2009

Diversity Graphs

P Blain

C Davis

Allen G. Holder
Trinity University, aholder@trinity.edu

J Silva

C Vinzant

Diversity Graphs

P. Blain[a], C. Davis[b], A. Holder[c,*], J. Silva[d] and C. Vinzant[e]

October 26, 2006

Abstract

Bipartite graphs have long been used to study and model matching problems, and in this paper we introduce the bipartite graphs that explain a recent matching problem in computational biology. The problem is to match haplotypes to genotypes in a way that minimizes the number of haplotypes, a problem called the Pure Parsimony problem. The goal of this work is not to address the computational or biological issues but rather to explore the mathematical structure through a study of the underlying graph theory.

Keywords: Diversity Graph, Graph Theory, Haplotype, Optimization, Parsimony, Partially Ordered Sets

---

[a] Swarthmore College Mathematics, Swarthmore, PA, pblain1@swarthmore.edu

[b] University of Utah Mathematics, Salt Lake, UT, davis@math.utah.ed

[c] Trinity University Mathematics, San Antonio, TX, aholder@trinity.edu

[d] University of Colorado Applied Mathematics, Denver, CO, jsilva2105@msn.com

[e] Oberlin College Mathematics, Oberlin, OH, cvinzant@oberlin.edu

[*] Corresponding Author.

– Research conducted at Trinity University, San Antonio, TX, with partial support of the National Science Foundation, grant DMS-0353488.

1 Introduction

The burgeoning field of computational biology is advancing the science of genetics and transforming traditional 'wet lab' research into computational efforts. The preponderance of the current research emphasizes computational aspects, which have made significant strides in projects such as the human genome project. These advances have the potential of redefining standard medical practice and have already proven to be a significant contribution to mankind.

One of the problems currently receiving attention is that of describing how genetic diversity propagates from one generation to the next. Such problems are called *haplotyping* problems, and a brief biological description is warranted. Genes are sequences of DNA that code for specific traits, with the vast majority of DNA being common among all individuals. The locations on the genome where diversity occurs are called single nucleotide polymorphisms (SNPs). Diploid organisms like humans have two distinct copies of each gene, one from each parent, which together describe a trait. A collection of SNPs in a single copy of a gene is called a *haplotype*, and a pair of haplotypes forms a *genotype*. Each SNP of a haplotype is in one of two states, denoted by -1 or 1, that corresponds to the two distinct nucleotide base pairs of the DNA. Each SNP of a genotype is in one of three states, -2, 0, or 2, where the SNP is -2 (resp, 2) if and only if each of the haplotypes that form the genotype have a -1 (resp. 1) at that SNP, and the SNP is 0 if and only if one of the haplotypes has a -1 and the other a 1 at that SNP.

Biologists are capable of efficiently determining an individual's genotype, but it is difficult and costly to determine the haplotypes. However, haplotypes are more valuable to biologists, and a haplotyping problem is to calculate the haplotypes knowing only the genotypes. In particular, finding small collections of haplotypes that explain the genotypes is biologically relevant. The problem of finding a smallest collection of haplotypes is called the *Pure Parsimony* problem and empirical evidence suggests that these minimum solutions naturally occur. The initial investigations into haplotyping were undertaken by Clark in [5], and since then there has been flourish of activity addressing computational issues [2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Theoretically we know that the parsimony problem is APX-hard and that practically it is difficult to solve on large data sets. Our goal is not directly computational, and instead we address the underlying structure of the problem through graph theory. We do solve the pure parsimony problem in a few well-defined instances.

2 Notation, Definitions and Preliminary Results

The results of this paper are graph theoretical, and a basic understanding of bipartite graph theory is expected. We point readers to [1] for a thorough development. The degree and neighborhood of node v are denoted by $\deg(v)$ and $N(v)$, respectively. The vector of ones is denoted by e, where length is decided by the context of its use. If x is a vector, then $\mathrm{diag}(x)$ is the symmetric matrix whose diagonal elements correspond to x and whose off-diagonal elements are zero. So, $\mathrm{diag}(e)$ is the identity matrix. For any real number C we define $C_+ = \max\{C, 0\}$.

We assume that haplotypes are of length n and that SNPs are indexed by $i = 1, 2, \ldots, n$. The set of all possible haplotypes of length n is the collection of sequences $\mathcal{H} = \{-1, 1\}^n$. Similarly, the collection of all genotypes of length n is $\{-2, 0, 2\}^n$. The arithmetic of *mating* haplotypes to form a genotype is simply coordinate-wise addition. So, if the maternal haplotype is $(-1, 1, -1, 1)$ and the fraternal haplotype is $(1, 1, -1, -1)$, the genotype is

$$(-1, 1, -1, 1) + (1, 1, -1, -1) = (0, 2, -2, 0). \tag{1}$$

We extend this coordinatewise addition so that we can generally add elements in $\{-2,-1,0,1,2\}^n$. Let $a,b \in \{-2,-1,0,1,2\}$ and define $\oplus$ so that

$$a \oplus b = \begin{cases} -2, & a<0, b<0 \\ 2, & a>0, b>0 \\ 0, & \text{otherwise.} \end{cases}$$

This binary operator is commutative, and to add elements of $\{-2,-1,0,1,2\}^n$ we perform the operation componentwise. For example,

$$(-1,1,-1,1) \oplus (1,0,-2,1) \oplus (1,-1,-1,1) = (0,0,-2,2).$$

Notice that $\oplus$ reduces to the typical addition in (1) if the terms on the left are in $\{-1,1\}^4$. Unfortunately, $\oplus$ does not generally satisfy a cancellation rule since $a \oplus 0 = b \oplus 0$ does not mean that $a = b$. For much of the paper simple addition as described in (1) is sufficient, and to distinguish $\oplus$ from $+$ we use $+$ in all instances where it appropriate.

SNP values of 0 are called *ambiguous* because the orientation of the 1 and -1 in the parental donations could be reversed. The question we address begins with a collection of genotypes and asks us to construct a collection of haplotypes that form the genotypes under this arithmetic. If there were no ambiguous SNPs, this process would be trivial, and hence, we assume that each genotype has at least one ambiguous SNP. The subset of $\{-2,0,2\}^n$ with this property is denoted $\mathcal{G}$. For any $\mathcal{G}' \subseteq \mathcal{G}$, we say $\mathcal{H}' \subseteq \mathcal{H}$ is a *solution* to $\mathcal{G}'$ if, for all $\mathbf{g} \in \mathcal{G}'$, there exist $\mathbf{h}', \mathbf{h}'' \in \mathcal{H}$ such that $\mathbf{g} = \mathbf{h}' + \mathbf{h}''$. A solution $\mathcal{H}$ to $\mathcal{G}'$ is *minimal* if $\mathcal{H} \backslash \{\mathbf{h}\}$ is not a solution to $\mathcal{G}'$, for all $\mathbf{h} \in \mathcal{H}$. We say $\mathcal{H}$ is a *minimum solution* if there exists no solution $\mathcal{H}'$ to $\mathcal{G}'$ such that $|\mathcal{H}'| < |\mathcal{H}|$.

Our intent is to study the underlying graph theory of finding solutions, and we introduce the concept of a *Diversity Graph*. Informally, a diversity graph is a labeled (or colored) bipartite graph with one set of nodes representing genotypes, the other set representing haplotypes, and edges representing the possible relationships between them.

Definition 2.1. *For $\mathcal{H}' \subset \mathcal{H}$ and $\mathcal{G}' \subset \mathcal{G}$, a bipartite graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$, and functions $\eta : \mathcal{V} \to \mathcal{H}'$ and $\gamma : \mathcal{W} \to \mathcal{G}'$, we say $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph on n SNPs if*

1. *η and γ are one-to-one,*
2. *for each $w \in \mathcal{W}$, there exists some $v \in \mathcal{V}$ such that $(v,w) \in \mathcal{E}$, and*
3. *$\mathcal{E}$ has the property that if $(v',w) \in \mathcal{E}$, there exists some $v'' \in \mathcal{V} \backslash \{v'\}$ such that $(v'',w) \in \mathcal{E}$ and $\mathbf{h}' + \mathbf{h}'' = \mathbf{g}$, where $\mathbf{h}' = \eta(v')$, $\mathbf{h}'' = \eta(v'')$, and $\mathbf{g} = \gamma(w)$.*

The requirement that η and γ be one-to-one ensures that each haplotype and genotype are represented by exactly one node. The rest of the definition guarantees that $\mathcal{H}'$ is a solution to $\mathcal{G}'$. If $\eta(v') + \eta(v'') = \gamma(w)$, we say that v' and v'' or $\eta(v')$ and $\eta(v'')$ are *mates* for w or $\gamma(w)$. Notice that a diversity graph is a labeled bipartite graph, and we make the distinction between the structure represented by the graph and the biology represented by the labeling. The elements of $\mathcal{H}$ and $\mathcal{G}$ are denoted by $\mathbf{h}$ and $\mathbf{g}$ or by $\eta(v)$ and $\gamma(w)$, where v and w are elements of $\mathcal{V}$ and $\mathcal{W}$. Different elements of $\mathcal{H}$ and $\mathcal{G}$ are indicated with superscripts and SNP locations are indicated with subscripts. We say that a bipartite graph *supports diversity* if there are sets $\mathcal{H}' \subseteq \mathcal{H}$ and $\mathcal{G}' \subseteq \mathcal{G}$, and functions η and γ that fulfill the definition.

