Mathematical Approaches to the Pure Parsimony Problem

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

July 29, 2005

Abstract

Given the genetic information of a population, the Pure Parsimony problem asks us to find the least of amount of genetic diversity necessary in the parent population to explain the offspring. The question is of importance to biologists studying genetic disease and has typically been attacked through integer programming. In this paper we study a graph theory formulation of the problem. In particular, we consider certain bipartite graphs with labeled nodes, called diversity graphs, that represent the formation of genotypes from haplotypes. We classify certain graphs that cannot be labeled to become diversity graphs and provide two equivalent notions of diversity graphs. Moreover, we show how to solve the problem in special situations and describe a possible algorithm for solving it completely.

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

---

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

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

[c] University of Colorado, Denver, CO, jsilva2105@msn.com

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

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

1 Introduction

Genes are sequences of DNA in an organism's genome that code for specific traits. While genes are of varying length, there are some particularly small sites on the genome, called single nucleotide polymorphisms (SNPs), in which genetic variation is observed. Biologists study these SNPs in humans so as to better understand genetic disease.

Humans are diploid organisms, meaning that we 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 in a haplotype is in one of two states, denoted by A or B, corresponding to the two distinct nucleotide base pairs in DNA. Each SNP in a genotype is in one of three states, A, B, or X, where the SNP is A (resp, B) if and only if each of the haplotypes that pair to form the genotype have an A (resp. B) in that SNP, and the SNP is X if and only if one of the haplotypes has an A and the other a B in that SNP. See [2] for reference.

Biologists are capable of determining an individual's genotype. It is more difficult and costly to determine the haplotypes that constitute each genotype, but this information is more valuable to biologists. Hence we search for ways to determine haplotypes that give rise to a set of genotypes. Given a set of genotypes, the Pure Parsimony problem asks us to find a smallest set of haplotypes such that every genotype is expressed as a pair of haplotypes; empirical evidence suggests that these minimum solutions are the solutions that naturally occur.

In this paper, we address the Pure Parsimony problem by recasting it in the language of graph theory. We do not solve the whole problem, but rather begin to describe the structure of those graphs that admit a solution to the problem. We also impose a partial ordering on the set of genotypes and haplotypes and find minimum solutions for chains of genotypes. We conclude with a possible algorithm for finding minimum haplotype solutions for any set of genotypes.

2 Notation and Definitions

We replace the alphabet $\{A, B, X\}$ with the set $\{-2, -1, 0, 1, 2\}$ and define $\mathcal{H}_n = \{\mathbf{h}_i = \langle h_{i1}, h_{i2}, \ldots, h_{in}\rangle : h_{ij} \in \{-1, 1\}\}$ to be the set of all haplotypes on n SNPs, and $\mathcal{G}_n = \{\mathbf{g}_i = \langle g_{i1}, g_{i2}, \ldots, g_{in}\rangle : g_{ij} \in \{-2, 0, 2\}$ and $g_{ij} = 0$ for some $1 \leq j \leq n\}$ to be the set of all genotypes on n SNPs. Thus, each genotype is expressed as the sum of the two haplotypes; often there are multiple pairs of haplotypes that sum to form each genotype. We say two pairs of haplotypes $(\mathbf{h}_a, \mathbf{h}_b), (\mathbf{h}_c, \mathbf{h}_d)$ are *unique* if $\{\mathbf{h}_a, \mathbf{h}_b\} \neq \{\mathbf{h}_c, \mathbf{h}_d\}$ and *disjoint* if $\{\mathbf{h}_a, \mathbf{h}_b\} \cap \{\mathbf{h}_c, \mathbf{h}_d\}$ is empty.

