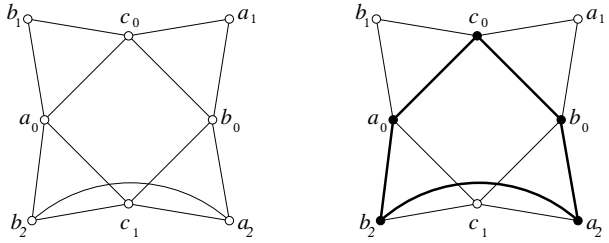


Fig. 42. If b_2 and a_2 are adjacent in $G(S)$, this creates a chordless cycle of length five.

Suppose there are no further adjacencies between vertices in Figure 39; then there must be additional edges formed by the final state c_2 of character c in order for $G[a, b, c]$ to be nontriangulatable (in order to satisfy condition (iii)). Now, state c_2 is adjacent to one or more of the edges with color pattern (a, b). If c_2 is adjacent to exactly one such edge, then the resulting graph $G[a, b, c]$ can be properly triangulated by adding the edge (a_0, b_0). Otherwise, state c_2 is adjacent to two or more edges. If the two edges share a vertex (i.e., the two edges are either (a_1, b_0) and (a_2, b_0) or (b_1, a_0) and (b_2, a_0)), then there is a cycle on two colors (as shown in Figure 43(a) and 43(b)), contradicting condition (ii).

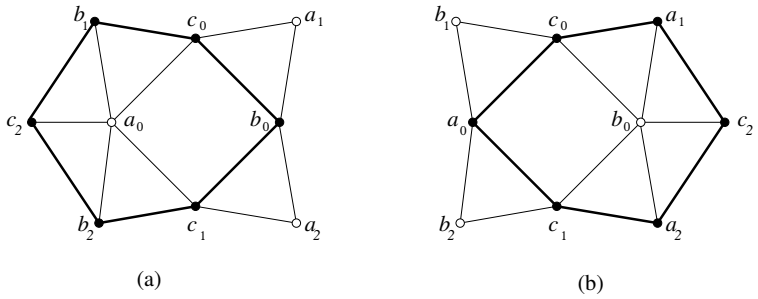


Fig. 43. If c_2 witnesses two adjacent edges in $G(S)$, this creates a chordless cycle on two colors.

Else if state c_2 is adjacent to two nonadjacent edges in $G(S)$ (Figure 44(a) and 44(b)), then this again creates a chordless cycle on two colors as shown in Figure 44(c), contradicting condition (ii).

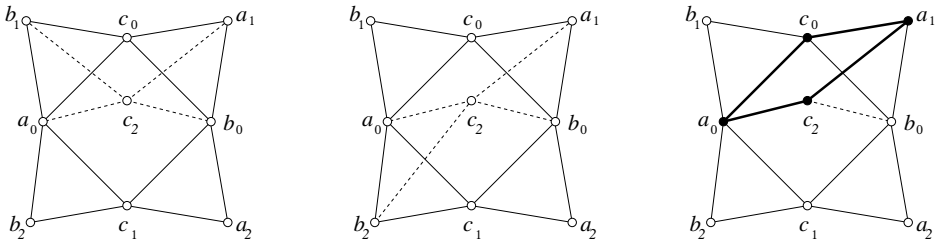


Fig. 44. If c_2 witnesses two nonadjacent edges in $G(S)$, this creates a chordless cycle on two colors.

In summary, the following are the minimal obstruction sets to the existence of perfect phylogenies for three-state characters up to relabeling of the character states.