An important observation is that the pure parsimony problem as stated assumes that $\mathcal{G}'$ is known, that $\eta(V) = \mathcal{H}$ and that $\mathcal{E}$ is as large as possible. However, the parsimony problem makes sense on other graphs, and in general we address the problem of starting with a diversity graph $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ and finding a smallest subset of $\mathcal{V}$, say $\mathcal{V}'$, such that

- $\eta(\mathcal{V}')$ is a solution to $\gamma(\mathcal{W}) = \mathcal{G}'$, and
- if v' and v'' are in $\mathcal{V}'$, then they are allowed to mate and form w if and only if (v', w) and (v'', w) are in $\mathcal{E}$.

So, when we say that $\mathcal{H}$ is a minimal or minimum solution we mean that it is a solution with respect to a diversity graph. If $\mathcal{V}$ and η are such that $\eta(\mathcal{V}) = \mathcal{H}$ and $\mathcal{E}$ is as large as possible, then we are considering the original parsimony problem.

Before continuing with an investigation into the bipartite graphs that support diversity, we establish some general results about diversity graphs. The first of these results shows how to order the elements of $\mathcal{H}$ so that we can conveniently pair them to form the genotype $(0, 0, \ldots, 0)$. The imposed ordering is *lexicographic*, meaning that $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n) < (\mathbf{h}'_1, \mathbf{h}'_2, \ldots, \mathbf{h}'_n)$ if the first component with different values satisfies $\mathbf{h}_i < \mathbf{h}'_i$. The proof of this lemma is simple and omitted.

Lemma 2.1. *If the elements of $\mathcal{H}$ are ordered lexicographically, then for unique i and j between 1 and 2^n we have that $h^i + h^{i+1} \neq h^j + h^{j+1}$ and that $h^j + h^{(2^n - j+1)} = (0, 0, \ldots, 0)$.*

Since haplotypes mate in unique pairs to form a genotype, the degree of every node in $\mathcal{V}$ is even. An immediate consequence of this observation is that not every bipartite graph supports diversity. Moreover, even if every node of a bipartite graph has even degree it does not mean the graph supports diversity. As an example, the complete bipartite graph $K_{2,2}$ does not support diversity because any η and γ that satisfies the third and fourth conditions of the definition violates the fact that γ is one-to-one. Theorem 2.1 does not characterize the graphs that support diversity, but it does establish a necessary condition.

Theorem 2.1. *Let $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ be a diversity graph. Let w^1 and w^2 be in $\mathcal{W}$, with $w^1 \neq w^2$. Then,*

$$\max\{\deg(w^1), \deg(w^2)\} \geq 2|N(w^1) \cap N(w^2)|.$$

Proof. Let $\mathcal{H}' = \{\eta(v) : v \in N(w^1) \cap N(w^2)\}$ and assume to the contrary that

$$\max\{\deg(w^1), \deg(w^2)\} < 2|\mathcal{H}'|.$$

Since haplotypes mate in unique pairs, there must be fewer than $|\mathcal{H}'|$ pairs of haplotypes mating to form each of $\gamma(w^1) = \mathbf{g}^1$ and $\gamma(w^2) = \mathbf{g}^2$. It follows that there exists $\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3$ and $\mathbf{h}^4$ in $\mathcal{H}'$ such that $\mathbf{h}^1 + \mathbf{h}^2 = \mathbf{g}^1$ and $\mathbf{h}^3 + \mathbf{h}^4 = \mathbf{g}^2$. Suppose that $\mathbf{g}^1_j = 2$, which means $\mathbf{h}_j = 1$ for all $\mathbf{h} \in \mathcal{H}'$. It follows that $\mathbf{g}^2_j = 2$. Similarly, if $\mathbf{g}^1_j = -2$, we have that $\mathbf{g}^2_j = -2$. Suppose that $\mathbf{g}^1_j = 0$. Then $\mathbf{g}^2_j$ can not be 2 or -2 because if so, the same argument would guarantee that $\mathbf{g}^1_j$ is 2 or -2, respectively. We conclude that $\mathbf{g}^2_j = 0$. Since j was arbitrary, we have the contradiction that $\mathbf{g}^1 = \mathbf{g}^2$. □

We now turn our direction to a matrix equation that is satisfied by every diversity graph. Consider the diversity graph $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$, where $|\mathcal{V}| = m$ and $|\mathcal{W}| = k$. List the elements of $\mathcal{V}$ as $v^1, v^2, \ldots, v^m$ and $\mathcal{W}$ as $w^1, w^2, \ldots, w^k$, and assume that $\eta(v^i) = \mathbf{h}^i$ and $\gamma(w^i) = \mathbf{g}^i$. Let H be the $m \times n$ matrix so that $H_{(i,j)} = h^i_j$, and let G be the $k \times n$ matrix defined by $G_{(i,j)} = g^i_j$. Also let E be the $m \times k$ biadjacency matrix —i.e. $E_{(i,j)} = 1$ if $(v^i, w^j) \in \mathcal{E}$ and $E_{(i,j)} = 0$ otherwise. Note that the column sums of E must be even from the definition of a diversity graph.

The $k \times n$ matrix product $E^T H$ aggregates the mating structure for each genotype. Without loss of generality, let $E_{(1,i)} = E_{(2,i)} = \ldots = E_{(t,i)} = 1$ and $E_{(t+1,i)} = E_{(t+2,i)} = \ldots = E_{(m,i)} = 0$. Then the ith row of $E^T H$ is $\eta(v^1) + \eta(v^2) + \ldots + \eta(v^t)$. From the definition of a diversity graph we know there are $t/2$ disjoint pairs, (v^p, v^q), with p and q no greater than t, such that $\eta(v^p) + \eta(v^q) = \gamma(w^i)$. This means that the ith row of $E^T H$ is $(t/2)\gamma(w^i)$. We have just established the following result.

Theorem 2.2. *If $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph, then $E^T H = diag\left(\frac{1}{2}E^T e\right) G$.*

The matrix equation in Theorem 2.2 succinctly separates the structure of the graph, explained by E, from the labeling of the graph, explained by H and G. Unfortunately, satisfying the matrix equation does not guarantee the graph is a diversity graph because the aggregated information ignores the need of a mating structure. As an example

$$\begin{aligned} E^T H &= \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix} \\ &= (2) \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \text{diag}\left(\frac{1}{2}E^T e\right) G. \end{aligned}$$

This labeling of $K_{4,1}$ does not lead to a diversity graph since no pair of haplotypes (no two rows of H) add to form the single genotype (the row of G).

We conclude this section with a discussion of a logical operator that helps address the failure of Theorem 2.2 to characterize graphs with the stated matrix equation. The *logical join* of a sequence of matrices is determined by the logical operator "or" over each component of these matrices. The component-wise logical join is defined so that $0 \vee 0 = 0$, $0 \vee 1 = 1$, and $1 \vee 1 = 1$. The set $\{A^1, A^2, \ldots, A^s\}$ is a *logical decomposition* of A if A is the logical join of the matrices in this set, denoted:

$$\bigvee_{1 \le i \le s} A^i = A^1 \vee A^2 \vee \ldots \vee A^s = A,$$

where we assume that all matrix elements are 0 or 1. For example, the matrices on the left are a logical decomposition of the matrix on the right,

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \vee \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Such decompositions are used in the next section to characterize the graphs that support diversity.

3 Graphs that Support Diversity

A natural goal is to characterize the bipartite graphs that support diversity, and the first result of this section does this for complete bipartite graphs.

Theorem 3.1. *The complete graph $K_{p,q}$ supports diversity if and only if p is even and $q = 1$.*

Proof. Assume that $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a complete diversity graph. Suppose that $|\mathcal{W}| > 1$. Then, from Theorem 2.1 we have for any w^1 and w^2 in $\mathcal{W}$ that

$$\max\{\deg(w^1), \deg(w^2)\} \ge 2|N(w^1) \cap N(w^2)| = 2|\mathcal{V}|,$$

which is a contradiction. So, $|\mathcal{W}| = 1$. The fact that genotypes need pairs of haplotypes guarantees that $|\mathcal{V}|$ is even.

Now assume that p is even and $q = 1$. Select n so that $|\mathcal{V}| < 2^n$ and let γ be such that $\gamma(w) = (0, 0, \ldots, 0)$. There are 2^{n-1} disjoint pairs of haplotypes that can mate to form $\gamma(w)$. Pick $|\mathcal{V}|/2$ of these pairs and let $\mathcal{H}'$ be the set of haplotypes in these pairs. Allowing $\eta : \mathcal{V} \to \mathcal{H}'$ to be a bijection, we see that $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph. □

From Theorem 3.1 we see that the probability of generating a complete bipartite graph of the form $K_{p,1}$ that supports diversity is one half (assuming that even and odd values of p are equally likely). In some ways, the next result extends this idea by showing that the probability of generating a random bipartite graph that supports diversity is low. The result decomposes the biadjacency matrix into matrices whose rows sums are all 2, which guarantees a mating structure.