Given a set of genotypes $\mathcal{G}'_n \subset \mathcal{G}_n$, we say $\mathcal{S} \subset \mathcal{H}_n$ is a *solution* to $\mathcal{G}'_n$ if, for all $\mathbf{g}_i \in \mathcal{G}'_n$, there exist $\mathbf{h}_k, \mathbf{h}_j \in \mathcal{S}$ such that $\mathbf{g}_i = \mathbf{h}_k + \mathbf{h}_j$. A solution $\mathcal{S}$ to $\mathcal{G}'_n$ is *irreducible* if $\mathcal{S}\backslash\{\mathbf{h}\}$ is not a solution to $\mathcal{G}'_n$ for all $\mathbf{h} \in \mathcal{S}$. We say $\mathcal{S}$ is *minimum* if there exists no solution $\mathcal{S}'$ to $\mathcal{G}'_n$ such that $|\mathcal{S}'| < |\mathcal{S}|$. Clearly,

every minimum solution is irreducible.

In order to use graph theory to approach the Pure Parsimony problem, we need to understand graphical representations of these biological situations. Diversity graphs were first described in [1]. Informally, a diversity graph is a labeled bipartite graph in which one set of nodes represents genotypes, the other set of nodes represents haplotypes that solve them, and the edges represent the parent-offspring relationship between them.

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

- *η and γ are injective,*
- *$\mathcal{W}$ is nonempty,*
- *for each $w \in \mathcal{W}$, there exists some $v \in \mathcal{V}$ such that $(v, w) \in \mathcal{E}$, and*
- *$\mathcal{E}$ has the property that if $(v_1, w) \in \mathcal{E}$, there exists some $v_2 \in \mathcal{V}$ such that $(v_2, w) \in \mathcal{E}$ and $\mathbf{h}_1 + \mathbf{h}_2 = \mathbf{g}$ where $\mathbf{h}_i = \eta(v_i)$ and $\mathbf{g} = \gamma(w)$.*

Requiring that η and γ be injective functions ensures that we represent each haplotype and genotype with exactly one node. The rest of the definition ensures that $\mathcal{H}'_n$ is a solution to $\mathcal{G}'_n$.

3 Forbidden Structures

Because haplotypes mate in unique pairs to form genotypes, the degree of every node in $\mathcal{V}$ must be even. From this restriction alone we see that in the set of all bipartite graphs, there are few that can be labeled to form diversity graphs.

We now describe a necessary condition of diversity graphs:

Theorem 3.1. *Let $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ be a diversity graph. Let $w_1, w_2 \in \mathcal{W}$ such that $w_1 \neq w_2$, and let $\mathcal{V}^* = \{v \in \mathcal{V} : (v, w_1), (v, w_2) \in \mathcal{E}\}$. Then $deg(w_1) \geq 2|\mathcal{V}^*|$ or $deg(w_2) \geq 2|\mathcal{V}^*|$.*

Proof. Assume to the contrary that $deg(w_1) < 2|\mathcal{V}^*|$ and $deg(w_2) < 2|\mathcal{V}^*|$. Let $\mathcal{H}^* = \{\eta(v) : v \in \mathcal{V}^*\}$. Since haplotypes mate in unique pairs, there must be fewer than $|\mathcal{V}^*|$ pairs of haplotypes mating to form each of $\mathbf{g_1}$ and $\mathbf{g_2}$. It follows that there exist some $\mathbf{h_1}, \mathbf{h_2} \in \mathcal{H}^*$ such that $\mathbf{h_1} + \mathbf{h_2} = \mathbf{g_1}$. Similarly, there exist $\mathbf{h_3}, \mathbf{h_4} \in \mathcal{H}^*$ such that $\mathbf{h_3} + \mathbf{h_4} = \mathbf{g_2}$. Suppose now that $g_{1j} = 1$; then $h_{ij} = 1$ for all $\mathbf{h_i} \in \mathcal{H}^*$. It follows that $h_{3j} = h_{4j} = 1$, and therefore $g_{2j} = 1$. Conversely, suppose that $g_{2j} = 1$; then $h_{1j} = h_{2j} = 1$ and $g_{1j} = 1$. Similarly, $g_{1j} = -1$ if and only if $g_{2j} = -1$. Finally, suppose that $g_{1j} = 0$; then $g_{2j} \neq 1$ and $g_{2j} \neq -1$ so $g_{2j} = 0$, and we see that $g_{1j} = 0$ if and only if $g_{2j} = 0$. Since j is an arbitrary SNP, we know that $\mathbf{g_1}$ and $\mathbf{g_2}$ are identical on every SNP, and thus $\gamma(w_1) = \gamma(w_2)$, contradicting the assumption that γ is an injective function. Hence $deg(w_1) \geq 2|\mathcal{V}^*|$ or $deg(w_2) \geq 2|\mathcal{V}^*|$. □