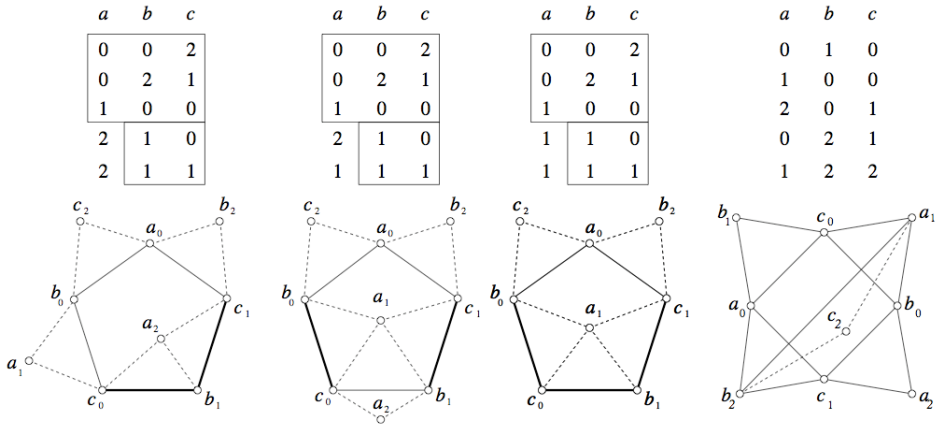


Fig. 45. Minimal obstruction sets for three-state characters up to relabeling.

6 Construction of Fitch-Meacham Examples

In this section, we examine in detail the class of Fitch-Meacham examples, which were first introduced by Fitch [13,14] and later generalized by Meacham [27]. The

goal of these examples is to demonstrate a lower bound on the number of characters that must be simultaneously examined in any test for perfect phylogeny. The natural conjecture generalizing our main result is that for any r, there is a perfect phylogeny on r-state characters if and only if there is one for every subset of r characters. We show here that such a result would be the best possible, for any r. While the general construction of these examples and the resulting lower bounds were stated by Meacham [27], to the best of our knowledge, the proof of correctness for these lower bounds has not been established. We fill this gap by explicitly describing the complete construction for the entire class of Fitch-Meacham examples and providing a proof for the lower bound claimed in [27].

For each integer r ($r \geq 2$), the Fitch-Meacham construction F_r is a set of $r+2$ sequences over r characters, where each character takes r states. We describe the construction of the partition intersection graph $G(F_r)$; the set of sequences F_r can be obtained from $G(F_r)$ in a straightforward manner, with each taxon corresponding to an r-clique in $G(F_r)$.

Label the r characters in F_r by $0, 1, \ldots r-1$; each vertex labeled by i will correspond to a state in character i. The construction starts with two cliques EC_1 and EC_2 of size r, called end-cliques, with the vertices of each clique labeled by $0, 1, \ldots r-1$. The vertex labeled i in EC_1 is adjacent to the vertex labeled $(i+1) \mod r$ in EC_2. For each such edge $(i, (i+1) \mod r)$ between the two end-cliques, we create a clique of size $r-2$ with vertices labeled by $\{0, 1, \ldots r-1\} \backslash \{i, (i+1) \mod r\}$. Every vertex in this $(r-2)$-clique is then attached to both i (in end-clique 1) and $(i+1) \mod r$ (in end-clique 2), creating an r-clique whose vertices are labeled with integers $0, 1, \ldots r-1$. There are a total of r such cliques, called *tower-cliques*, and denoted by TC_1, TC_2, $\ldots \mathrm{TC}_r$. Note that for each i ($0 \leq i \leq r-1$), there are exactly r vertices labeled by i; we give each such vertex a distinct state, resulting in r states for each character.

Note that the graph corresponding to the four gamete obstruction set is an instance of the Fitch-Meacham construction with $r = 2$. In this case, the four binary sequences $00, 01, 10, 11$ have two states, two colors and four taxa and the partition intersection graph for these sequences is precisely the graph $G(F_2)$. Note that in this case, every subset of $r-1 = 1$ characters has a perfect phylogeny, while the entire set of characters does not. Similarly, the fourth graph shown in Figure 45 illustrating the obstruction set for 3-state input is the graph $G(F_3)$ corresponding to the Fitch-Meacham construction for $r = 3$ (in the figure, $\mathrm{EC}_1 = \{a_0, b_2, c_1\}$ and $\mathrm{EC}_2 = \{a_1, b_0, c_0\}$). As shown in Section 5, every $r-1 = 2$ set of characters in the corresponding input set allows a perfect phylogeny while the entire set of characters does not. The following theorem generalizes this property to the entire class of Fitch-Meacham examples. Because the theorem was stated without proof in [27], we provide a proof of the result here.

