Haplotyping and Minimum Diversity Graphs

C. Davis[†] and A. Holder[‡]

July 17, 2003

Abstract

Haplotyping is the process of reconstructing the genetic information donated by a prior generation to form a current population. Haplotyping is important because it allows us to study how traits are passed from one generation to another, which in turn allows us to find genetic markers that describe a current population's susceptibility to diseases. Our goal is to study the underlying graph theory problem, and we study the bipartite graphs, called diversity graphs, that describe haplotyping. In particular, we investigate the problem of finding the minimum number of haplotypes that can reconstruct a population, called the Pure Parsimony problem. The graph theory representation provides significant insight if the number of mates is restricted.

Keywords: Haplotyping, Parsimony, Diversity Graph, Longest Paths, Optimization

---

[†] University of Utah, Department of Mathematics, Salt Lake City, UT, USA. Research conducted at Trinity University.

[‡] Trinity University Mathematics, San Antonio, TX, USA.

1 Introduction

With the successes of the human genome project, understanding how the genetic composition of a population is formed from a prior generation is becoming increasingly important. Indeed, the search for genetic markers that indicate an individual's susceptibility to diseases and the design of patient-specific drugs depend on the ability to reconstruct genetic information donated by prior generations. This reconstruction is called haplotyping, and the problem is to start with genotypic information from a current population and find a collection of haplotypes (genetic donations from the previous generation) that explain the current population.

Genes are linear sequences of deoxyribonucleic acid (DNA) that code the genotype of every living organism. In diploid organisms, such as humans, genes are paired to describe physical traits. The binary alphabet $\{A, B\}$ is used to describe genes, and so a gene is a sequence like $ABABBA$. The positions are single nucleotide polymorphisms (SNPs), and for this example the first SNP is an A, the second is a B, and so on. If this gene is paired with $AAABBB$, the resulting physical trait is $AXABBX$, where an A indicates that both SNPs are an A, a B implies that both SNPs are a B, and an X means that one SNP is an A and the other is a B. An individual's genotype is a sequence of paired genes.

Physical traits are formed by the parents' donated genes, and a haplotype is the sequence of genes donated by a single parent, see Figure 1. Genotyping a population is the process of collecting gene sequences from individuals in the population. So, when a population is genotyped, a sequence of paired genes expressed over the alphabet $\{A, B, X\}$ is collected from each individual. It is important to note that we do **not** obtain the haplotypes donated by the parents—i.e., we do not know the genetic donations made by the parents that form the physical traits. The haplotyping problem is to construct a set of haplotypes that can be paired to form the genotypes of the sampled population. Clark [5] first introduced the problem and suggested the following haplotyping technique, known as Clark's Rule. Select a genotype and construct two haplotypes that form the genotype. Repeatedly select one of the remaining genotypes and construct new haplotypes only if the genotype cannot be formed by pairing previously constructed haplotypes.

Since Clark's original work, several researchers have suggested other techniques and variations on the problem [2, 3, 4, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17]. The problem we address is called the Pure Parsimony problem, which means we are searching for the smallest number of haplotypes that can be paired to form our collection of genotypes. This NP-Hard problem is suggested in [6] and is first investigated in [9] by Gusfield. In this last work, Gusfield introduces an integer programming formulation and describes techniques to reduce the problem so that biologically relevant problems can be solved. Our interest is different in that we explore the graphs that explain the Pure Parsimony problem. In Section 2 we introduce the concept of a *diversity graph* and establish several of its properties. In Section 3 we study the Pure Parsimony problem with a restricted mating structure. The most important result of this section shows that a restricted form of the Pure Parsimony problem can be solved by decomposing the graph into cycles and longest paths.

2 Notation & Basic Results

Let G be the set of genotypes collected from a population and H be a set of haplotypes. Throughout this paper h is a haplotype and g is a genotype, where h_i^j is SNP i on haplotype h^j and g_i^j is SNP i on genotype g^j (note that if only one haplotype or genotype is considered, the superscript j is disregarded). We assume that each haplotype and genotype has n SNPs. For parent haplotypes h^1 and h^2 and offspring genotype g, we have the following at each SNP:

- g_i = A if, and only if, $h_i^1 = h_i^2 =$ A.
- g_i = B if, and only if, $h_i^1 = h_i^2 =$ B.
- g_i = X if, and only if, either $h_i^1 =$ A and $h_i^2 =$ B, or $h_i^1 =$ B and $h_i^2 =$ A.

We say that $h^1 \oplus h^2 = g$ provided that h^1, h^2, and g adhere to these rules. For example, let $h^1 = AABAAB$ and $h^2 = ABBABB$. Then, $h^1 \oplus h^2 = g = AXBAXB$. It is easy to see that $\oplus$ is a binary operation with the property that $h^i \oplus h^j = h^i \oplus h^k$ implies $h^j = h^k$. Parental haplotypes that contribute genetic information to the same offspring's genotype are called *mates*. That is, if $h^1 \oplus h^2 = g$, we say that h^1 mates with h^2 to form g. Furthermore, we say that h^1 *resolves* g if $h^1 \oplus h^2 = g$ for some h^2. This concept is extended to sets, and we say that H resolves G if for each $g \in G$, there is an h^1 and h^2 in H such that $h^1 \oplus h^2 = g$.

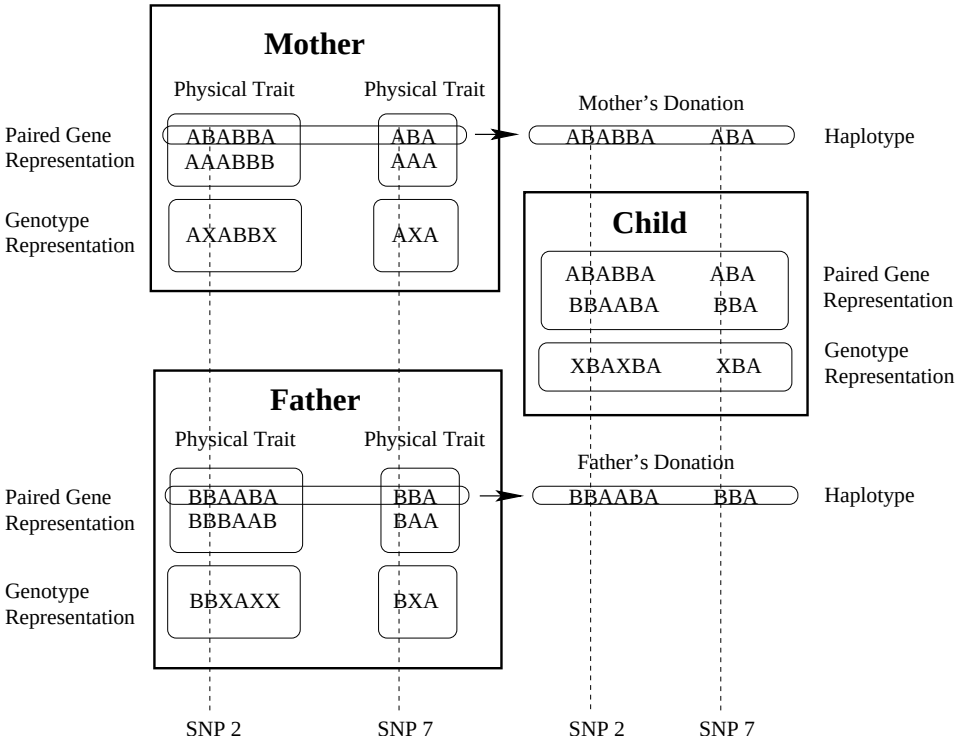

Figure 1: The mother and father's haplotype donations combine to form two physical traits in the child. SNP 2 and 7 are highlighted by dashed lines.