Theorem 3.2. *The bipartite graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ supports diversity if and only if the biadjacency matrix E has a logical decomposition $E^1, E^2, \ldots, E^s$ so that*

- $e^T E^k = 2e^T$ *for all* $1 \leq k \leq s$,
- *there exists an* $H \in \{-1, 1\}^{|\mathcal{V}| \times n}$ *with distinct rows and the property that* $(E^1)^T H = (E^2)^T H = \ldots = (E^s)^T H$, *and*
- *the rows of* $(E^k)^T H$ *are distinct.*

Proof. Assume that $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph, and let $s = (1/2)\max\{\deg(w) : w \in \mathcal{W}\}$. Order the elements of $\mathcal{W}$ so that $\deg(w^1) \geq \deg(w^2) \geq \ldots \geq \deg(w^{|\mathcal{W}|})$. We construct the matrices $E^1, E^2, \ldots, E^s$ that form a desired logical decomposition. The neighborhood of each w^j can be written as the disjoint union of $(1/2)\deg(w^j)$ pairs,

$$N(w^j) = \bigcup_k \{v^{k_{j'}}, v^{k_{j''}}\}, \tag{2}$$

where $1 \leq k \leq (1/2)\deg(w^j)$. Let every element of the first column of each E^k be a zero except for the $k_{1'}$ and $k_{1''}$ positions, which are both set to 1. If the $\deg(w^1) = \deg(w^2)$, form the second column of each E^k similarly, replacing $1'$ and $1''$ with $2'$ and $2''$. Otherwise, $\deg(w^1) > \deg(w^2)$ and this construction terminates once k reaches $(1/2)\deg(w^2)$. In this case, let the second column of E^k, for $(1/2)\deg(w^2)+1 \leq k \leq s$, be the same as the second column of E^1. Continue in this fashion through the remaining nodes in $\mathcal{W}$, duplicating columns from E^1 as needed. From (2) we have that $E_{(i,j)} = 1$ if and only if $E^k_{(i,j)} = 1$ for at least one k, from which we conclude that $E = E^1 \vee E^2 \vee \ldots \vee E^s$.

Let H and G be the matrices in Theorem 2.2. Each column of E^k corresponds to an element of $\mathcal{W}$ that in turn corresponds to a genotype under γ. Moreover, each column of E^k contains two 1s that identify a pair of haplotypes (via η) that mate to form the genotype. So, each column sum of every E^k sums to 2 and $(E^k)^T H = G$, for every k. From the fact that γ is one-to-one we have that the rows of $(E^k)^T H$ are distinct.

Now assume that E has a logical decomposition, say $E^1, E^2, \ldots E^s$, that satisfies the three conditions. List the elements of $\mathcal{V}$ from 1 to $|\mathcal{V}|$ and define $\eta(v^i)$ to be the i^{th} row of H. The fact that the rows of H are unique ensures that η is one-to-one. Similarly, list the nodes in $\mathcal{W}$ from 1 to $|\mathcal{W}|$ and let $\gamma(w^i)$ be the i^{th} row of $(E^k)^T H$, which is common for $1 \leq k \leq s$. The assumption that the rows of $(E^k)^T H$ are distinct guarantees that γ is one-to-one. From the condition that each column sum of E^k is 2, we have that each column of E has at least two ones. This means that there are at least two elements of V that are adjacent to each element of W. The same condition together with the definition of η and γ further guarantee that if $(v', w) \in \mathcal{E}$, then there is a v'' so that $(v'', w) \in \mathcal{E}$ and $\eta(v') + \eta(v'') = \gamma(w)$. □

The logical decomposition in Theorem 3.2 extends Theorem 2.2 to characterize graphs that support diversity by adding the necessary condition that the edge structure must accommodate a mating structure. The fact that $(E^1)^T H = (E^2)^T H = \ldots = (E^s)^T H$ shows that every pairwise differences $(E^i)^T - (E^j)^T$ must share a non-trivial null space, which is restrictive and demonstrates that bipartite graphs that support diversity are rare.

The last two results indicate that the structural requirements needed to support diversity are important and that most bipartite graphs can not be labeled to represent a population. We point out that this is true even with the number of SNPs being arbitrary, which is somewhat counter intuitive because the complexity of a mating scheme can increase as the number of SNPs grows. We conclude this section by showing that we can add nodes and edges to any bipartite graph so that it does support diversity. We only consider adding nodes to $\mathcal{V}$ since in real problems the genotypic information corresponding to $\mathcal{W}$ is defined by the biological data.

For $w \in W$, define

$$T(w) = \bigcup_{w' \neq w} \left(N(w) \cap N(w')\right).$$

So, $T(w)$ is the collection of nodes in the neighborhood of w that are also in the neighborhood of another node in W. We extend the neighborhood of each w so that the number of points in $N(w)\backslash T(w)$ plus the number of points in the extension is at least the number of points in $T(w)$. Let $\hat{\mathcal{V}}(w)$ be a collection of nodes whose cardinality is either $(2|T(w)| - |N(w)|)_+$ or $1 + (2|T(w)| - |N(w)|)_+$ to ensure that $|N(w) \cup \hat{\mathcal{V}}(w)|$ is even or 0. Notice that if $2|T(w)| \leq |N(w)|$, then $N(w)$ is not extended. The extended vertex and edge sets are

$$\bar{\mathcal{V}} = \mathcal{V} \cup \left(\bigcup_{w \in W} \hat{\mathcal{V}}(w) \right) \quad \text{and} \quad \bar{\mathcal{E}} = \mathcal{E} \cup \left(\bigcup_{w \in \mathcal{W}} \{(v, w) : v \in \hat{\mathcal{V}}(w)\} \right).$$

Lemma 3.1. *Any bipartite graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ with no isolated nodes can be extended and labeled to become a diversity graph by adding no more than $\sum_{w \in \mathcal{W}} |\hat{\mathcal{V}}(w)|$ nodes to $\mathcal{V}$. In particular, $(\bar{\mathcal{V}}, \mathcal{W}, \bar{\mathcal{E}})$ is an extension of $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ that adds this number of nodes to $\mathcal{V}$ that supports diversity.*

Proof. The proof follows by induction on $|\mathcal{W}|$. Let $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a bipartite graph with no isolated nodes such that $|\mathcal{W}| = 1$. Let $W = \{w\}$ and notice that $T(w) = \emptyset$. Hence, $(2|T(w)| - |N(w)|)_+ = 0$, and we add a single node to $\mathcal{V}$ if and only if $|\mathcal{V}|$ is odd. The resulting $\bar{\mathcal{V}}$ has an even number of nodes, and from Theorem 3.1 we know that that this graph supports diversity.

Assume the result holds if $|\mathcal{W}| \leq k$. Let $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a bipartite graph with no isolated nodes such that $|\mathcal{W}| = k + 1$. Select $w^1 \in W$, and let $(\mathcal{V}', \mathcal{W}', \mathcal{E}')$ be the subgraph of $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ with the vertices in $\{w^1\} \cup N(w^1)\backslash T(w^1)$ and edges incident to w^1 removed. Extend the subgraph so that $(\bar{\mathcal{V}}', \mathcal{W}', \bar{\mathcal{E}}', \eta', \gamma')$ is a diversity graph, where the image sets of η' and γ' are in $\{-1, 1\}^{n'}$ and $\{-2, 0, 2\}^{n'}$, respectively. The functional descriptions of η and γ below depend on η' and γ', and a slight abuse of notation is used to describe this dependence. As an example, if $\eta'(v) = (1, -1, 1)$, we assume that $(\eta'(v), 1, 1, 1) = (1, -1, 1, 1, 1, 1)$, which allows us to embed η' into a larger collection of haplotypes.

The argument is established in 2 cases, each of which constructs η and γ so that $(\bar{\mathcal{V}}, \mathcal{W}, \bar{\mathcal{E}}, \eta, \gamma)$ is a diversity graph. Notice that the number of nodes added to $\mathcal{V}$ is additive over $\mathcal{W}$, and hence,

$$\sum_{w \in \mathcal{W}} |\hat{\mathcal{V}}(w)| = |\hat{\mathcal{V}}(w^1)| + \sum_{w \in \mathcal{W}'} |\hat{\mathcal{V}}(w)|.$$

This fact guarantees that the constructions below add the maximum number of vertices allowed by the result.

Case 1: Suppose that $T(w^1) = \emptyset$. Then, $(2|T(w^1)| - |N(w^1)|)_+ = 0$, and $|\hat{\mathcal{V}}(w^1)|$ is either 0 or 1 depending on whether $|N(w^1)|$ is even or odd, respectively. If $|N(w^1)|$ is even (odd), then no nodes are (a single node is) added to $(\mathcal{V}', \mathcal{W}, \mathcal{E}')$. Let $|N(w^1) \cup \hat{\mathcal{V}}(w^1)| = 2p$ for the natural number p. Let k be a natural number so that $2^k > 2p$. List the elements of $\{-1, 1\}^k$

lexicographically as $\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^{2^k}$. Then, denoting the elements of $N(w^1) \cup \hat{\mathcal{V}}(w^1)$ as v^i for $i = 1, 2, \ldots, 2p$, we define η and γ by

$$\eta : \bar{\mathcal{V}} \to \{-1,1\}^{n'+k} : v \mapsto \begin{cases} (\eta'(v), 1, 1, \ldots, 1), & v \in \mathcal{V}' \\ (1, 1, \ldots, 1, \mathbf{h}^{i+1}), & v = v^i, i = 1, 2, \ldots p \\ (1, 1, \ldots, 1, \mathbf{h}^{2^k - i + p}), & v = v^i, i = p+1, p+2, \ldots 2p \end{cases}$$

and

$$\gamma : \mathcal{W} \to \{-2, 0, 2\}^{n'+k} : w \mapsto \begin{cases} (\gamma'(v), 2, 2, \ldots, 2), & w \in \mathcal{W}' \\ (2, 2, \ldots 2, 0, 0, \ldots 0), & w = w^1, \end{cases}$$

where $\gamma(w^1)$ has k zeros. We mention that Lemma 2.1 is used to guarantee that the last k elements of $\eta(v^i)$, for $i = 1, 2, \ldots, 2p$, can be paired to satisfy the definition of a diversity graph.