Let $K_{m,n}$ be the complete bipartite graph with node sets of size m and n.

Corollary 3.1. *There exist functions η, γ such that $(K_{m,n}, \eta, \gamma)$ is a diversity graph if and only if $m+n$ is odd and $1 \in \{m, n\}$.*

Proof. Assume that $m+n$ is even; then either $1 \notin \{m, n\}$ or both m and n are odd. If both m and n are odd, then $deg(w)$ is odd for all $w \in \mathcal{W}$, so there exist no η, γ such that $(K_{m,n}, \eta, \gamma)$ is a diversity graph. If $1 \notin \{m, n\}$; then there exist at least two genotypes with the same set of parent haplotypes, and so, by the above theorem, there exist no η, γ such that $(K_{m,n}, \eta, \gamma)$ is a diversity graph.

Now assume that $m+n$ is odd and $1 \in \{m, n\}$. Then choose $\mathcal{W}$ and $\mathcal{V}$ so that $|\mathcal{W}| = 1$, and $|\mathcal{V}|$ is even. Pick γ such that $\gamma(w) = \langle g_1, g_2, \ldots, g_n \rangle$ where $g_i = 0$ for all i and $2^n \geq |\mathcal{V}|$. There are 2^{n-1} disjoint pairs of haplotypes that can mate to form $\mathbf{g}$. Pick $|\mathcal{V}|/2$ of these pairs and let $\mathcal{H}'_n$ be the set of all haplotypes in these pairs. Setting $\eta : \mathcal{V} \to \mathcal{H}'_n$ to be a bijection, we see that $(K_{m,n}, \eta, \gamma)$ is a diversity graph. □

4 Equivalent Representations

Having multiple representations for the problem offers greater insight into possible solutions.

Given a diversity graph, we can look at the adjacency matrix of the underlying bipartite graph. If $\mathcal{H}'_n \subseteq \mathcal{H}_n$ consists of m haplotypes, let H be the $m \times n$ matrix where $(H)_{ij} = h_{ij}$ and $(H)_{ij}$ is the i, j^{th} component of H. Likewise, if $\mathcal{G}'_n \subseteq \mathcal{G}_n$ consists of k genotypes, let G be the $k \times n$ matrix where $(G)_{ij} = g_{ij}$. Let E be the $k \times m$ matrix where $(E)_{ij} = 1$ if $(v_j, w_i) \in \mathcal{E}$, and $(E)_{ij} = 0$ otherwise. Note that the row sums of E must be even by definition of diversity graph.

Consider the $k \times n$ matrix product EH. Without loss of generality, let $(E)_{ia_1} = (E)_{ia_2} = \ldots (E)_{ia_t} = 1$ and all other entries in the i^{th} row of E be zero. Then $(EH)_{ij} = h_{a_1 j} + h_{a_2 j} + \ldots + h_{a_t j}$. Since $(E)_{ia_1} = (E)_{ia_2} = \ldots (E)_{ia_t} = 1$, we know that $(v_{a_s}, w_i) \in \mathcal{E}$ for all $1 \leq s \leq t$. By definition of a diversity graph, it follows that there are $t/2$ disjoint pairs, (v_b, v_c) from $\{v_{a_s} : 1 \leq s \leq t\}$ such that $\mathbf{h_b} + \mathbf{h_c} = \mathbf{g_i}$. Then without loss of generality, let

$$\begin{aligned} h_{a_1 j} + h_{a_2 j} &= g_{ij} \\ h_{a_3 j} + h_{a_4 j} &= g_{ij} \\ &\vdots \\ h_{a_{t-1} j} + h_{a_t j} &= g_{ij}. \end{aligned}$$