Throughout we assume a familiarity with graph theory, and we suggest [1] as a reference. A *diversity graph* is a bipartite graph with vertex sets H and G and with its edge set defined by the mating structure. The definition is based on the operation $\oplus$ and is biologically based.

Definition 1 *A bipartite graph $D = (H, G, E)$ is a diversity graph if*

- *G is nonempty,*
- *each genotype in G is resolved by some haplotype in H, and*
- *E has the property that if $(h^1, g) \in E$, then there exists an $h^2 \in H$ such that $(h^2, g) \in E$ and $h^1 \oplus h^2 = g$.*

This definition does not require that a genotype be connected to every possible pair of haplotypes, but it does require that every genotype be connected to at least one pair of mates. The neighborhood of a genotype g is $N(g) = \{h : (h, g) \in E\}$, and the degree of g, denoted $deg(g)$, is the number of edges emanating from g. The third condition in the definition of a diversity graph guarantees that $deg(g)$ is even for each genotype.

A bipartite graph is a diversity graph if we can label the vertices so that the edges represents a valid mating structure. So, the bipartite graph (V, W, E) is a diversity graph if we 1) label each node in V with a sequence of n SNPs over the alphabet $\{A, B\}$, 2) label the nodes in W with sequences of n SNPs over the alphabet $\{A, B, X\}$, and 3) guarantee that the edge set satisfies the last two conditions of Definition 1. For example, consider the bipartite graphs $(\{v^1, v^2\}, \{w^1\}, \{(v^1, w^1), (v^2, w^1)\})$ and $(\{v^1, v^2\}, \{w^1\}, \{(v^1, w^1)\})$. The first graph is a diversity graph because we can label v^1, v^2, and w^1 with an A, which clearly creates a diversity graph. The second graph is not a diversity graph because it does not have enough edges to satisfy the third condition of Definition 1.

There are graphs with the property that every node has even degree but that are not diversity graphs. An example is $K(2, 2)$ (the complete bipartite graph with both vertex sets containing two nodes), which has the property that all nodes in W have degree 2 but the nodes cannot be labeled so that the edge set satisfies $\oplus$. In fact, any graph containing $K(2, 2)$ cannot be a diversity graph for the same reason, and hence, $K(m, n)$ is not a diversity graph if m and n are at least 2.

The set of all possible haplotypes with n SNPs is $\mathcal{H}$, and $\mathcal{E}$ is the largest edge set between $\mathcal{H}$ and G that is allowed by the third condition of Definition 1. A subtle convention is that if a genotype has no SNPs with a value of X (and hence a haplotype mates with itself to form the genotype), we include the edge between the haplotype and genotype twice. So, for the genotype $ABBABB$, the edge set contains $(ABBABB, ABBABB)$ twice.

The Pure Parsimony problem may be stated as finding a subgraph of $(\mathcal{H}, G, \mathcal{E})$ with the following properties: 1) G is a vertex set, 2) the subgraph is a diversity graph, and 3) the subcollection of $\mathcal{H}$ is as small as possible. Any subgraph with these properties is said to be optimal and is denoted (H^*, G, E^*). There are typically several optimal subgraphs, each of which has the property that the fewest number of haplotypes required to resolve G is $|H^*|$.

A SNP is *ambiguous* if it has a value of X, and a gene is ambiguous if at least two of its SNPs are ambiguous. Non-ambiguous genotypes are special because they are resolved by a unique pair of haplotypes. For example, the genotype $ABABA$, is uniquely resolved as $ABABA \oplus ABABA$ and the genotype $AXBBA$ is uniquely resolved as $ABBBA \oplus AABBA$. If r_g is the number of ambiguous SNPs on genotype g, there are at most 2^{r_g} haplotypes that can resolve g. So, in any diversity graph we have for every genotype that $deg(g) \leq 2^{r_g}$. This bound is tight for $(\mathcal{H}, G, \mathcal{E})$.

Theorem 1 shows that ordering the elements of $\mathcal{H}$ lexicographically provides a convenient way to list the 2^{n-1} pairs of haplotypes that mate to form the genotype that is ambiguous at every SNP. This result is used in Lemma 1 to show that any bipartite graph with no isolated nodes can be extended to a diversity graph.

Theorem 1 *If the elements of $\mathcal{H}$ are lexicographically ordered (where $A < B$), we have for $1 \leq j \leq 2^n$ that $h^j \oplus h^{(2^n - j+1)} = XX...X$.*

Proof: List the haplotypes lexicographically so that $h^j < h^{j+1}$ (so $h^1 = AA...A$). Create a second list of the haplotypes by exchanging every A and B. Notice that in this new list $h^j > h^{j+1}$ for all j. So, the second list is the reverse of the first list.

Fix j between 1 and 2^n. Let k be any SNP such that $h^j_k = A$. So, if we start at $h^1_k = A$ and travel down the first list j haplotypes, we get an A at h^j_k. Recall that reading the first list in reverse order is the same as reading the second list top down. Hence, moving up j haplotypes from $h^{2^n}_k = B$ in the first list is the same as moving down the second list j haplotypes, and the construction of the second list guarantees that we stop at a B. This means $h^{(2^n-j+1)}_k = B$. So, we have that $h^j_k = A$ if, and only if, $h^{(2^n-j+1)}_k = B$. Hence, $h^j_k \oplus h^{(2^n-j+1)}_k = X$. A similar argument works if $h^j_k = B$. ■

As an example, for $n = 3$ the first list is lexicographically ordered from top to bottom and the second list is formed by exchanging A's and B's. If we apply $\oplus$ componentwise, each application produces XXX.

$$\begin{pmatrix} AAA \\ AAB \\ ABA \\ ABB \\ BAA \\ BAB \\ BBA \\ BBB \end{pmatrix} \oplus \begin{pmatrix} BBB \\ BBA \\ BAB \\ BAA \\ ABB \\ ABA \\ AAB \\ AAA \end{pmatrix} = \begin{pmatrix} XXX \\ XXX \\ XXX \\ XXX \\ XXX \\ XXX \\ XXX \\ XXX \end{pmatrix}.$$