Case 2: Suppose $T(w^1) \neq \emptyset$. The difficulty with this case lies in the fact that $\eta'(v)$ is defined for $v \in T(w^1)$. Notice that

$$(|N(w^1)| - 2|T(w^1)|) + (2|T(w^1)| - |N(w^1)|)_+ = (|N(w^1)| - 2|T(w^1)|)_+ \geq 0,$$

which guarantees that there are enough nodes in $(N(w^1) \cup \hat{\mathcal{V}}(w^1)) \backslash T(w^1)$ to be uniquely paired with the nodes in $T(w^1)$. Let $\{Z, Z^C\}$ be a two set partition of $(N(w^1) \cup \hat{\mathcal{V}}(w^1)) \backslash T(w^1)$ so that $|Z| = |T(w^1)|$. Notice that the definition of $\hat{\mathcal{V}}(w^1)$ guarantees that both $|T(w^1) \cup Z|$ and $|Z^C|$ are even.

List the elements of $T(w^1)$, Z and Z^C so that

$$T(w^1) = \{v^1, v^2, \ldots, v^{|T(w^1)|}\}, \tag{3}$$

$$Z = \{v^{|T(w^1)|+1}, v^{|T(w^1)|+2}, \ldots, v^{2|T(w^1)|}\}, \text{ and} \tag{4}$$

$$Z^C = \{v^{2|T(w^1)|+1}, v^{2|T(w^1)|+2}, \ldots, v^{|\hat{\mathcal{V}}(w^1)|}\}. \tag{5}$$

Let $|N(w^1) \cup \hat{\mathcal{V}}(w^1)| = 2p$ for the natural number p and let k be such that $2^k > p$. Label $\{-1,1\}^k$ lexicographically as $\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^{2^k}$. Define $\eta : \bar{\mathcal{V}} \to \{-1,1\}^{n'+k}$ so that

$$v \mapsto \begin{cases} (\eta'(v), 1, 1, \ldots, 1), & v \in \mathcal{V}' \backslash T(w^1) \\ (\eta'(v^i), \mathbf{h}^{i+1}), & v = v^i, 1 \leq i \leq |T(w^1)| \\ (-\eta'(v^{i-|T(w^1)|}), \mathbf{h}^{2^k - i + |T(w^1)|}), & v = v^i, |T(w^1)| + 1 \leq i \leq 2|T(w^1)| \\ (\eta'(v^i), \mathbf{h}^{i-|T(w^1)|}), & v = v^i, 2|T(w^1)| + 1 \leq i \leq |\hat{\mathcal{V}}(w^1)|, i \text{ odd} \\ (-\eta'(v^{i-1}), \mathbf{h}^{2^k - i + 2|T(w^1)|}), & v = v^i, 2|T(w^1)| + 1 \leq i \leq |\hat{\mathcal{V}}(w^1)|, i \text{ even} \end{cases}$$

and $\gamma : \mathcal{W} \to \{-2, 0, 2\}^{n'+k}$ so that

$$w \mapsto \begin{cases} (\gamma'(w), 2, 2, \ldots, 2), & w \in \mathcal{W}' \\ (0, 0, \ldots, 0), & w = w^1. \end{cases}$$

□

We conclude this section by showing that any bipartite graph, including those with isolated nodes, can be extended and labeled to become a diversity graph. The result is an extension of Lemma 3.1, but the edges added between isolated nodes are handled outside the definition of $\bar{\mathcal{E}}$.

Theorem 3.3. *Any bipartite graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ can be extended and labeled to become a diversity graph by adding no more than*

$$\sum_{w \in \mathcal{W}} |\hat{\mathcal{V}}(w)| + (2M_{\mathcal{W}} - M_{\mathcal{V}})_+ + M_{\mathcal{V}}(\bmod 2)$$

nodes to $\mathcal{V}$, where $M_{\mathcal{V}}$ and $M_{\mathcal{W}}$ are the number of isolated nodes in $\mathcal{V}$ and $\mathcal{W}$, respectively.

Proof. Let $\mathcal{V}_I$ and $\mathcal{W}_I$ be the isolated nodes in $\mathcal{V}$ and $\mathcal{W}$ and let $(\mathcal{V}', \mathcal{W}', \mathcal{E}')$ be the subgraph of $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ with these nodes removed. Extend $(\mathcal{V}', \mathcal{W}', \mathcal{E}')$ as in Lemma3.1 so that $(\bar{\mathcal{V}}', \mathcal{W}', \bar{\mathcal{E}}', \eta', \gamma')$ is a diversity graph. Since $N(w) = T(w) = \emptyset$ for all $w \in \mathcal{W}_I$, we have that

$$\sum_{w \in \mathcal{W}'} |\hat{\mathcal{V}}(w)| = \sum_{w \in \mathcal{W}} |\hat{\mathcal{V}}(w)|,$$

and we conclude that the extension of $(\mathcal{V}', \mathcal{W}', \mathcal{E}')$ to $(\bar{\mathcal{V}}', \mathcal{W}', \bar{\mathcal{E}}')$ adds $\sum_{w \in \mathcal{W}} |\hat{\mathcal{V}}(w)|$ nodes to $(\mathcal{V}, \mathcal{W}, \mathcal{E})$. Let n' be such that the images of η' and γ' are in $\{-1, 1\}^{n'}$ and $\{-2, 0, 2\}^{n'}$. Let k be a natural number so that $2^k > |\mathcal{W}_I|$. Let $\hat{\mathcal{V}}_I$ be a set of nodes of size $(2M_{\mathcal{W}} - M_{\mathcal{V}})_+ + M_{\mathcal{V}}(\bmod 2)$, which guarantees that $|\mathcal{V}_I \cup \hat{\mathcal{V}}_I|$ is even and at least twice the size of $\mathcal{W}_I$. List the elements in $\mathcal{W}_I$ as $w^1, w^2, \ldots, w^{M_W}$ and the elements in $\mathcal{V}_I \cup \hat{\mathcal{V}}_I$ as $v^1, v^2, \ldots, v^q$, where $q = M_V + (2M_{\mathcal{W}} - M_{\mathcal{V}})_+ + M_{\mathcal{V}}(\bmod 2)$. For $i = 1, 2, \ldots, M_W$, add (v^{2i-1}, w^i) and (v^{2i}, w^i) to $\bar{\mathcal{E}}'$. Notice that this may leave some isolated nodes in $\mathcal{V}_I$, which is allowed by the definition. Let $\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^{2^k}$ be a lexicographic ordering of $\{-1, 1\}^k$. Define $\eta : \bar{\mathcal{V}}' \cup \mathcal{V}_I \cup \hat{\mathcal{V}}_I \to \{-1, 1\}^{n'+k}$ by

$$v \mapsto \begin{cases} (\eta'(v), 1, 1, \ldots, 1), & v \in \bar{\mathcal{V}}' \\ (1, 1, \ldots, 1, \mathbf{h}^{i+1}), & v = v^i, i = 1, 2, \ldots, q/2 \\ (1, 1, \ldots, 1, \mathbf{h}^{2^k - i + (q/2)}), & v = v^i, i = q/2 + 1, q/2 + 2, \ldots, q \end{cases}$$

and $\gamma : \mathcal{W} \to \{-2, 0, 2\}^{n'+k}$ so that

$$w \mapsto \begin{cases} (\gamma', 2, 2, \ldots, 2), & w \in \mathcal{W}' \\ (2, 2, \ldots, 2, \eta(v^{2i-1}) + \eta(v^{2i})), & w = w^i, i = 1, 2, \ldots, M_W \in \mathcal{W}_I, \end{cases}$$

where the uniqueness of $\eta(v^{2i-1}) + \eta(v^{2i})$ follows from Lemma 2.1. □

4 Algorithms and Solutions for the Pure Parsimony Problem

Although the pure parsimony is generally difficult, there are cases where a closed form solution exists. Throughout this section we assume that $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph with the property that γ maps $\mathcal{W}$ onto $\mathcal{G}'$. We also assume that $\mathcal{H}^* \subseteq \eta(\mathcal{V})$ is a minimal solution relative to $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$.

We begin by establishing the intuitive result that a minimal solution has cardinality $2|\mathcal{G}'|$ if and only if the neighborhoods of the elements in $\mathcal{W}$ are disjoint. Although this fact is nearly obvious, we include a proof for completeness. The following lemma supports the result.