Summing the above equations, we find that

$$\tfrac{t}{2} g_{ij} = h_{a_1 j} + h_{a_2 j} + \ldots + h_{a_t j} = (EH)_{ij}.$$

Since t is the sum of the i^{th} row of E, $(t/2)g_{ij}$ is the i, j^{th} entry of the $k \times n$ matrix product $\mathrm{diag}(\frac{1}{2}Ee)G$. From the above argument, we conclude the following:

Theorem 4.1. *Given a diversity graph $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$, the corresponding matrices E, H, and G satisfy the equation $EH = diag(\frac{1}{2}Ee)G$.*

Unfortunately, the converse is not always true; that is, there exist matrices E, H, and G like those above such that $EH = \text{diag}(\frac{1}{2}Ee)G$ but for which there is no corresponding diversity graph. See Figure 1.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix} = (2) \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \tag{1}$$

Figure 1: This matrix equation does not correspond to a diversity graph because no two haplotypes sum to form the genotype. There does, however, exist a labeling such that the edge structure $K_{4,1}$ is a diversity graph. Note that there also exist edge structures that satisfy the matrix equation but for which no such labeling exists.

The matrix equation $EH = \text{diag}(\frac{1}{2}Ee)G$ is not sufficient to ensure a diversity graph because it ignores the need for a mating structure. To remedy this and algebraicly express haplotype pairing that occurs in diversity graphs, we can use a *logical decomposition* of edge matrices.

The *logical join* of a series of matrices is determined by the logical operator "or" over each component of these matrices. The component-wise logical join is defined so that $0 \vee 0 = 0, 0 \vee 1 = 1, 1 \vee 1 = 1$. 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:

$$\bigvee_{1 \leq i \leq s} A_i = A_1 \vee A_2 \vee \ldots \vee A_s = A$$

For example,

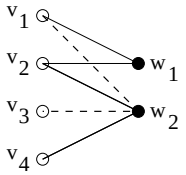
$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \vee \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{2}$$

Figure 2: A logical decomposition

By decomposing the edge matrix of a bipartite graph into matrices whose rows sums are all 2, we express mating structures.

Another way of expressing mating structures is through paired matchings. Let $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ be a bipartite graph such that $\mathcal{W}$ is non-empty and for every $w \in \mathcal{W}$ there is a $v \in \mathcal{V}$ such that $(v, w) \in \mathcal{E}$. We say that $M = \{(v_i, v_j, w_k)\}$ is a *paired matching* on $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ if for every $(v_1, w) \in \mathcal{E}$, there is unique $v_2 \in \mathcal{V}$ such that $(v_2, w) \in \mathcal{E}$ and $(v_1, v_2, w) \in M$. Therefore (v_1, v_2, w) and (v_1, v_3, w) can be elements of M only if $v_2 = v_3$.

For example, consider the bipartite graph below, where $\mathcal{V} = \{v_1, v_2, v_3, v_4\}$, $\mathcal{W} = \{w_1, w_2\}$, and $\mathcal{E} = \{(v_1, w_1), (v_2, w_1), (v_1, w_2), (v_2, w_2), (v_3, w_2), (v_4, w_2)\}$. On this bipartite graph, $M = \{(v_1, v_2, w_1), (v_1, v_3, w_2), (v_2, v_4, w_2)\}$ is a paired matching.