The next Lemma shows that bipartite graphs with no isolated vertices can be extended to a diversity graph, and it provides a tight upper bound on the number of nodes that need to be added. Consider the bipartite graph (V, W, E), and let V be the vertex set that we extend and W be the set that remains unchanged (so V and W become the haplotype and genotype set, respectively, after labeling). For $w \in W$, define

$$T(w) = \bigcup_{w' \neq w} [N(w) \cap N(w')].$$

So, $T(w)$ is the collection of nodes in the neighborhood of w that are also in the neighborhood of another node in W. For any real number C we define C_+ by

$$C_+ = \begin{cases} 0 & \text{if} \quad C < 0 \\ C & \text{if} \quad C \geq 0. \end{cases}$$

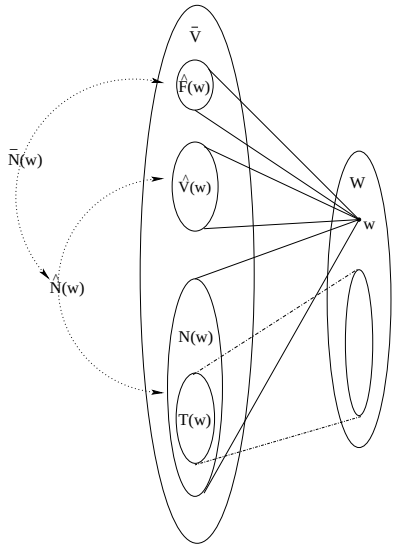

Figure 2: Geometric representation of the extensions.

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 exactly the number of points in $T(w)$, except for a specific case. Let $\hat{V}(w)$ be a set of cardinality $(2|T(w)| - |N(w)|)_+$, and let $\hat{N}(w) = N(w) \cup \hat{V}(w)$. Notice that if $2|T(w)| \leq |N(w)|$, $N(w)$ is not extended. This leads to problems because $|\hat{N}(w)|$ can be odd. We remedy this by extending $\hat{N}(w)$ with $\hat{F}(w)$, which is empty if $|\hat{N}(w)|$ is even and contains a single node if $|\hat{N}(w)|$ is odd. Formally, let

$$\theta = \{w : |\hat{N}(w)| = 2p + 1 \text{ for } p = 1, 2, 3, \ldots\}$$

and

$$\chi_\theta(w) = \begin{cases} 1, & w \in \theta \\ 0, & w \notin \theta. \end{cases}$$

We set $|\hat{F}(w)| = \chi_\theta(w)$. Union $\hat{V}(w)$ and $\hat{F}(w)$ with each w neighborhood so that the new neighborhood of each w is

$$\bar{N}(w) = N(w) \cup \hat{V}(w) \cup \hat{F}(w) = \hat{N}(w) \cup \hat{F}(w).$$

The extended vertex set is

$$\bar{V} = V \cup \left(\bigcup_{w \in W} \hat{V}(w) \right) \cup \left(\bigcup_{w \in W} \hat{F}(w) \right).$$

Notice that after the nodes have been labeled, we consider v to be a haplotype and w to be a genotype, and hence we may apply the operation $\oplus$ to the nodes. If $deg(w) = 1$, it is possible for w to be resolved by a single v after labeling if we include an additional edge so that $v \oplus v = w$. This situation requires a double edge, and we let $\hat{E}(w)$ be the multiset

$$\{(v, w), (v, w) : deg(w) = 1, v \in N(w)\backslash T(w)\}.$$

The expanded edge set is

$$\bar{E} = E \cup \{(v, w) : v \in \bar{N}(w)\} \cup \hat{E}(w).$$

Note that throughout we have denoted an extension with the $\hat{}$ notation and an extended set with the $\bar{}$ notation. See Figure 2 for a geometric depiction of the definitions.

Lemma 1 *Any bipartite graph (V, W, E) with no isolated nodes can be extended and labeled to*

become a diversity graph by adding no more than

$$\sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+$$

nodes to V. In particular, $(\bar{V}, W, \bar{E})$ is an extension of (V, W, E) with this number of nodes added to V that can be labeled to become a diversity graph.

Proof: The proof follows by induction on $|W|$. Let (V, W, E) be a bipartite graph with no isolated nodes such that $|W| = 1$. So, $W = \{w\}$. Notice that $T(w) = \hat{V}(w) = \emptyset$, and $\hat{N}(w) = N(w)$. There are three cases to consider.

Case 1: Suppose $deg(w) = 1$. Then, $|\hat{V}(w)| = (2|T(w)| - |N(w)|)_+ = (0-1)_+ = 0$. Since $deg(w) = 1$, $\chi_\theta(w) = |\hat{F}(w)| = 0$. So, $\bar{N}(w) = N(w) \cup \hat{V}(w) \cup \hat{F}(w) = N(w) = \{v\}$. Hence, $\bar{V} = V$ and

$$|\bar{V}| - |V| = 0 = \chi_\theta(w) + (2|T(w)| - |N(w)|)_+.$$

From $\hat{E}(w) = \{(v, w), (v, w)\}$ we see that $\bar{E} = E \cup \hat{E}(w) = \hat{E}(w)$. Label w and v by setting $w = v = A$, so that $v \oplus v = w$. With this labeling, $(\bar{V}, W, \bar{E})$ becomes a diversity graph.

Case 2: Suppose $deg(w) = 2p$ for some $p \in \mathbb{N}$. Then, $|\hat{V}(w)| = (2|T(w)| - |N(w)|)_+ = (-2p)_+ = 0$. Notice that since $\chi_\theta(w) = 0$, $\bar{N}(w) = \bar{V} = V$. Likewise, $\hat{E}(w) = \emptyset$, and $\bar{E} = E$. We now have that

$$(\bar{V}, W, \bar{E}) = (V, W, E) \quad \text{and} \quad |\bar{V}| - |V| = 0 = \chi_\theta(w) + (2|T(w)| - |N(w)|)_+.$$

Let $n \in \mathbb{N}$ be such that $2p \leq 2^n$. Set w to be a sequence of Xs of length n. Let the nodes in $N(w)$ be ordered from v^1 to v^{2p}. Set $v_i^1 = A$ for all SNPs i. For $j \leq p-1$, let v^{j+1} be the next lowest permutation lexicographically after v^j, where $A < B$. For $j > p$ use the lexicographic ordering established in Theorem 1 by setting v^{2p-j+1} to be the haplotype $2^n - j + 1$ from Theorem 1, so that we have $v^j \oplus v^{(2p-j+1)} = XX...X$. Now, every node in $N(w)$ mates with another to form w, and $(\bar{V}, W, \bar{E})$ is labeled.

Case 3: Suppose $deg(w) = 2p-1$ for some $p = 2, 3, \ldots$. Then, $|\hat{V}(w)| = (2|T(w)| - |N(w)|)_+ = (-(2p-1))_+ = 0$ and $\chi_\theta(w) = 1$. So, $\bar{N}(w) = \bar{V} = V \cup \hat{F}(w)$. So,

$$|\bar{V}| - |V| = 1 = \chi_\theta(w) + (2|T(w)| - |N(w)|)_+.$$

Also since $\hat{E}(w) = \emptyset$, we have that $\bar{E} = E$. Note that $|\bar{V}| = 2p$. So we may use the labeling scheme from Case 2 to see that $(\bar{V}, W, \bar{E})$ may be labeled.