Theorem 5. *[27] For every $r \geq 2$, F_r is a set of input sequences over r state characters such that every $r-1$ subset of characters allows a perfect phylogeny while the entire set F_r does not allow a perfect phylogeny.*

Proof. We first show that $G(F_r)$ does not allow a proper triangulation for any r. As observed above, $G(F_2)$ is a four cycle on two characters and therefore, does not allow a proper triangulation (since any proper triangulation for a graph containing cycles must have at least three colors). Suppose $G(F_r)$ is properly triangulatable for some $r \geq 3$, let s be the smallest integer such that $G(F_s)$ has a proper triangulation, and let $G'(F_s)$ be a minimal proper triangulation of $G(F_s)$.

For each tower-clique TC_i in $G(F_s)$, consider the set of vertices in TC_i that are not contained in either end-clique; call these vertices *internal tower-clique vertices* and the remaining two tower vertices *end tower-clique vertices*. Note that the removal of the two end tower-clique vertices disconnects the internal tower-clique vertices from the rest of the graph. This implies that the internal tower-clique vertices cannot be part of any chordless cycle: otherwise, such a chordless cycle C must contain *both* end tower-clique vertices i and $(i+1) \mod s$. However, the two end tower-clique vertices are connected by an edge and therefore induce a chord in C, a contradiction since C is a chordless cycle.

In the graph $G(F_s)$, onsider the following cycle of length four: $s-2$ (in EC_1) $\rightarrow s-1$ (in EC_1) $\rightarrow 0$ (in EC_2) $\rightarrow s-1$ (in EC_2) $\rightarrow s-2$ (in EC_1). This four-cycle has a unique proper triangulation, which forces the edge e between vertex $s-2$ in EC_1 and vertex 0 in EC_2 to be included in $G'(F_s)$. Consider removing all vertices labeled $s-1$ from $G'(F_s)$, and for the two vertices labeled $s-1$ in end-cliques EC_1 and EC_2, remove all interior tower-clique vertices (but not end tower-clique vertices) adjacent to $s-1$. Then edge e between vertices $s-2$ and 0 is still present and we can expand e into a tower-clique of size $s-1$ (by forming a clique with new vertices $1, 2, \ldots s-3$ adjacent to both $s-2$ and 0 of the two end-cliques).

In the resulting graph, the vertices are exactly those of $G(F_{s-1})$ and all edges in $G(F_{s-1})$ are present. Furthermore, if there is a chordless cycle in this graph, then it would create a chordless cycle in $G'(F_s)$ since no internal tower-clique vertex can be part of any chordless cycle (and in particular, the new vertices $1, 2, \ldots s-3$ cannot be part of any chordless cycle). Therefore, the resulting graph is a proper triangulation for $G(F_{s-1})$, a contradiction since s was chosen to be the smallest integer such that $G(F_s)$ allows a proper triangulation.

To prove the second part of the theorem, we show that in F_r, any subset of $r-1$ characters does allow a perfect phylogeny by proving that the partition intersection graph on any subset of $r-1$ characters has a proper triangulation. By the symmetry of the construction of F_r, we can assume without loss of generality that the $r-1$ characters under consideration are $\{0, 1, \ldots r-2\}$. Consider the graph obtained by connecting every vertex i $(0 \leq i \leq r-3)$ in EC_1 to every vertex j satisfying $j > i$ in EC_2. Note the asymmetry between the first and second end-cliques in this construction and observe that none of the added edges are between characters with the same label.