For a paired matching M on $(\mathcal{V}, \mathcal{W}, \mathcal{E})$, let $\mathcal{C}_M$ be the set of all functions $c : \mathcal{V} \to \mathcal{H}_1$ such that for every $v_1, v_2, v_3, v_4 \in \mathcal{V}$ and $w \in \mathcal{W}$, if (v_1, v_2, w) and (v_3, v_4, w) belong to M, $c(v_1) + c(v_2) = c(v_3) + c(v_4)$. As there are finitely many such functions, we index the elements of $\mathcal{C}_M$ by denoting them c_i for $1 \leq i \leq |\mathcal{C}_M|$. For every $c_i \in \mathcal{C}_M$, define the fuction $d_i : \mathcal{W} \to \mathcal{G}_1$ such that $d_i(w) = c_i(v_1) + c_i(v_2)$ for $(v_1, v_2, w) \in M$ and let $\mathcal{D}_M$ be the collection of all such functions. Note that these functions are well-defined because if (v_1, v_2, w) and (v_3, v_4, w) belong to M, then $c_i(v_1) + c_i(v_2) = c_i(v_3) + c_i(v_4)$.

For every $c_i \in \mathcal{C}_M$, let $H(c_i)$ be the set of all (v_j, v_k) such that $j \neq k$ and $c_i(v_j) = c_i(v_k)$. Define $H(M)$ to be the intersection of $H(c_i)$ over all $c_i \in \mathcal{C}_M$. Similary, let $G(c_i)$ be the set of all (w_j, w_k) such that $j \neq k$ and $d_i(w_j) = d_i(w_k)$, and define $G(M)$ to be the intersection of $G(c_i)$ over all $c_i \in \mathcal{C}_M$.

For our paired matching in the example above, consider the functions $c_1, c_2 \in \mathcal{C}_M$ and corresponding $d_1, d_2 \in \mathcal{D}_M$ given below:

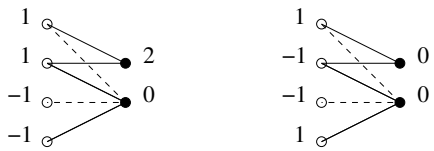


Figure 3: On the left, $c_1(v_1) = c_1(v_2) = 1$ and $c_1(v_3) = c_1(v_4) = -1$. It follows that $d_1(w_1) = 2$ and $d_1(w_2) = 0$. On the right, $c_2(v_1) = c_2(v_4) = 1$ and $c_2(v_2) = c_2(v_3) = -1$, and thus $d_2(w_1) = d_2(w_2) = 0$.

Because $c_1(v_1) = c_1(v_2)$ and $c_1(v_3) = c_1(v_4)$, $H(c_1) = \{(v_1, v_2), (v_3, v_4)\}$. Similarly $H(c_2) = \{(v_1, v_4), (v_2, v_3)\}$. Since $d_1(w_1) \neq d_1(w_2)$ and $d_2(w_1) = d_2(w_2)$, $G(c_1)$ is empty and $G(c_2) = \{(w_1, w_2)\}$. It follows that for this paired matching, both $H(M)$ and $G(M)$ are empty.

Theorem 4.2. *Given a bipartite graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ the following are equivalent:*

1. *There exist functions η and γ such that $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph.*
2. *The edge matrix E has non-trivial dimension and a logical decomposition $E_1, E_2, \ldots, E_s$ where*

- $E_i e = 2e$ *for all* $1 \leq i \leq s$*;*
- *there exists an* H *with distinct rows and the property that* $E_1 H = E_2 H = \ldots = E_s H$*; and*
- *the rows of* $E_1 H$ *are distinct.*

3. $\mathcal{W}$ *is non-empty, for every* $w \in \mathcal{W}$ *there exists* $v \in \mathcal{V}$ *such that* $(v, w) \in \mathcal{E}$*, and there is a paired matching* M *on* $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ *such that both* $H(M)$ *and* $G(M)$ *are empty.*

Proof. (1) $\Rightarrow$ (2): For every $w_j \in \mathcal{W}$, define the set $\mathcal{N}(w_j) = \{v \in \mathcal{V}\text{: } (v, w_j) \in \mathcal{E}\}$. Let E be the edge matrix of the graph $(\mathcal{V}, \mathcal{W}, \mathcal{E})$ and note that because $\mathcal{W}$, and thus $\mathcal{V}$, is non-empty, E has non-trivial dimension.