Assume that if $|W| \leq k$, the result holds. Let (V, W, E) be a bipartite graph with no isolated nodes such that $|W| = k+1$. Select $w^1 \in W$, and let (V_k, W_k, E_k) be the subgraph of (V, W, E) with w^1 and its edges and neighboring vertices removed. Extend and label (V_k, W_k, E_k) so that $\overline{D}_k = (\bar{V}_k, W_k, \bar{E}_k)$ becomes a diversity graph. Let n_k be the number of SNPs in the labeling, and let

$$s_k = \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+.$$

Notice that the mating structure is unchanged if we lengthen the number of SNPs for each $w \in W_k$ and $v \in \bar{V}_k$ by assigning the same value at every additional SNP. For instance, if $v^1 \oplus v^2 = ABB \oplus BAB = XXB = w$, adding two As at the end provides $v^1 \oplus v^2 = ABBAA \oplus BABAA = XXBAA = w$. Each of the following cases lengthens the number of SNPs to $n_k + p$ in this fashion.

Case 1: Suppose $|N(w^1)| = 1$ and $T(w^1) = \emptyset$. Then $|\hat{V}(w^1)| = (2|T(w^1)| - |N(w^1)|)_+ = 0$. Therefore, $|\hat{N}(w^1)| = 1$ and $\chi_\theta(w^1) = 0$. Thus, $\bar{N}(w^1) = N(w^1)$, which means $\bar{V} = \bar{V}_k \cup V$. So, we have the following:

$$\begin{aligned}
&|\bar{V}| - |V| \\
&= s_k + 0 + 0 \\
&= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\
&= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+.
\end{aligned}$$

Since $\hat{E}(w^1) = \{(v, w^1), (v, w^1)\}$, we have that $\bar{E} = E \cup \hat{E}(w^1)$. Increase the number of SNPs in the nodes in W_k and $\bar{V}_k$ by adding SNP $n_k + 1$ containing an A to every node. Likewise, let the number of SNPs in w^1 and $v^1 \in N(w^1)$ be $n_k + 1$, and let $v_i^1 = w_i^1 = B$ for all i. Notice that w^1 and v^1 are unique since they disagree with every node in W_k and $\bar{V}_k$ at SNP $n_k + 1$. Since $v^1 \oplus v^2 = w$, we see that $(\bar{V}, W, \bar{E})$ is labeled.

Case 2: Suppose $|N(w^1)| = 1$ and $T(w^1) \neq \emptyset$. So, $T(w^1) = N(w^1)$. Then, $|\hat{V}(w^1)| = (2|T(w^1)| - |N(w^1)|)_+ = 1$, which means $|\hat{N}(w^1)| = 2$ and $\chi_\theta(w^1) = 0$. Therefore, $\bar{N}(w^1) = \hat{N}(w^1) = N(w^1) \cup \hat{V}(w^1)$ and $\bar{V} = \bar{V}_k \cup V \cup \hat{V}(w^1)$. We now have that

$$\begin{aligned} & |\bar{V}| - |V| \\ &= s_k + 0 + 1 \\ &= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\ &= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+. \end{aligned}$$

In this case, we see that $\hat{E}(w^1) = \emptyset$ and $\bar{E} = E \cup \{(v, w^1) : v \in \hat{V}(w^1)\}$. Let $\{v^1\} = N(w^1)$ and $\{v^2\} = \hat{V}(w^1)$. Lengthen the number of SNPs in each node in W and $\bar{V}$ by adding an A at SNP $n_k + 1$. Since v^1 is in $\bar{V}_k$, it is already labeled through the first n_k SNPs. So let $v_i^2 = v_i^1 = w_i^1$ in every SNP $i \neq n_k + 1$. Set $v_{n_k+1}^2 = B$ and $w_{n_k+1}^1 = X$. Now $v^1 \oplus v^2 = w^1$. Since we already know that $(\bar{V}_k, W_k, \bar{E}_k)$ has been labeled, we have that $(\bar{V}, W, \bar{E})$ is labeled.

Case 3: Suppose $|N(w^1)| = 2p$ for some $p \in \mathbb{N}$ and $T(w^1) = \emptyset$. Then $(2|T(w^1)| - |N(w^1)|)_+ = 0$ and $|\hat{N}(w^1)| = 2p$. So, $\chi_\theta(w^1) = 0$ and $\bar{N}(w^1) = V$. Hence, $\bar{V} = \bar{V}_k \cup V$ and

$$\begin{aligned} & |\bar{V}| - |V| \\ &= s_k + 0 \\ &= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\ &= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+. \end{aligned}$$

Also, since $\hat{E}(w^1) = \emptyset$, $\bar{E} = E$. Increase the number of SNPs in the nodes in W and $\bar{V}$ to $n_k + p$. For every $w \in W_k$ and $v \in \bar{V}_k$, let $w_i = v_i = A$ for $n_k < i \leq n_k + p$. For every $v \in N(w^1)$, let $v_i = B$ for $n_k < i \leq n_k + p$. Next, let $w_i^1 = X$ for $1 \leq i \leq n_k$ and $w_i^1 = B$ for $n_k < i \leq n_k + p$. Order the nodes in $N(w^1)$ from v^1 to v^{2p}, and set $v_i^1 = A$ for all SNPs i. For $j \leq p - 1$, let v^{j+1} be the next lowest lexicographic permutation after v^j, where $A < B$. For $j \geq p$ use the lexicographic ordering from Theorem 1 by setting v^{2p-j+1} to be the Theorem's value for $h^{2^n - j+1}$. Then we have that $v^j \oplus v^{(2p-j+1)} = XX...X$, and $(\bar{V}, W, \bar{E})$ is labeled.

Case 4: Suppose $|N(w^1)| = 2p - 1$ for $p = 2, 3, \ldots$ and $T(w^1) = \emptyset$. Since $(2|T(w^1)| - |N(w^1)|)_+ = 0$, we have that $\hat{V}(w^1) = \emptyset$. From the assumption that $|\hat{N}(w^1)| = 2p - 1$, we see that $\chi_\theta(w^1) = 1$. Thus, $\bar{N}(w^1) = N(w^1) \cup \hat{F}(w^1)$ and $\bar{V} = \bar{V}_k \cup V \cup \hat{F}(w^1)$. We now have that

$$\begin{aligned} & |\bar{V}| - |V| \\ &= s_k + 1 + 0 \\ &= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\ &= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+. \end{aligned}$$

In this case, $\bar{E} = E \cup \{(v, w^1) : v \in \hat{F}(w^1)\}$. Notice that with the extension, $|\bar{N}(w^1)| = 2p$. Thus, we can use the labeling scheme described in the previous case to label $(\bar{V}, W, \bar{E})$.