Lemma 4.1. *Suppose that $T(w) \neq \emptyset$ for some $w \in \mathcal{W}$. Then, $\mathcal{H}^*$ contains an element of $\bigcup_{w \in \mathcal{W}} \eta(T(w))$.*

Proof. Assume that $T(w) \neq \emptyset$ for some $w \in \mathcal{W}$ and suppose that $\mathcal{H}^*$ does not contain an element of $\bigcup_{w \in \mathcal{W}} \eta(T(w))$. Then for each w there is a v' and v'' in $N(w) \setminus \bigcup_{w \in \mathcal{W}} \eta(T(w))$ so that $\eta(v') + \eta(v'') = \gamma(w)$. This implies that $|\mathcal{H}^*| = 2|\mathcal{G}'|$. However, we know that $T(w)$ is nonempty for some w, which

means there exists w^1 and w^2 such that $\eta(v^1)+\eta(v^2)=\gamma(w^1)$ and $\eta(v^1)+\eta(v^3)=\gamma(w^2)$ for some v^1, v^2, and v^3 in $\mathcal{V}$. This means that

$$\big(\mathcal{H}^*\backslash\{\eta(v):\eta(v)\in\mathcal{H}^*,\{(v,w^1),(v,w^2)\}\cap\mathcal{E}\neq\emptyset\}\big)\cup\{\eta(v^1),\eta(v^2),\eta(v^3)\}$$

is a solution to $\mathcal{G}'$ whose cardinality is at most $|\mathcal{H}^*|-1$, which is a contradiction. □

Theorem 4.1. *We have that $|\mathcal{H}^*|=2|\mathcal{G}'|$ if and only if $N(w')\cup N(w'')=\emptyset$ for all w' and w'' in $\mathcal{W}$.*

Proof. The fact that $|\mathcal{H}^*|=2|\mathcal{G}'|$ if $N(w')\cup N(w'')=\emptyset$ for all w' and w'' in $\mathcal{W}$ is clear. Assume that $|\mathcal{H}^*|=2|\mathcal{G}'|$, and suppose for the sake of obtaining a contradiction that $T(w)\neq\emptyset$ for some $w\in\mathcal{W}$. From Lemma 4.1 we have that $\mathcal{H}^*$ contains an element in $\bigcup_{w\in\mathcal{W}}\eta(T(w))$. Let w^1 and w^2 be such that $\eta(v^1)+\eta(v^2)=\gamma(w^1)$ and $\eta(v^1)+\eta(v^3)=\gamma(w^2)$, for some v^1, v^2, and v^3. Let $\mathcal{W}'=\mathcal{W}\backslash\{w^1,w^2\}$, and let $\mathcal{V}'=\cup_{w\in\mathcal{W}'}N(w)$. Furthermore, let $(\mathcal{V}')^*$ be such that $\eta((\mathcal{V}')^*)$ is a minimum solution to $\gamma(\mathcal{W}')$ with respect to $(\mathcal{V}',\mathcal{W}',\mathcal{E}')$, where $\mathcal{E}'$ is $\mathcal{E}$ with the edges incident to w^1 and w^2 removed. Then, $|\eta((\mathcal{V}')^*)|\leq 2|\mathcal{G}'|$. We know that we can resolve $\mathcal{G}'$ by including v^1, v^2, and v^3 in $(H')^*$. Since all three haplotypes might not be required, we have that $2|\mathcal{G}'|=|\mathcal{H}^*|\leq|(\mathcal{H}')^*|+3$. So,

$$2|\mathcal{G}'|=|\mathcal{H}^*|\leq|(\mathcal{V}')^*|+3\leq 2|\mathcal{W}'|+3=2(|\mathcal{W}|-2)+3=2|\mathcal{G}'|-1.$$

Since this is a contradiction, we have that $T(w)=\emptyset$ for all w, and consequently, $N(w')\cap N(w'')=\emptyset$, for all $w'\neq w''$. □

We continue our investigation by exploring the effects of restricting the number of times a haplotype can be used to form a genotype. This makes sense realistically since in many populations the mating structure is not random. For example, many species have a unique mate for life, which means their haplotypes are only used in conjunction with the haplotypes of another individual. To make this precise, we reduce the edge set of the initial graph. For any $\mathcal{E}'\subseteq\mathcal{E}$ we define the degree of v with respect to $\mathcal{E}'$ to be $\deg_{\mathcal{E}'}(v)=|\{(v,w):(v,w)\in\mathcal{E}'\}|$. For the diversity graph $(\mathcal{V},\mathcal{W},\mathcal{E},\eta,\gamma)$ we let $\mathcal{V}_m^*\subseteq\mathcal{V}$ be any solution to

$$\min\{|\mathcal{V}'|:\mathcal{V}'\subseteq\mathcal{V},\eta(\mathcal{V}')\text{ solves }\gamma(\mathcal{W}),\max\{\deg_{\mathcal{E}'}(v)\leq m:v\in\mathcal{V}'\}\text{ for some }\mathcal{E}'\subseteq\mathcal{E}\}. \tag{6}$$

The value of this optimization problem is denoted $\phi(m)=|\mathcal{V}_m^*|$, and if the problem is infeasible, we let $\phi(m)=\infty$. As an example, consider the graph in Figure 1, which is easily seen to support diversity. Since $\deg(w^i)=2$ for all i except 3, the only solution is $\eta(\{v^i:i=1,2,\ldots,9\})$. If $m=1$, then each v can be adjacent to at most one w with respect to $\mathcal{E}'$. This means we must be able to associate a unique pair in $\mathcal{V}$ with each element of $\mathcal{W}$. Biologically this means that each parent can donate one of its two haplotypes to a unique child. Since this is impossible for this graph, we have that $\phi(1)=\infty$. Notice that in general we have $\phi(1)$ is either $2|\mathcal{W}|$ or ∞ depending on whether or not (6) is feasible. The situation is more complex if $m>1$, and one of the main goals of this section is to show that $\phi(2)$ can be calculated by decomposing an acyclic diversity graph into longest paths.

At some threshold, increasing m does not change the cardinality of $\mathcal{V}_m^*$. For instance, if a haplotype is not compatible with more than m genotypes, then allowing it to mate with $m+1$ haplotypes provides no additional benefit. Hence, for some m, $\phi(m)=\phi(m+k)$ for every natural number k. Moreover, increasing the number of possible mates that any haplotype is allowed never causes an increase in $\phi(m)$, and hence, ϕ is non-increasing. The smallest m such that $\phi(m)=\phi(m+k)$, for all $k\in\mathbb{N}$, is denoted by m^*. Clearly we have that

$$m^*\leq\max\{\deg(v):v\in\mathcal{V}\}\leq|\mathcal{W}|.$$

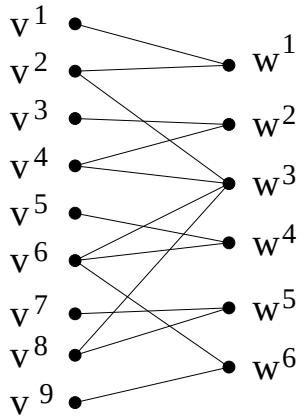

Figure 1: A graph for which $\mathcal{V}_1^* = \emptyset$, $\phi(1) = \infty$, $\mathcal{V}_2^* = \mathcal{V}$, $\phi(2) = 9$, and $m^* = 2$.

An important observation is that $\phi(m^*)$ is the solution to the pure parsimony problem. So, if we knew how ϕ grew as m increased and how to bound m^*, then we could estimate the size of a biologically relevant collection of haplotypes. Unfortunately, we do not know the answer to either of these questions at this point, but these and related questions have future research promise. We initiate the investigation by studying $\phi(2)$ and $\phi(|\mathcal{W}|)$ if $m^* = |\mathcal{W}|$, the latter of which is addressed below.

Theorem 4.2. *If $m^* = |\mathcal{W}|$, then $\phi(m^*) = |\mathcal{W}| + 1$.*

Proof. Let $m^* = |\mathcal{W}|$. Then, there exists $v' \in \mathcal{V}_m^*$ such that for each $w^i \in \mathcal{W}$ there is a unique $v^i \in \mathcal{V}_m^* \backslash \{v'\}$ that satisfies $\eta(v') + \eta(v^i) = \eta(w^i)$. This means that $\phi(m^*) \geq |\mathcal{W}| + 1$, and since $\eta(\{v', v^1, v^2, \ldots, v^{|\mathcal{W}|}\})$ solves $\mathcal{G}'$, we conclude that $\phi(m^*) = |\mathcal{W}| + 1$. □

Our next goal is to calculate $\phi(2)$ for acyclic graphs. The key observation in this case is that the most complicated subgraphs induced by a solution are paths. To motivate this intuition, consider the diversity graph in Figure 2. Notice that both $\eta(\{v^1, v^3\})$ and $\eta(\{v^2, v^4\})$ are solutions to $\gamma(\{w^1\})$ but that the path v^1, w^1, v^3 has the advantage over v^2, w^1, v^4 since the single node v^5 can be appended to the path so that $\eta(\{v^1, v^3, v^5\})$ solves $\gamma(\{w^1, w^2\})$. If we had instead selected v^2, w^1, v^4, then both v^3 and v^5 would have been needed so that $\eta(\{v^2, v^3, v^4, v^5\})$ solved $\gamma(\{w^1, w^2\})$. It is clear in this example that $\phi(2) = 3$ and that $m^* = 2$. The intuition is that we want to decompose the graph into longest paths, a process explained by the algorithm in Table 1. The fact that this technique minimizes the number of paths is established in Theorem 4.3. The proof is by induction on $|\mathcal{W}|$ and relates $\phi(2)$ as defined on $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ to $\phi(2)$ as defined on one of its subgraphs.