Let s be the maxmimum degree of all $w \in \mathcal{W}$ and construct the matrices $E_1, E_2, \ldots, E_s$ as follows. Pick $w_j \in \mathcal{W}$ and let v_{a_i} be the i^{th} element of $\mathcal{N}(w_j)$. Because $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph, we know that given $(v_{a_i}, w_j) \in \mathcal{E}$, there is some $v_{b_i} \in \mathcal{V}$ such that $(v_{b_i}, w_j) \in \mathcal{E}$ and $\eta(v_{a_i}) + \eta(v_{b_i}) = \gamma(w_j)$.

For every $1 \leq i \leq |\mathcal{N}(w_j)|$, in the j^{th} row of E_i, put 1's in the ${a_i}^{th}$ and ${b_i}^{th}$ columns and 0's in all the rest. For every $|\mathcal{N}(w_j)| < i \leq s$, put 1's in the ${a_1}^{th}$ and ${b_1}^{th}$ columns and 0's in the other columns of the j^{th} row of E_i.

Note that if the j, k^{th} component of E is 1 then $v_k \in \mathcal{N}(w_j)$; thus there is some $E_i \in \{E_1, E_2, \ldots, E_s\}$ such that the $(E_i)_{jk} = 1$. Also, if $(E)_{jk}$ is 0 then $v_k \notin \mathcal{N}(w_j)$, meaning that for all $E_i \in \{E_1, E_2, \ldots, E_s\}$, $(E_i)_{jk}$ must equal 0. Therefore $E_1, E_2, \ldots, E_s$ is a logical decomposition of E. Note that for all $1 \leq i \leq s$, the row sums of E_i are 2.

Let H be a matrix such that $(H)_{i*} = \eta(v_i)$, where $(H)_{i*}$ is the i^{th} row of H. Because η is an injective function, the rows of H must be distinct. Once again consider $w_j \in \mathcal{W}$. For every $1 \leq i \leq |\mathcal{N}(w_j)|$, $(E_i H)_{j*} = \eta(v_{a_i}) + \eta(v_{b_i}) = \gamma(w_j)$. Similarly for $|\mathcal{N}(w_j)| < i \leq s$, $(E_i H)_{j*} = \eta(v_{a_1}) + \eta(v_{b_1}) = \gamma(w_j)$. So the j^{th} rows of $E_1 H, E_2 H, \ldots, E_s H$ are identical. Therefore $E_1 H = E_2 H = \ldots = E_s H$. Because $(E_1 H)_{i*} = \gamma(w_i)$ and γ is an injective function, the rows of $E_1 H$ are distinct.

(2) $\Rightarrow$ (3): Let M be the set of all (v_i, v_j, w_k) such that for some $E_q \in \{E_1, E_2, \ldots, E_s\}$, $(E_q)_{ki} = (E_q)_{kj} = 1$. Let $(v_i, w_k) \in \mathcal{E}$. Then $(E)_{ki} = 1$ and for some $E_q \in \{E_1, E_2, \ldots, E_s\}$, $(E_q)_{ki} = 1$. By assumption, $E_q e = 2e$, so there exists some $j \neq i$ such that $(E_q)_{kj} = 1$. It follows that $(v_i, v_j, w_k) \in M$. Suppose that there exists v_l such that $(v_i, v_l, w_k) \in M$. Then for some $E_r \in \{E_1, E_2, \ldots, E_s\}$, $(E_r)_{ki} = (E_r)_{kl} = 1$. The k^{th} row of $E_q H$ equals $(H)_{i*} + (H)_{j*}$ and the k^{th} row of $E_r H$ equals $(H)_{i*} + (H)_{l*}$. Since by assumption $E_q H = E_r H$, $(H)_{i*} + (H)_{j*}$ equals $(H)_{i*} + (H)_{l*}$ and thus $(H)_{j*} = (H)_{l*}$. Because the rows of H are distinct, j must equal l. Therefore M is a paired matching.