Case 5: Suppose $|N(w^1)| = p$ for $p = 2, 3, \ldots$ and $T(w^1) \neq \emptyset$. Let $\hat{V}(w^1)$ have cardinality $(2|T(w^1)| - |N(w^1)|)_+$, and extend the neighborhood of w^1 by setting $\hat{N}(w^1) = N(w^1) \cup \hat{V}(w^1)$. By examining the following cases, we see that $|\bar{N}(w^1)|$ is always even.

Case i: Suppose $p \in \mathbb{N}$ and $2|T(w^1)| \geq p$. If $|\hat{V}(w^1)| = (2|T(w^1)| - |N(w^1)|)_+$ is even, then $|N(w^1)|$ is even, and hence $|\hat{N}(w^1)| = p + |\hat{V}(w^1)|$ is even. Similarly, if $|\hat{V}(w^1)| = (2|T(w^1)| - |N(w^1)|)_+$ is odd, then $p = |N(w^1)|$ is odd and $|\hat{N}(w^1)| = p + |\hat{V}(w^1)|$ is even. In either case $\chi_\theta(w^1) = 0$, and $\bar{N}(w^1) = \hat{N}(w^1)$. Hence, $|\bar{N}(w^1)|$ is even. Since $\bar{V} = \bar{V}_k \cup V \cup \hat{V}(w^1)$, the overall extension satisfies,

$$\begin{aligned}
&|\bar{V}| - |V| \\
&= s_k + 0 + (2|T(w^1)| - |N(w^1)|)_+ \\
&= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\
&= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+.
\end{aligned}$$

Since $\hat{E}(w^1) = \emptyset$, we have in this case that $\bar{E} = E \cup \{(v, w^1) : v \in \hat{V}(w^1)\}$.

Case ii: Suppose p is even and $2|T(w^1)| < p$. Then, $|\hat{V}(w^1)| = 0$, and $|\hat{N}(w^1)|$ is even. Therefore, $\chi_\theta(w^1) = 0$ and $|\bar{N}(w^1)|$ is even. Thus, $\bar{V} = \bar{V}_k \cup V$ and

$$\begin{aligned}
&|\bar{V}| - |V| \\
&= s_k + 0 + 0 \\
&= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\
&= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+.
\end{aligned}$$

Since $\hat{E}(w^1) = \emptyset$, the extended edge set is $\bar{E} = E$.

Case iii : Suppose p is odd and $2|T(w^1)| < p$. Then, $|\hat{V}(w^1)| = 0$, and $|\hat{N}(w^1)|$ is odd. So, $|\hat{F}(w^1)| = \chi_\theta(w^1) = 1$, which gives us that $|\bar{N}(w^1)| = |\hat{N}(w^1)| + \chi_\theta(w^1)$ is even. Hence $\bar{V} = \bar{V}_k \cup V \cup \hat{F}(w^1)$, and so

$$\begin{aligned}
&|\bar{V}| - |V| \\
&= s_k + 1 + 0 \\
&= \sum_{w \in W_k} \chi_\theta(w) + \sum_{w \in W_k} (2|T(w)| - |N(w)|)_+ + \left[\chi_\theta(w^1) + (2|T(w^1)| - |N(w^1)|)_+\right] \\
&= \sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+.
\end{aligned}$$

Since $\hat{E}(w^1) = \emptyset$, we have that $\bar{E} = E \cup \{(v, w^1) : v \in \hat{F}(w^1)\}$.

From these cases we see that the size of the extended neighborhood is even and the extensions match the number specified in the lemma statement.

Let every node in $\bar{V}$ and W have $n_k + p$ SNPs. Note that W_k and $\bar{V}_k$ have values assigned to SNPs 1 through n_k by assumption. In W_k and $\bar{V}_k$, let SNPs $n_k + 1$ through $n_k + p$ all contain As. Let $w_i^1 = X$ for all SNPs i, and note that w^1 is unique from every other genotype. Also, notice that each $v \in T(w^1)$ is already labeled. For every $v \in T(w^1)$, label a $v' \in \hat{N}(w^1) \backslash T(w^1)$ such that $v \oplus v' = w^1$. This yields a value for v' that differs from every other node in $\bar{V}_k$ since SNPs $n_k + 1$ through $n_k + p$ of v^1 must contain Bs by construction. If $2|T(w^1)| \geq p$, there are no unlabeled nodes remaining in $\hat{N}(w^1)$. Since $\chi_\theta(w^1) = 0$ in this case, we have labeled $(\bar{V}, W, \bar{E})$. However, if $2|T(w^1)| < p$, then while every node in $\hat{V}(w^1)$ is labeled, $N(w^1)$ still contains unlabeled nodes. Let $\bar{F}(w^1)$ be the set of remaining unlabeled nodes, and notice that $|\bar{F}(w^1)|$ is even by construction. Let $|\bar{F}(w^1)| = 2f$ and order the nodes in $\bar{F}(w^1)$ from v^1 to v^{2f}. For v^1 through v^f, let the first $n_k + 1$ SNPs contain Bs and let SNP $n_k + 2$ be an A. Conversely, for v^{f+1} through v^{2f} let the first $n_k + 1$ SNPs contain As and let SNP $n_k + 2$ be a B. So, $\bar{F}(w^1)$ looks like,

$$
\begin{array}{ll}
\{ & \text{BBB...BBA} - - \ldots -, \\
& \text{BBB...BBA} - - \ldots -, \\
& \qquad \vdots \\
& \text{BBB...BBA} - - \ldots -, \\
& \text{AAA...AAB} - - \ldots -, \\
& \text{AAA...AAB} - - \ldots -, \\
& \qquad \vdots \\
& \text{AAA...AAB} - - \ldots - \quad \}.
\end{array}
$$

Lexicographically assign SNPs n_k+3 through n_k+p on v^1 through v^{2^f} as demonstrated in Case 2 of the base case, but alter the indices so that $v^j \oplus v^{2f-j+1} = w^1$ where $1 \leq j \leq 2f$. Thus, if $v^1 \in \bar{F}(w^1)$, we have for $v^{2f-j+1} \in \bar{F}(w^1)$ that $v^1 \oplus v^{2f-j+1} = w^1$. Therefore, $(\bar{V}, W, \bar{E})$ is labeled.

■

Figure 3 demonstrates the extension of a graph (V, W, E) where $|W| = 2$, $|N(w^1)| = 5$, $|N(w^2)| = 2$, and $|T(w^1)| = |T(w^2)| = 2$. Since $2|T(w^1)| < |N(w^1)|$, we have that $|\hat{V}(w^1)| = 0$, and since $|\hat{N}(w^1)| = 3$, we see that $\chi_\theta(w^1) = 1$. So, the neighborhood of w^1 is extended by one node via $\hat{F}(w^1)$. For w^2, we have that $|N(w^2)| = |T(w^2)| = 2$, so $|\hat{V}(w^1)| = 2$ and $\chi_\theta(w^2) = 0$. Hence, the neighborhood of w^2 is extended by the two nodes in $\hat{V}(w^2)$. We mention that not all of the extended nodes are required because it is possible to label one of w^1 or w^2 with all Xs and eliminate the need for $\hat{V}(w^2)$. We now show that any bipartite graph, including those with isolated nodes, can be extended and labeled to become a diversity graph. The result is a simple extension of Lemma 1.