Theorem 4.3. *Let $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ be an acyclic diversity graph. Then the algorithm in Table 1 calculates $\phi(2)$, and in particular, if k is the number of paths found by the algorithm, then $\phi(2) = |\mathcal{W}| + k$.*

Proof. If $|\mathcal{W}| = 1$, the algorithm in Table 1 clearly finds an optimal solution. Assume the result is true as long as $|\mathcal{W}| \leq q$. Let $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ be an acyclic diversity graph with $|\mathcal{W}| = q+1$. Apply the algorithm in Table 1 to $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ and let $P_1, P_2, ..., P_k$ be the paths in non-increasing length found by the algorithm. Denote the last path as $P_k = v^1, w^1, v^2, w^2, \ldots, w^r, v^{r+1}$. Let $\mathcal{W}' = \mathcal{W} \backslash \{w^r\}$ and $\mathcal{E}' = \mathcal{E} \backslash \{(w^r, v) : v \in N(w^r)\}$.

An Algorithm to Decompose the acyclic bipartite graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ into the Fewest Paths

Step 1: Set $k = 0$ and $(\mathcal{V}_k, \mathcal{W}_k, \mathcal{E}_k) = (\mathcal{V}, \mathcal{W}, \mathcal{E})$.

Step 2: Find the longest path in $(\mathcal{V}_k, \mathcal{W}_k, \mathcal{E}_k)$, say P_k. If no path exists, set $P_k = \emptyset$.

Step 3: If $P_k = \emptyset$, stop.

Step 4: Set $(\mathcal{V}_{k+1}, \mathcal{W}_{k+1}, \mathcal{E}_{k+1}) = (\mathcal{V}_k, \mathcal{W}_k, \mathcal{E}_k) \backslash P_v$.

Step 5: Increase k by 1.

Step 6: Go to Step 2.

Table 1: Theorem 4.3 shows that this algorithm calculates $\phi(2)$. The removal of the path in Step 4 means that all nodes and edges in P_k are removed.

Case 1: Suppose $P_k \neq v^1, w^1, v^2$. Then, the algorithm applied to $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$ finds the paths $P_1, P_2, ..., P_k'$, where $P_k' = v^1, w^1, v^2, w^2, \ldots, w^{r-1}, v^r$ —i.e. the last path is missing w^r and v^{r+1}. In this case, the algorithm terminates with k paths for both $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ and $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$. From the induction hypothesis we have that $\phi(2) = |\mathcal{W}'| + k$ for $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$. Let $\hat{\mathcal{V}}_2^*$ be a solution to (6) for the subgraph $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$. Then, $\hat{\mathcal{V}}_2^* \subseteq \mathcal{V}$, $|\hat{\mathcal{V}}_2^*| = |\mathcal{W}'| + k$, $\eta(\hat{\mathcal{V}}_2^*)$ solves $\gamma(\mathcal{W}')$, and $|N(w) \cap \hat{\mathcal{V}}_2^*| \leq 2$. Since $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ is acyclic we know that (v^1, w^r) and (v^r, w^r) are not both in $\mathcal{E}$. Moreover, w^r cannot be adjacent to any of the terminal nodes $P_1, P_2, \ldots, P_{v-1}$ since this would violate the fact that each of these is a longest path. We conclude that adding w^r back to $\mathcal{W}'$ forces $\phi(2)$ for $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ to be at least one greater than $\phi(2)$ for $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$. Notice that $\eta(\hat{\mathcal{V}}_2^* \cup \{v^{r+1}\})$ is a solution to $\gamma(\mathcal{W})$ that is feasible to (6) for $(\mathcal{V}, \mathcal{W}, \mathcal{E})$. Since

$$|\hat{\mathcal{V}}_w^* \cup \{v^{r+1}\}| = |\mathcal{W}'| + k + 1 = |\mathcal{W}| + k,$$

we have that $\phi(2)$ for $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ is $|\mathcal{W}| + k$.

Case 2: Suppose $P_k = v^1, w^1, v^2$ —i.e. $r = 1$. Then the algorithm applied to $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$ produces the paths $P_1, P_2, \ldots, P_{k-1}$, and we have that $\phi(2) = |\mathcal{W}'| + k - 1$ for $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$. Let $\hat{\mathcal{V}}_2^*$ be as in Case 1 with the cardinality condition replaced by $|\hat{\mathcal{V}}_2^*| = |\mathcal{W}'| + k - 1$. Notice that w^1 cannot be adjacent to any of the terminal nodes of $P_1, P_2, \ldots, P_{k-1}$, as this would violate the fact that these are longest paths. So, adding w^1 back to $\mathcal{W}'$ forces $\phi(2)$ for $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ to be at least 2 greater than $\phi(2)$ for $(\mathcal{V}, \mathcal{W}', \mathcal{E}')$. Since $\eta(\hat{\mathcal{V}}_2^* \cup \{v^1, v^2\})$ is a solution to $\gamma(\mathcal{W})$ that is feasible to (6) for $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ that additionally satisfies

$$|\hat{\mathcal{V}}_w^* \cup \{v^{r+1}\}| = |\mathcal{W}'| + (k-1) + 2 = |\mathcal{W}| + k,$$

we have that $\phi(2)$ for $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ is $|\mathcal{W}| + k$.

□

We mention that this proof does not readily extend to graphs with cycles. The problem is that cycles can share nodes, and hence the removal of a longest cycle can destroy other cycles. Although a proof currently alludes the authors, we suspect the following is true.

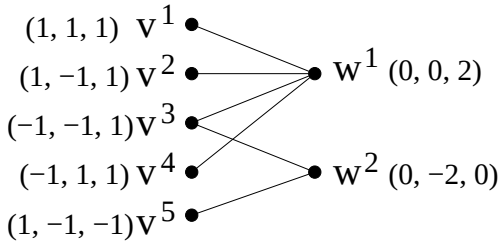

Figure 2: In this example $\eta(\{v^1, v^3, v^5\} = \mathcal{V}_2^*$ and $\phi(2) = 3$.

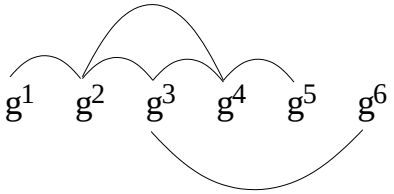

Figure 3: The consecutive genotypes of a path in $(\mathcal{H}, G, \mathcal{E})$ are indicated with an arc. So, there is path that contains the sequence g^2, h', g^4, h'', g^5, but there is no path that contains g^1, h, g^3.

Conjecture 4.1. *Let $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ be a diversity graph and $(\mathcal{V}', \mathcal{W}', \mathcal{E}')$ be the subgraph with all cycles removed. Then, if k is the number of paths identified by the algorithm in Table 1, we have that $\phi(2) = |\mathcal{W}| + k$.*

The insight from Theorem 4.3 is that the solutions of a restricted form of the Pure Parsimony problem are representable as a collection of paths. However, this technique has two shortcomings. First, the longest path problem is NP-Complete, making each step of the algorithm difficult. So, the technique describes the nature of a solution but does not provide an efficient solution procedure. The second shortcoming is that the algorithm is not capable of finding every optimal solution. To see this, consider the following collection of genotypes,

$$\left.\begin{array}{rcl} \mathbf{g}^1 = \gamma(w^1) & = & (2, 0, -2, -2, -2, -2) \\ \mathbf{g}^2 = \gamma(w^2) & = & (0, 2, 0, 0, -2, -2) \\ \mathbf{g}^3 = \gamma(w^3) & = & (-2, 0, 2, 0, -2, 0) \\ \mathbf{g}^4 = \gamma(w^4) & = & (-2, 0, 0, 2, 0, -2) \\ \mathbf{g}^5 = \gamma(w^5) & = & (-2, -2, -2, 0, 2, -2) \\ \mathbf{g}^6 = \gamma(w^6) & = & (-2, -2, 0, -2, -2, 2). \end{array}\right\} \tag{7}$$

Assume that $|\mathcal{V}| = 2^6$, that $\eta(\mathcal{V}) = \{-1, 1\}^6$, and that $\mathcal{E}$ is the largest edge set possible. Notice that a path may contain the sequence $\mathbf{g}^i, \mathbf{h}^i, \mathbf{g}^{i+1}$ if and only if there is no SNP where $\mathbf{g}^i$ has a value of 2 or -2 and $\mathbf{g}^{i+1}$ has the other value. So, in the above example there is no $\mathbf{h}$ such that the path $\mathbf{g}^1, \mathbf{h}, \mathbf{g}^3$ exists in the diversity graph because the first SNP of $\mathbf{g}^1$ is a 2 and the first SNP of $\mathbf{g}^3$ is a -2. However, there is an $\mathbf{h}$ so that the path $\mathbf{g}^1, \mathbf{h}, \mathbf{g}^2$, is in the diversity graph because there is no SNP where $\mathbf{g}^1$ and $\mathbf{g}^2$ have different values of 2 and -2. If we compare each pair of genotypes in a similar fashion, we find that the paths pass through the genotypes as indicated in Figure 3. From this figure we see that there is not a path or cycle through every genotype, but that there are several two path solutions. From Theorem 4.3 we know that $\phi(2) = 6 + 2 = 8$. Up to reversing the order of the genotypes, there are four optimal progressions through the genotypes, see Table 2. Our algorithm finds the first solution indicated in Table 2, as the first path is as long as possible. None of the other paths have this property, and so the algorithm is not capable of finding these solutions.