Let n be the number of columns of H. For $1 \leq j \leq n$, define the fuction c_j such that for every $v_i \in \mathcal{V}$, $c_j(v_i) = (H)_{ij}$. Suppose that for $v_1, v_2, v_3, v_4 \in \mathcal{V}$ and $w_k \in \mathcal{W}$, both (v_1, v_2, w_k) and (v_3, v_4, w_k) are elements of M. Then for some $E_q, E_r \in \{E_1, E_2, \ldots, E_s\}$, $(E_q)_{k1} = (E_q)_{k2} = 1$ and $(E_r)_{k3} = (E_r)_{k4} = 1$. It follows that the k^{th} row of $E_q H$ is $(H)_{1*} + (H)_{2*}$ and the k^{th} row of $E_r H$

is $(H)_{3*} + (H)_{4*}$. By assumption, $E_qH = E_rH$, so $(H)_{1*} + (H)_{2*}$ must equal $(H)_{3*} + (H)_{4*}$. Thus for $1 \leq j \leq n$, $c_j(v_1) + c_j(v_2) = c_j(v_3) + c_j(v_4)$. Therefore $c_j \in \mathcal{C}_M$ for $1 \leq j \leq n$.

Similarly for $1 \leq j \leq n$, define d_j such that for every $w_i \in \mathcal{W}$, $d_j(w_i) = c_j(v_1) + c_j(v_2)$, where $(v_1, v_2, w_i) \in M$. Note that $d_j \in \mathcal{D}_M$. If $(v_1, v_2, w_i) \in M$, then for some $E_k \in \{E_1, E_2, \ldots, E_s\}$, $(E_k)_{i1} = (E_k)_{i2} = 1$. Thus for $1 \leq j \leq n$, $(H)_{1j} + (H)_{2j}$ equals $(E_kH)_{ij}$, which by assumption equals $(E_1H)_{ij}$. Since $(H)_{1j} + (H)_{2j} = c_j(v_1) + c_j(v_2) = d_j(w_i)$, we conclude that for all $w_i \in \mathcal{W}$ and $1 \leq j \leq n$, $d_j(w_i) = (E_1H)_{ij}$.

Let $v_1, v_2 \in \mathcal{V}$ where $v_1 \neq v_2$. The rows of H are distinct, so there exists a column j in which $(H)_{1*}$ and $(H)_{2*}$ differ. Thus $c_j(v_1) \neq c_j(v_2)$. Because $c_j \in \mathcal{C}_M$, (v_1, v_2) cannot be an element of $H(M)$. This holds for all $v_1, v_2 \in \mathcal{V}$, so $H(M)$ is empty. Since $d_j \in \mathcal{D}_M$ and the rows of E_1H are distinct, it follows by a similar argument that $G(M)$ is empty.

(3) $\Rightarrow$ (1): Let $\eta^* : \mathcal{V} \rightarrow \mathcal{H}_n$ be a function such that $\eta^*(v_j) = \langle c_1(v), c_2(v), \ldots, c_n(v) \rangle$, where $n = |\mathcal{C}_M|$. Let $\mathcal{H}'_n$ be the image of $\mathcal{V}$ under η^*, and define $\eta : \mathcal{V} \rightarrow \mathcal{H}'_n$ such that $\eta^*(v_j) = \eta(v_j)$. Suppose that for some $v_j, v_k \in \mathcal{V}$, $\eta(v_j)$ equals $\eta(v_k)$. Then for all $c_i \in \mathcal{C}_M$, $c_i(v_j) = c_i(v_k)$. If j and k are distinct, then (v_j, v_k) is an element of $H(M)$. By assuption we know that $H(M) = \emptyset$, so j must equal k. Thus η is injective.