Theorem 2 *Any bipartite graph (V, W, E) can be extended and labeled to become a diversity graph by adding no more than*

$$\left[\sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+ \right] + (M_V - M_W)_+,$$

where M_V and M_W are the number of isolated nodes in V and W, respectively.

Proof: Let (V, W, E) be a bipartite graph, and let V_I and W_I be collections of isolated nodes from V and W, respectively. Also, let (V', W', E) be the subgraph of (V, W, E) with the isolated nodes removed. Extend and label (V', W', E) as in Lemma 1 so that $(\bar{V}', W, \bar{E}')$ is a diversity graph. This extension requires adding

$$\sum_{w \in W'} \chi_\theta(w) + \sum_{w \in W'} (2|T(w)| - |N(w)|)_+ \tag{1}$$

nodes to V'. For every node in W_I we have that $\chi_\theta(w) = (2|T(w)| - |N(w)|)_+ = 0$, so we can write (1) as

$$\sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+. \tag{2}$$

Let n' be the number of SNPs used in Lemma 1 and $\hat{V}_I$ be a set of cardinality $(M_V - M_W)_+$. Also let $\bar{V}_I = V_I \cup \hat{V}_I$ and notice that $|\bar{V}_I| \geq |W_I|$. Add $|\bar{V}_I|$ SNPs so that every node has $n' + |\bar{V}_I|$ SNPs. For each node in $\bar{V}'$ and W, label SNPs $n'+1$ through $n'+|\bar{V}_I|$ with an A so that $(\bar{V}', W, \bar{E}')$ is still a diversity graph.

For each node in $\bar{V}_I$ and W_I, label SNP $n'+|\bar{V}_I|$ with a B. Order the nodes in W_I and $\bar{V}_I$ from w^1 to w^{M_W} and v^1 to $v^{M_V+|\hat{V}_I|}$. For SNPs 1 through $n'+|\bar{V}_I|-1$, label the nodes in W_I and $\bar{V}_I$ lexicographically. So $w^1 = v^1 = AA...AB$, $w^2 = v^2 = AA...BB$, and so on. Since $|\bar{V}_I|$ is at least as large as $|W_I|$, this labeling scheme continues to label all the nodes in $\bar{V}_I$ once the nodes in W_I are labeled. Extend $\bar{E}'$ to the multiset $\bar{E}$ such that $\bar{E} = \bar{E}' \cup \{(v, w), (v, w) : v \oplus v = w \in W_I\}$. Setting $\bar{V} = \bar{V}' \cup \hat{V}_I$, we see that $(\bar{V}, W, \bar{E})$ is labeled and and that V is extended by

$$\left[\sum_{w \in W} \chi_\theta(w) + \sum_{w \in W} (2|T(w)| - |N(w)|)_+ \right] + (M_W - M_V)_+.$$

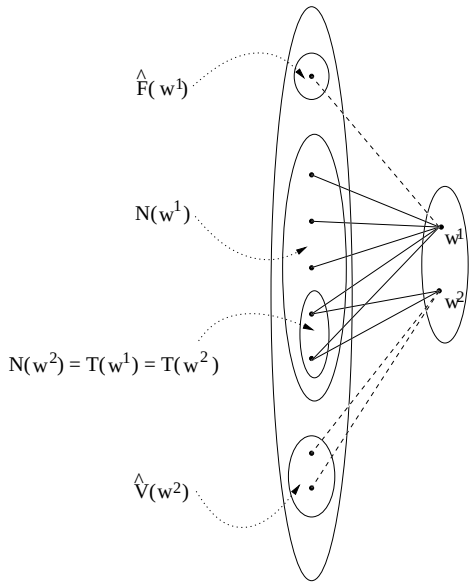

Figure 3: Example with all extensions from Lemma 1 used.

We next show that $2|G|$ haplotypes are needed to resolve G exactly if the neighborhoods of the genotypes partition $\mathcal{H}$. In support of this result, Lemma 2 shows that every H^* contains haplotypes that resolve multiple genotypes, provided that any such haplotypes exist.

Lemma 2 *Suppose that $T(g) \neq \emptyset$ for some $g \in G$. Then, H^* contains an element of $\bigcup_{g \in G} T(g)$.*

Proof: Let $T(g) \neq \emptyset$ for some $g \in G$. Suppose that H^* does not contain an element of $\bigcup_{g \in G} T(g)$. Then, to resolve each g we must select two elements from $N(g) \backslash \bigcup_{g \in G} T(g)$, provided that g has at least one ambiguous SNP. If g contains no Xs, we select one element from $N(g) \backslash \bigcup_{g \in G} T(g)$. This implies that $|H^*| = 2|G| - u$, where u is the number of genotypes with no ambiguous SNPs. However, we know that $T(g)$ is nonempty for some g, which means there exists g^1 and g^2 such that $h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^3 = g^2$ for some h^1, h^2, and h^3. If we replace the four haplotypes that resolve g^1 and g^2 with h^1, h^2, and h^3, then we have resolved G with $2|G| - u - 1$ haplotypes, which contradicts the definition of H^*. Hence, H^* contains an element of $\bigcup_{g \in G} T(g)$. ■

Theorem 3 *Assume every g has one or more ambiguous SNPs. Then, $|H^*| = 2|G|$ if, and only if, the neighborhoods of the genotypes together with the set of isolated haplotypes partitions $\mathcal{H}$.*

Proof: ($\Leftarrow$) Let all g have at least one ambiguous SNP, and let H_I be the set of isolated haplotypes. Assume the neighborhoods of the genotypes and H_I partition $\mathcal{H}$. Then, there does not exist an h that resolves both g^1 and g^2, with $g^1 \neq g^2$. Since all genotypes are ambiguous in some SNP, there is no g such that $h \oplus h = g$, for some h. So, two distinct haplotypes must mate to form every genotype. Since $\mathcal{H}$ is partitioned, H^* has exactly two distinct haplotypes from each genotype's neighborhood. Therefore, $|H^*| = 2|G|$.