Our last discussion approaches the pure parsimony problem through lattice theory and requires the more general $\oplus$ as discussed in Section 2. Let $\preceq$ be a binary relation such that $2 \preceq 2$, $2 \preceq 0$, $-2 \preceq -2$, $-2 \preceq 0$, and $0 \preceq 0$. Then $\{-2, 0, 2\}^n$ is a poset under componentwise comparisons of $\preceq$. A subset of $\mathcal{G}'$ for which all elements are comparable forms a chain of genotypes. For example, the

First Path's Genotype Progression	Second Path's Genotype Progression
$(\mathbf{g}^1, \mathbf{g}^2, \mathbf{g}^3, \mathbf{g}^4, \mathbf{g}^5)$	$(\mathbf{g}^6)$
$(\mathbf{g}^1, \mathbf{g}^2, \mathbf{g}^4, \mathbf{g}^5)$	$(\mathbf{g}^3, \mathbf{g}^6)$
$(\mathbf{g}^1, \mathbf{g}^2, \mathbf{g}^3, \mathbf{g}^6)$	$(\mathbf{g}^4, \mathbf{g}^5)$
$(\mathbf{g}^6, \mathbf{g}^3, \mathbf{g}^4, \mathbf{g}^5)$	$(\mathbf{g}^1, \mathbf{g}^2)$

Table 2: Ways in which the genotypes for the example in (7) can be listed in two distinct paths.

following four genotypes form a chain,

$$(-2, 2, 0, 2, -2) \preceq (0, 2, 0, 0, -2) \preceq (0, 2, 0, 0, 0) \preceq (0, 0, 0, 0, 0).$$

Chains have the property that as we look up the chain from smaller to greater elements that once a 2 or -2 becomes a 0 it remains 0. The following lemma and theorem solve the pure parsimony problem in the special case that $\mathcal{G}'$ is a chain.

Lemma 4.2. *For the diversity graph $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ assume that $\gamma(\mathcal{W}) = \mathcal{G}'$ forms a chain under $\preceq$. Let $\eta(\mathcal{V}')$ be a minimal solution and assume that $\mathbf{g} \in \mathcal{G}$ has the property that $\gamma(w) \prec \mathbf{g}$, for all $w \in \mathcal{W}$. Then there does not exist v' and v'' in $\mathcal{V}'$ such that $\eta(v') + \eta(v'') = \mathbf{g}$.*

Proof. If $|\mathcal{G}'| = 1$, then $\mathcal{V}' = \{v', v''\}$ and the result follows because $\eta(v') + \eta(v'') \in \mathcal{G}'$ but $\mathbf{g} \notin \mathcal{G}'$. So, assume that $|\mathcal{G}'| \geq 2$. Suppose for the sake of attaining a contradiction that there is a v' and v'' in $\mathcal{V}'$ such that $\eta(v') + \eta(v'') = \mathbf{g}$. Because $\eta(\mathcal{V}')$ is a minimal solution to $\mathcal{G}'$, there are no isolated nodes in $\mathcal{V}'$. Since $\mathbf{g} \notin \mathcal{G}'$, this implies that there exists $\hat{v}'$ and $\hat{v}''$ in $\mathcal{V}'$ such that $\eta(v') + \eta(\hat{v}')$ and $\eta(v'') + \eta(\hat{v}'')$ are distinct elements of $\mathcal{G}'$. Without loss of generality, we assume that $\eta(v') + \eta(\hat{v}') \prec \eta(v'') + \eta(\hat{v}'')$.

Since $\eta(v'') + \eta(\hat{v}'') \prec \mathbf{g}$ and $\mathbf{g} = \eta(v') + \eta(v'')$, we have from the definition of $\oplus$ that

$$\eta(v') \oplus \eta(v'') \oplus \eta(\hat{v}'') = \eta(v') \oplus \eta(v'').$$

Similarly, because $\eta(v') + \eta(\hat{v}') < \eta(v'') + \eta(\hat{v}'')$ and

$$\eta(v') \oplus \eta(\hat{v}') \oplus \eta(v'') \oplus \eta(\hat{v}'') = \eta(v'') \oplus \eta(\hat{v}''),$$

we have that

$$\eta(v') \oplus \eta(v'') \oplus \eta(\hat{v}'') = \eta(v'') \oplus \eta(\hat{v}'').$$

It follows that

$$\mathbf{g} = \eta(v') + \eta(v'') = \eta(v'') + \eta(\hat{v}''),$$

which is a contradiction. □

Theorem 4.4. *Assume that $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph with $\eta(\mathcal{V}) = \mathcal{H}$ and $\mathcal{E}$ as large as possible and that $\gamma(\mathcal{W}) = \mathcal{G}'$ is a chain under $\preceq$. Then a minimum solution has cardinality $|\mathcal{G}'| + 1$.*

Proof. List the elements of $\mathcal{G}'$ as $\mathbf{g}^1, \mathbf{g}^2, \ldots, \mathbf{g}^{|\mathcal{G}'|}$ and let $w^1, w^2, \ldots, w^{|\mathcal{G}'|}$ be such that $\gamma(w^i) = \mathbf{g}^i$, for $i = 1, 2, \ldots, |\mathcal{G}'|$. We first construct a solution to $\mathcal{G}'$ with cardinality $|\mathcal{G}'| + 1$. Choose $v \in \mathcal{V}$ so that $\eta(v) \prec \gamma(w^1) = \mathbf{g}^1$. Then for every $\mathbf{g}^i \in \mathcal{G}'$ there is a unique $v^i \in \mathcal{V}$ such that $\eta(v) + \eta(v^i) = \gamma(w^i) = \mathbf{g}^i$. Then, $\eta(\{v, v^1, v^2, \ldots, v^{|\mathcal{G}'|}\})$ solves $\mathcal{G}'$ and has cardinality $|\mathcal{G}'| + 1$.

We now show by induction on $|\mathcal{G}'|$ that there does not exists a solution with cardinality less than $|\mathcal{G}'|+1$. This fact is clear if $|\mathcal{G}'|=1$, and we assume the claim is true when $|\mathcal{G}'|\leq k$. Assume that $\mathcal{G}'$ is a chain of length $k+1$, and let $\eta(\mathcal{V}')$ be a minimum solution. Let $\mathcal{G}''=\mathcal{G}'\backslash\{\gamma(w^{k+1})\}$, where we assume that the elements of $\mathcal{G}'$ are ordered so that

$$\gamma(w^1)\prec\gamma(w^2)\prec\ldots\prec\gamma(w^k)\prec\gamma(w^{k+1}).$$

Since $\eta(\mathcal{V}')$ solves $\mathcal{G}''$ and a minimum solution to $\mathcal{G}''$ has cardinality $|\mathcal{G}''|+1=k+1$, we have that $|\eta(\mathcal{V}')|\geq k+1$. Suppose for sake of attaining a contradiction that $|\eta(\mathcal{V}')|=k+1$. From the induction hypothesis $\eta(\mathcal{V}')$ is a minimum solution to $\mathcal{G}'$. However, $\gamma(w^i)\prec\gamma(w^{k+1})$ for $i=1,2,\ldots,k$, and from Lemma 4.2, this leads to the contradiction that there is no v' and v'' in $\mathcal{V}'$ such that $\eta(v')+\eta(v'')=\gamma(w^{k+1})$. Hence $|\mathcal{V}'|\geq k+2=|\mathcal{G}'|+1$. Since we have already demonstrated that a solution of size $|\mathcal{G}'|+1$ exists, the proof is complete. □

Corollary 4.1 follows immediately from Theorem 4.4 and provides a bound on the pure parsimony problem.

Corollary 4.1. *Let $(\mathcal{V},\mathcal{W},\mathcal{E},\eta,\gamma)$ be a diversity graph such that $\eta(\mathcal{V})=\mathcal{H}$ and $\mathcal{E}$ is as large as possible. Partition $\mathcal{W}$ into $\mathcal{W}^1,\mathcal{W}^2,\ldots,\mathcal{W}^q$, where each $\gamma(\mathcal{W}^i)$ is a chain ordered by $\preceq$. Then a minimum solution has cardinality no greater than $|\mathcal{G}'|+q$.*

The best bound provided by Corollary 4.1 is the one that minimizes q. An interesting question for future research is whether or not calculating the minimum value of q actually solves the pure parsimony problem in some cases.

Instead of addressing the smallest size of a solution to a chain, the next result and its corollary considers how large a minimal solution can be.

Theorem 4.5. *Let $(\mathcal{V},\mathcal{W},\mathcal{E},\eta,\gamma)$ be a diversity graph such that $\eta(\mathcal{V})=\mathcal{H}$, $\mathcal{E}$ is as large as possible, and $\gamma(\mathcal{W})=\mathcal{G}'$ is a chain with respect to $\prec$. Assume that the elements of $\mathcal{W}$ are ordered so that*

$$\gamma(w^1)\prec\gamma(w^2)\prec\ldots\prec\gamma(w^{|\mathcal{G}'|}).$$

Assuming that $\gamma(w^2)$ has 3 or more zero SNPs, we have that there is a minimal solution to $\mathcal{G}'$ with cardinality $2|\mathcal{G}'|$.

Proof. The proof is by induction on $|\mathcal{G}'|$. The result clearly holds if $|\mathcal{G}'|=1$, and we assume the result is true for $|\mathcal{G}'|\leq k$. Assume that $|\mathcal{G}'|=k+1$, and let $\eta(\mathcal{V}'')$ be a minimal solution to $\mathcal{G}''=\mathcal{G}'\backslash\{\gamma(w^{k+1})\}$ whose cardinality is $2|\mathcal{G}''|$. From Lemma 4.2, $\eta(\mathcal{V}'')$ does not solve $\mathcal{G}'$ because there is no v' and v'' in $\mathcal{V}''$ such that $\eta(v')+\eta(v'')=\gamma(w^{k+1})$. We show that there is a minimal solution $\eta(\mathcal{V}')$ such that $\mathcal{V}''\subseteq\mathcal{V}'$ and $|\mathcal{V}''|+2=|\mathcal{V}'|$.