Suppose the resulting graph contains a chordless cycle C. Then C cannot contain three or more vertices in either end-clique and cannot contain any internal tower-clique vertices (as noted earlier), so must have length exactly four

with two vertices in each end-clique. It cannot be the case that two nonadjacent vertices of C are in the same end-clique, since these vertices would be adjacent and C would not be chordless. Therefore, cycle C must be formed as follows: i (in EC_1) $\rightarrow j$ (in EC_2) $\rightarrow j'$ (in EC_2) $\rightarrow i'$ (in EC_1). Since i and j are adjacent, we have $i < j$ and since i' and j' are adjacent, we have $i' < j'$. If $i < j'$, then i and j' are adjacent and the cycle C is not chordless, a contradiction. Therefore, $i' < j' \leq i < j$, which implies i' and j are adjacent and the cycle C is not chordless, again a contradiction. It follows that there are no chordless cycles and the added edges form a proper triangulation for the partition intersection graph on the subset of $r-1$ characters $\{0, 1, \ldots r-2\}$. □

7 Conclusion

We have studied the structure of the three state perfect phylogeny problem and shown that there is a necessary and sufficient condition for the existence of a perfect phylogeny for three state characters using triples of characters. This extends the extremely useful Splits Equivalence Theorem and four-gamete condition. The obvious extension of our work would be to discover similar results for r-state characters for $r \geq 4$.

Until this work, the notion of a conflict, or incompatibility, graph has been defined for two state characters only (using the four gamete condition). Our generalization of the four gamete condition therefore allows us to generalize this notion to incompatibility on three state characters. The resulting incompatibility structure will be a hypergraph with edges corresponding to pairs and triples of characters that do not allow a perfect phylogeny. These hypergraphs can be used to solve algorithmic and theoretical problems for three state characters analogous to those for binary characters. Examples of such problems include the minimum character removal problem and the problem of finding lower bounds on the minimum number of recombinations for three state characters.

In addition, there are several theoretical and practical results known for two state characters that are still open for characters on three or more states. For instance, it is known that the problem of constructing near-perfect phylogenies for two state characters is fixed parameter tractable; the analogous problem is open for characters on three or more states. Similarly, the question of finding decomposition theorems for recurrent mutations and recombinations remains open for three or more states. With the recent increase in collection of polymorphism data such as micro/mini-satellites, there is a need for the analysis of perfect phylogenies to be extended to multiple state characters. Our work lays a solid theoretical foundation we hope will help with this effort.

Acknowledgements The authors gratefully acknowledge G. Blelloch, R. Ravi, R. Schwartz, and T. Warnow for stimulating discussions and suggestions. This research was partially supported by NSF grants SEI-BIO 0513910, CCF-0515378, and IIS-0803564.