($\Rightarrow$) Let $|H^*| = 2|G|$, and suppose for the sake of obtaining a contradiction that $T(g) \neq \emptyset$ for some $g \in G$. Then by Lemma 2, H^* contains an element in $\bigcup_{g \in G} T(g)$. Let g^1 and g^2 be such that $h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^3 = g^2$, for some h^1, h^2, and h^3. Let $G' = G \backslash \{g^1, g^2\}$, and let $H' = \bigcup_{g \in G'} N(g)$. Furthermore, let $(H')^*$ be such that

$$|(H')^*| = \min\{|H| : H \subseteq \mathcal{H}, H \text{ resolves } G'\}.$$

Clearly $|(H')^*| \leq 2|G'|$. We know that we can resolve G' by including h^1, h^2, and h^3 in $(H')^*$. Since all three haplotypes might not be required, we have that $2|G| = |H^*| \leq |(H')^*| + 3$. So,

$$\begin{aligned} 2|G| = |H^*| &\leq |(H')^*| + 3 \\ &\leq 2|G'| + 3 \\ &= 2|(|G| - 2)| + 3 \\ &= 2|G| - 1. \end{aligned}$$

Since this is a contradiction, we have that $T(g) = \emptyset$ for all g, and consequently, $N(g^i) \cap N(g^j) = \emptyset$, for all $i \neq j$. Moreover, for every genotype we have that $H_I \cap N(g) = \emptyset$. The result follows since $\mathcal{H} = \left(\bigcup_{g \in G} N(g)\right) \cup H_I$. ■

3 Restricting the Mating Structure

We continue our investigation by exploring the effects of restricting the mating structure. We now constrain our optimization problem so that the maximum number of mates that any h can have is m. A smallest haplotype set that resolves G with this restriction is denoted by H^*_m. We define $\phi(m)$ to be the function that relates m to the cardinality of H^*_m.

Definition 2 *For $(\mathcal{H}, G, \mathcal{E})$, we let $\phi(m)$ be $|H^*_m|$, where no haplotype can have more than m mates.*

If $m = 1$, each haplotype can mate with at most one haplotype (remember that a haplotype can mate with itself). Biologically this means each parent can donate one of two haplotypes to a unique child, so this haplotype cannot be used to form another child. So, for $m = 1$ the neighborhoods of the genotypes in an optimal subgraph are disjoint, and the smallest number of haplotypes that can resolve G is $\phi(1) = 2|G| - u$, where u is the number of genotypes with no ambiguous SNPs. The situation is more complex if $m > 1$, and the main result of this section shows that $\phi(2)$ can be calculated by decomposing a diversity graph into cycles and longest paths.

At some threshold, increasing m does not change the cardinality of H^*_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 H^*_m. Thus, $\phi(m) \geq \phi(m+1)$ for all m, and ϕ is non-increasing. The smallest m such that $\phi(m) = \phi(m+k)$, for all $k \in \mathbb{N}$, is denoted by m^*. So, if $m \geq m^*$, we have that $\phi(m) = \phi(m^*)$. Notice that no haplotype can mate with more than $|G|$ haplotypes, and hence $m^* \leq |G|$. Alternatively, if no haplotype reconciles more than one genotype, $m^* = 1$. The next Theorem shows that if $m^* = |G|$, the number of haplotypes needed to resolve G is either $|G|$ or $|G| + 1$.

Theorem 4 *If $m^* = |G|$, we have that*

$$\phi(m^*) = \begin{cases} |G|, & \text{if } h \oplus h = g \text{ for some } h \in H^*_{m^*}, \\ |G| + 1, & \text{otherwise.} \end{cases}$$

Proof: Let $m^* = |G|$. Then, there exists $h' \in H^*_{m^*}$ such that h' resolves every g. Since $H^*_{m^*}$ resolves G, for each g^i there is an $h^i \in H^*_{m^*}$ such that $h' \oplus h^i = g^i$. We have two cases.

Case 1: Suppose that $h' \oplus h' \notin G$. Then, h' mates with a unique $h^i \in H^*_{m^*} \backslash \{h'\}$ to resolve each $g^i \in G$. Hence, $\phi(m^*) = |H^*| = |G| + 1$.

Case 2: Suppose $h' \oplus h' \in G$. In this case we have that h' mates with $|G| - 1$ haplotypes in $h^i \in H^*_{m^*} \backslash \{h'\}$ to resolve the genotypes in $G \backslash \{h' \oplus h'\}$. Hence, $\phi(m^*) = |H^*_{m^*}| = 1 + (|G| - 1) = |G|$.

■

From Theorem 4 we see that there are situations where calculating m^* solves the Pure Parsimony problem. Unfortunately, further relations between m^* and the diversity graph $(\mathcal{H}, G, \mathcal{E})$ are not clear, and related questions are left for future research.

For the remainder of this section we focus on calculating $\phi(2)$. The key observation for $m = 2$ is that the most complicated subgraphs allowed are cycles and paths. As an example, let $(\mathcal{H}, G, \mathcal{E})$ contain the cycle $h^1, g^1, h^2, g^2, \ldots, g^p, h^1$. Then, the collection of genotypes $\{g^1, g^2, \ldots, g^p\}$ is resolved by $\{h^1, h^2, \ldots, h^p\}$, where both sets have the same cardinality. None of the haplotypes can mate with another haplotype because each is already mating with two haplotypes. Similarly, if $h^1, g^1, h^2, g^2, \ldots, g^p, h^{p+1}$ is a path in $(\mathcal{H}, G, \mathcal{E})$, only the haplotypes

Algorithm to Decompose an acyclic $(\mathcal{H}, G, \mathcal{E})$ into Paths

Step 1: Set $v = 0$ and $(H_v, G_v) = (\mathcal{H}, G)$.

Step 2: Find the largest path in (H_v, G_v), say P_v. If no path exists, set $P_v = \emptyset$.

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

Step 4: Index v by 1.

Step 5: Set $(H_{v+1}, G_{v+1}) = (H_v, G_v) \backslash P_v$.

Step 6: Index v by 1.

Step 7: Go to Step 2.

Table 1: An algorithm that solves the Pure Parsimony problem for acyclic graphs under the restriction that haplotypes can have at most two mates. From Theorem 5 we have that $\phi(2) = |G| + v$ at the completion of the algorithm.

at the end of the path can mate to form another genotype. Unlike cycles, the p genotypes $\{g^1, g^2, \ldots, g^p\}$ are resolved by the $p+1$ haplotypes $\{h^1, h^2, \ldots, h^{p+1}\}$. So, the genotypes in cycles are resolved by an equal number of haplotypes, and the remaining genotypes are resolved along paths, each of which has one more haplotype than genotype. So, if v is the number of paths, the cardinality of the haplotype set that resolves G is $|G| + v$. Consequently, to find an optimal subgraph of $(\mathcal{H}, G, \mathcal{E})$, we first identify the genotypes in cycles and then proceed to resolve the remaining genotypes with the fewest number of paths. The algorithm in Table 1 "decomposes" any acyclic diversity graph (H, G, E) into paths by iteratively finding the longest paths through the unresolved genotypes. The fact that this technique minimizes the number of paths is established in Theorem 5.