Because $\gamma(w^{k+1})$ is the $k+1$ element in a chain, it has at least $k+1$ ambiguous SNPs, and thus $|N(w^{k+1})|\geq 2^{k+1}$. Since we assumed that $\gamma(w^2)$ has at least 3 zero SNPs, $|N(w^{k+1})|>2^{k+1}$. For $j\geq 1$, we have $2j\leq 2^j$, and thus $2k<|N(w^{k+1})|/2$. Since $|N(w^{k+1})|/2$ is the number of adjacent pairs to w^{k+1} and $|\mathcal{V}''|=2k$, there is a v' and v'' in $N(w^{k+1})\backslash N(\mathcal{V}'')$ such that $\eta(v')+\eta(v'')=\gamma(w^{k+1})$. This means that $\eta(\mathcal{V}''\cup\{v',v''\})$ is a minimal solution to $\mathcal{G}'$ whose cardinality is $2|\mathcal{G}'|$. □

The condition of $\gamma(w^2)$ having at least 3 zero SNPs is not imposed because this proof requires it, but rather, it is needed by any proof due to the following example. Let $\mathcal{G}'=\{(-2,0),(0,0)\}$. Then a four element solution would have the form $\{(-1,1),(-1,-1),(x,1),(y,-1)\}$, where x is either 1 or -1 and y is the other. In either case an element is duplicated, and hence there is no solution of size 4.

The following corollary establishes that under the conditions of Theorem 4.5, there is a minimal solution of every cardinality between the minimum value of $|\mathcal{G}'|+1$ and the maximum value of $2|\mathcal{G}'|$.

Corollary 4.2. *Let $(\mathcal{V}, \mathcal{G}', \mathcal{E}, \eta, \gamma)$ be a diversity graph satisfying the condition of Theorem 4.5. Then, there is a minimal solution of cardinality j for $|\mathcal{G}'| + 1 \leq j \leq 2|\mathcal{G}'|$.*

Proof. For $1 \leq i \leq |\mathcal{G}'|$, let

$$\mathcal{G}'_1 = \{\gamma(w^1), \gamma(w^2), \ldots, \gamma(w^{|\mathcal{G}'|-i+1})\}$$

and

$$\mathcal{G}'_2 = \{\gamma(w^{|\mathcal{G}'|-i+2}), \gamma(w^{|\mathcal{G}'|-i+3}), \ldots, \gamma(w^{|\mathcal{G}'|})\}$$

be subchains of $\mathcal{G}'$. By Theorem 4.4 there is a solution $\eta(\mathcal{V}_1)$ to $\mathcal{G}'_1$ of cardinality $|\mathcal{G}'| - i + 2$ and by Theorem 4.5 there is a solution $\eta(\mathcal{V}_2)$ to $\mathcal{G}'_2$ of cardinality $2i - 2$. If $i = 1$, notice that $\mathcal{G}'_1 = \mathcal{G}'$ and that $\mathcal{G}'_2 = \emptyset$. In this case Theorem 4.4 establishes that we can indeed find a solution of cardinality $|\mathcal{G}'| + 1$. For other values of i we have that if $\mathcal{V}_1$ and $\mathcal{V}_2$ are disjoint, then $\mathcal{V}_1 \cup \mathcal{V}_2$ is a minimal solution whose cardinality is $|\mathcal{G}'| + i$, for $1 \leq i \leq |\mathcal{G}'|$. So, all that is left to show is that $\mathcal{V}_1$ and $\mathcal{V}_2$ may be selected so that they are disjoint. We accomplish this by showing that as i increases to $i+1$ that there are always enough elements of $\mathcal{V}$ to allow $\mathcal{V}_1$ and $\mathcal{V}_2$ to be disjoint.

For $i = 1, 2, \ldots, |\mathcal{G}'|$ we have that $|\mathcal{G}'| - i + 2 \leq 2^{|\mathcal{G}'|-i+2}$. As in the proof of Theorem 4.5, we have that

$$2^{|\mathcal{G}'|-i+2} < |N(w^{|\mathcal{G}'|-i+2})|/2,$$

which guarantees that there are v' and v'' in $N(w^{|\mathcal{G}'|-i+2}) \backslash \eta(\mathcal{V}_1)$ such that $\eta(v') + \eta(v'') = \gamma(w^{|\mathcal{G}'|-i+2})$. So, as i increases from i to $|\mathcal{G}'|$, we are guaranteed to be able to select disjoint $\mathcal{V}_1$ and $\mathcal{V}_2$. □

5 Directions for Future Research

The goal of this paper was to establish an initial investigation into the structure of haplotyping problems by studying the underlying graph theory. We have shown that the structural requirements of the problem are meaningful and that the majority of bipartite graphs are incapable of representing the underlying biology We have further established a solution to the pure parsimony problem in a few cases, and in particular we have shown that ordering the genotypes with $\preceq$ and decomposing $\mathcal{G}'$ into chains bounds the problem. During the writing of this paper the authors had other questions that were left unanswered, many of which promise to be fruitful continued research:

- Although the matrix equation and the logical decomposition stated in Theorem 3.2 characterize the graphs that support diversity, we would have enjoyed a more graph theoretical characterization. A theorem like $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ supports diversity if and only if it does not contain a certain structure would have been particularly appealing.

- We conjecture that m^* is 2 for acyclic diversity graphs, which means that the pure parsimony problem is solve by the algorithm in Table 1. This together with a proof of Conjecture 4.1 may highlight the class of diversity graphs for which $m^* = 2$.

- Decomposing the genotypes into chains ordered by $\preceq$ bounds the optimal value of the pure parsimony problem, but this bound can likely be reduced by investigating how the solutions to the individual chains can interact. Moreover, we do not yet know how to decompose the genotypes into the fewest number of chains. This bound could be useful in the integer programming formulation of the problem, and numerical work should be explored.

- Investigating how $\phi(m)$ decreases and estimating m^* are exciting new avenues. If we can accomplish both of these, then we will be able to estimate the solution to the pure parsimony problem.

References

[1] A. S. Asratian, T. M. J. Denley, and R. Häggkvist. *Bipartite Graphs and Their Applications*. Cambridge University, New York, NY, 1998.

[2] V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. Technical Report CSE-2002-21, University of California, Davis, Computer Science, 2002. Augmented version to appear in the Journal of Computational Biology.

[3] R. H. Chung and D. Gusfield. Empirical exploration of perfect phylogeny haplotyping and haplotypers. Technical report, University of California, Computer Science, 2003. To appear in the Proceedings of the 2003 Cocoon Conference.

[4] R. H. Chung and D. Gusfield. Perfect phylogeny haplotyper: Haplotype inferral using a tree model. *Bioinformatics*, 19(6):780–781, 2003.

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

[6] D. Gusfield. A practical algorithm for optimal inference of haplotypes from diploid populations. *Proceedings of the Eight International Conference on Intelligent Systems for Molecular Biology*, 2000.

[7] D. Gusfield. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Journal of Computational Biology*, 8(3):305–324, 2001.

[8] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. *Proceedings of RECOMB 2002: The Sixth Annual International Conference on Computational Biology*, pages 166–175, 2002.

[9] D. Gusfield. Haplotyping by pure parsimony. Technical Report CSE-2003-2, University of California, Davis, 2003. To appear in the Proceedings of the 2003 Combinatorial Pattern Matching Conference.

[10] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz. SNPs problems, complexity and algorithms. In *European Symposium on Algorithms*, volume 2161 of *Lecture Notes in Computer Science*, pages 182–193. Springer-Verlag, 2001.

[11] G. Lancia and M. Perlin. Genotyping of pooled microsatellite markers by combinatorial optimization techniques. *Discrete Applied Mathematics*, 88(1-3):291–314, 1998.

[12] G. Lancia, M. Pinotti, and R. Rizzi. Haplotyping populations by pure parsimony. complexity, exact and approximation algorithms. *INFORMS Journal on Computing*, 16(4):348–359, 2004.

[13] S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *American Journal of Human Genetics*, 71:1129–1137, 2002.

[14] R. Lippert, R. Schwartz, G. Lancia, and S. Istrail. Algorithmic strategies for the SNPs haplotype assemply problem. *Briefings in Bioinformatics*, 3(1):23–31, 2001.

[15] T. Niu, Z. S. Quin, X. Xu, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002.

[16] D. Qian and L. Beckmann. Minimum-recombinant haplotyping in pedigrees. *American Journal of Human Genetics*, 70:1434–1445, 2002.

[17] R. Rizzi, V. Bafna, S. Istrail, and G. Lancia. Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. In R. Guigo and D. Gusfield, editors, *Algorithms in Bioinformatics: Proceedings of the Second International Workshop on Algorithms on Bioinformatics, WABI 2002, Rome, Italy, September 17-21, 2002*, volume 2452 of *Lecture Notes in Computer Science*, pages 29–43. Springer-Verlag Berlin Heidelberg, 2002.

[18] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.

[19] C. F. Xu, K. Lewis, K. L. Cantone, P. Khan, C. Donnelly, N. White, N. Crocker, P. R. Boyd, D. V. Zaykin, and I. J. Purvis. Effectiveness of computational methods in haplotype prediction. *Human Genetics*, 110:148–156, 2002.