References

1. V. Bafna and V. Bansal. Improved recombination lower bounds for haplotype data. *In proceedings of Research in Computational Molecular Biology (RECOMB)*, 2005.
2. V. Bafna and V. Bansal. The number of recombination events in a sample history: Conflict graph and lower bounds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:78–90, Apr-Jun, 2004.
3. V. Bafna, D. Gusfield, G. Hannenhalli, and S. Yooseph. A note on efficient computation of haplotypes via perfect phylogeny. *Journal of Computational Biology*, 11:858–866, 2004.
4. V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. *Journal of Computational Biology*, 10:323–340, 2003.
5. G. E. Blelloch, K. Dhamdhere, E. Halperin, R. Ravi, R. Schwartz, and S. Sridhar. Fixed parameter tractability of binary near-perfect phylogenetic tree reconstruction. *International Colloquium on Automata, Languages and Programming*, 2006.
6. H. Bodlaender and T. Kloks. A simple linear time algorithm for triangulating three-colored graphs. *J. Algorithms*, 15(1):160–172, 1993.
7. P. Bonizzoni. A linear-time algorithm for the perfect phylogeny haplotype problem. *Algorithmica*, 48:267–285, 2007.
8. P. Buneman. A characterization of rigid circuit graphs. *Discrete Math*, 9:205–212, 1974.
9. Z. Ding, V. Filkov, and D. Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping problem. *J. of Computational Biology*, 13:522–553, 2006.
10. E. Eskin, E. Halperin, and R. M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, pages 1–20, 2003.
11. G. Estabrook, C. Johnson, and F. McMorris. A mathematical formulation for the analysis of cladistic character compatibility. *Math Bioscience*, 29, 1976.
12. J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass., 2004.
13. W.M. Fitch. Toward finding the tree of maximum parsimony. *The Eighth International Conference on Numerical Taxonomy (Estabrook, G. F., ed.), San Francisco: W. H. Freeman and Company*, pages 189–220, 1975.
14. W.M. Fitch. On the problem of discovering the most parsimonious tree. *American Naturalist*, 11:223–257, 1977.
15. D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21:19–28, 1991.
16. D. Gusfield. Haplotyping as a perfect phylogeny: Conceptual framework and efficient solutions. *Research in Computational Molecular Biology*, 2002.
17. D. Gusfield. Optimal, efficient reconstruction of Root-Unknown phylogenetic networks with constrained and structured recombination. *JCSS*, 70:381–398, 2005.
18. D. Gusfield and V. Bansal. A fundamental decomposition theory for phylogenetic networks and incompatible characters. *Research in Computational Molecular Biology*, 2005.
19. D. Gusfield, V. Bansal, V. Bafna, and Y. Song. A decomposition theory for phylogenetic networks and incompatible characters. *Journal of Computational Biology*, 14(10):1247–1272, Dec 2007.
20. D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinformatics and Computational Biology*, 2(1):173–213, 2004.

21. 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, Special issue on Computational Biology, 2007*, 155:806–830, 2007.
22. E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 2004.
23. R. Hudson and N. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.
24. D. Huson, T. Klopper, P. J. Lockhart, and M. A. Steel. Reconstruction of reticulate networks from gene trees. *Research in Computational Molecular Biology*, 2005.
25. R. M. Idury and A. A. Schäffer. Triangulating three-colored graphs in linear time and linear space. *SIAM J. Discret. Math.*, 6(2):289–293, 1993.
26. S. Kannan and T. Warnow. Triangulating three-colored graphs. *SODA '91: Proceedings of the second annual ACM-SIAM symposium on Discrete algorithms*, pages 337–343, 1991.
27. C. Meacham. Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. *Nato ASI series vol G1 on Numerical Taxonomy, Springer Verlag*, 1983.
28. R. V. Satya and A. Mukherjee. An optimal algorithm for perfect phylogeny haplotyping. *Journal of Computational Biology*, 13:897–928, 2006.
29. R.V. Satya, A. Mukherjee, G. Alexe, L. Parida, and G. Bhanot. Constructing near-perfect phylogenies with multiple homoplasy events. *Bioinformatics*, 22:e514–i522, 2006. Bioinformatics Suppl., Proceedings of ISMB 2006.
30. C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.
31. S. Sridhar, G. E. Blelloch, R. Ravi, and R. Schwartz. Optimal imperfect phylogeny reconstruction and haplotyping. *In proceedings of Computational Systems Bioinformatics (CSB)*, 2006.
32. S. Sridhar, K. Dhamdhere, G. E. Blelloch, E. Halperin, R. Ravi, and R. Schwartz. Simple reconstruction of binary near-perfect phylogenetic trees. *International Workshop on Bioinformatics Research and Applications*, 2006.
33. S. Sridhar, K. Dhamdhere, G. E. Blelloch, E. Halperin, R. Ravi, and R. Schwartz. Algorithms for efficient near-perfect phylogenetic tree reconstruction in theory and practice. *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, 2007.