Theorem 5 *The algorithm in Table 1 finds an optimal subgraph of the acyclic diversity graph (H, G, E). Moreover, if v is the number of paths found by the algorithm, $\phi(2) = |G| + v$.*

Proof: The case where the genotypes are resolved in cycles is If $|G| = 1$, the algorithm in Table 1 clearly finds an optimal solution. Assume the algorithm finds an optimal subgraph for all diversity graphs such that $|G| = k$. Let (H, G, E) be an acyclic diversity graph with $|G| = k + 1$. Apply the algorithm in Table 1 to (H, G, E). Notice that every path must start and end with haplotypes since every g must be connected to two haplotypes. So, for each path P_i, $|H_i| = |G_i| + 1$. Let $P_1, P_2, ..., P_v$ be the paths in non-increasing length found by the algorithm, and let $P_v = h_{v_1}, g_{v_1}, h_{v_2}, g_{v_2}, \ldots, g_{v_r}, h_{v_{r+1}}$. Remove the last genotype g_{v_r} from G and set $G' = G \backslash \{g_{v_r}\}$. Form a new diversity graph (H, G', E'), where E' is E with the edges incident to g_{v_r} removed.

Case 1: Suppose $P_v \neq h_{v_1} g_{v_1} h_{v_2}$. Then, the algorithm applied to (H, G', E') finds the paths $P_1, P_2, ..., P'_v$, where $P'_v = h_{v_1}, g_{v_1}, h_{v_2}, g_{v_2}, \ldots, g_{v_r - 1} h_{v_r}$—i.e., the last path is missing the last haplotype and genotype. In this case it takes the same number of paths to resolve the genotypes. From the induction hypothesis we know that these paths form an optimal subgraph of (H, G', E'). Adding g_{v_r} back to G' we can do no better than adding one additional haplotype to reconcile g_{v_r}. Moreover, g_{v_r} cannot be added to any of the paths $P_1, P_2, \ldots, P_{v-1}$ since this would violate the fact that each of these is a longest path through the unresolved genotypes. Thus, $P_1, P_2, ..., P_v$, comprise an optimal subgraph of (H, G, E), and $\phi(2) = |G| + v$.

Case 2: Suppose $P_v = h_{v_1} g_{v_1} h_{v_2}$. Then the algorithm applied to (H, G', E') produces the paths $P_1, P_2, \ldots, P_{v-1}$, which form an optimal subgraph of (H, G', E') by the induction hypothesis. Notice that g_{v_1} cannot be added to any of $P_1, P_2, \ldots, P_{v-1}$, as this would violate the fact that these are the longest paths through the unresolved haplotypes. So, adding g_{v_1} back to G' forces us to use a new path to resolve g_{v_1}. Since P_v accomplishes this task, the paths $P_1, P_2, ..., P_v$, form an optimal subgraph. Hence, $\phi(2) = |G| + v$.

■

The statement of Theorem 5 holds for any acyclic diversity graph, and hence, the result is true for $(\mathcal{H}, G, \mathcal{E})$ if it is acyclic. If $(\mathcal{H}, G, \mathcal{E})$ contains a cycle, the algorithm in Table 1 can still be used to find $\phi(2)$, but we need to first remove all the genotypes that can be resolved along

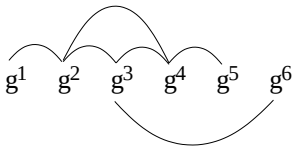

Figure 4: The consecutive genotypes of a path in $(\mathcal{H}, G, \mathcal{E})$ are indicated with an arc. So, there is path that contains the sequence g^2, h', g^4, h'', g^5, but there is no path that contains g^1, h, g^3.

cycles. Unfortunately, a similar technique of finding longest cycles is not guaranteed to identify all of the genotypes that can be resolved in cycles.

The insight from Theorem 5 is that the solutions of a restricted form of the Pure Parsimony problem are representable as a collection of cycles and paths. However, while Theorem 5 highlights the fact that an optimal subgraph can be calculated by first finding the genotypes that can be resolved along cycles and then iteratively removing longest paths, this technique has two shortcomings. First, the longest cycle problem is NP-Hard because it is equivalent to showing that a subgraph is Hamiltonian. So, while the technique describes the nature of a solution, it does not theoretically 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,

$$\begin{aligned} g^1 &= AXBBBB \\ g^2 &= XAXXBB \\ g^3 &= BXAXBX \\ g^4 &= BXXAXB \\ g^5 &= BBBXAB \\ g^6 &= BBXBBA. \end{aligned}$$

To understand this example, notice that a path may contain the sequence g^i, h^i, g^{i+1} if, and only if, there is no SNP where g^i has a value of A or B and g^{i+1} has the other value. So, in the above example no path contains the sequence g^1, h, g^3 because the first SNP of g^1 is an A and the first SNP of g^2 is a B. However, there is a path containing g^1, h, g^2, as there is no SNP where g^1 and g^2 have different values of A and B. If we compare each pair of genotypes in a similar fashion, we find that the paths in $(\mathcal{H}, G, \mathcal{E})$ must pass through the genotypes as indicated in Figure 4. 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 5 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.

4 Conclusion & Future Directions

Using genotypic information of a population, we have explored the fundamental question: what is the minimum amount of diversity in the previous generation that can explain the current population's diversity? Our goal was not to investigate heuristics or model designs like much of the previous literature, but rather, we were interested in studying the mathematical structure underlying this biological question. This endeavor led us to the concept of a diversity graph, which is defined in biological terms. The two main results of this work are

- that any bipartite graph can be extended and labeled to become a diversity graph, and
- that if the mating structure between haplotypes is restricted in a way that does not allow a haplotype to have more than two mates, then the minimum diversity of the previous population can be found by decomposing the diversity graph into cycles and paths.

There are many interesting questions yet to be answered, and we suggest the following:

- How fast does $\phi(m)$ grow? If we had good lower bounds on $\phi(m) - \phi(m+1)$, we could use this to estimate $|H^*|$. For example, if $\phi(m) - \phi(m+1) \geq m\lambda$, then since $\phi(1) = 2|G|$,

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

Table 2: Ways in which the genotypes of $(\mathcal{H}, G, \mathcal{E})$ can be listed in two distinct paths.

we know that

$$\phi(3) \leq \phi(2) - 2\lambda \leq (\phi(1) - \lambda) - 2\lambda = 2|G| - 3\lambda.$$

If we have biological information that indicates no parental haplotype has been passed to more than three children, then we know that the maximum number of haplotypes required to resolve the current population is $2|G| - 3\lambda$.

- We see from Theorem 4 that knowing m^* can solve the Pure Parsimony problem in some cases. Moreover, knowing m^* is beneficial in all cases as this removes many subgraphs from consideration. So, in an integer programming formulation of the Pure Parsimony problem, m^* provides a cut that may help reduce solution times. Finding bounds on m^* is an interesting area of future work.
- Randomized coloring algorithms have been efficient on many classes of graphs, and it may be that finding longest paths and cycles can be thought of as a coloring problem. If so, then these techniques could be used to approximate the algorithm in Table 1, with the hope being that substantial biological models could be addressed.

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] H. Greenberg, W. E. Hart, and G. Lancia. Opportunities for combinatorial optimization in computational biology. Technical report, University of Colorado at Denver, Mathematics Department, Denver, CO, 2002.

[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] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.