Similarly, let $\gamma^* : \mathcal{W} \rightarrow \mathcal{G}_n$ be a function such that $\gamma^*(w_j) = \langle d_1(w_j), d_2(w_j), \ldots, d_n(w_j) \rangle$. Let $\mathcal{G}'_n$ be the image of $\mathcal{W}$ under γ^*, and define $\gamma : \mathcal{W} \rightarrow \mathcal{G}'_n$ such that $\gamma^*(w_j) = \gamma(w_j)$. Because $G(M) = \emptyset$, γ is also injective.

Let $(v_1, w) \in \mathcal{E}$. We know that there exists $v_2 \in \mathcal{V}$ such that $(v_2, w) \in \mathcal{E}$ and $(v_1, v_2, w) \in M$. Note that $\gamma(w) = \langle d_1(w_j), d_2(w_j), \ldots, d_n(w_j) \rangle$. Because for all $c_i \in \mathcal{C}_M$, $d_i(w_j)$ equals $c_i(v_1) + c_i(v_2)$, we conclude that $\gamma(w) = \langle c_1(v_1) + c_1(v_2), c_2(v_1) + c_2(v_2), \ldots, c_n(v_1) + c_n(v_2) \rangle$. Hence $\gamma(w) = \eta(v_1) + \eta(v_2)$.

From this we see that for every $(v_1, w) \in \mathcal{E}$ there is a $v_2 \in \mathcal{V}$ such that $(v_2, w) \in \mathcal{E}$ and $\eta(v_1) + \eta(v_2) = \gamma(w)$, and by assumption, $\mathcal{W}$ is non-empty and for every $w \in \mathcal{W}$ there is a $v \in \mathcal{V}$ such that $(v, w) \in \mathcal{E}$. Hence $(\mathcal{V}, \mathcal{W}, \mathcal{E}, \eta, \gamma)$ is a diversity graph.

□

5 Solving Chains

In order to find minimum haplotype solutions for genotype sets with certain characteristics, we impose a partial ordering on our genotype and haplotype sets.

We begin by recalling some elementary terms from lattice theory. A *partially ordered set*, or *poset*, is a set P together with a binary operation $\leq$ such that the following statements are true for all $x, y, z \in P$:

- $x \leq x$ (reflexivity);

- if $x \leq y$ and $y \leq x$, then $x = y$ (antisymmetry);
- if $x \leq y$ and $y \leq z$, then $x \leq z$ (transitivity).

If either $x \leq y$ or $y \leq x$, we say that x and y are *comparable*; if not, we say x and y are *incomparable*. A *chain* is a subset C of a poset P such that x and y are comparable for all $x, y \in C$; an *antichain* is a subset A of a poset P such that x and y are incomparable for all $x, y \in A$.

Given a subset $\{x_1, x_2, \ldots, x_n\} \subset P$, we define the *join* of these elements to be their least upper bound. Note that if $x \leq y$, then $x \vee y = y$.

We now approach the Pure Parsimony Problem through lattice theory. Let $K = \{A, B, X\}$. Let $\leq$ be a binary relation such that $A \leq A, A \leq X, B \leq B, B \leq X$, and $X \leq X$; then K together with $\leq$ is a poset. Define $\mathcal{K}_n = \{\mathbf{k} = \langle k_1, k_2, \ldots, k_n \rangle : k_i \in K \text{ for all } i\}$ and write $\mathbf{k}_a \leq \mathbf{k}_b$ if and only if $k_{aj} \leq k_{bj}$ for all $1 \leq j \leq n$; then $\mathcal{K}_n$ together with $\leq$ is a poset.

Let $\mathcal{H}_n = \{\mathbf{h} = \langle h_1, h_2, \ldots, h_n \rangle : h_i \in \{A, B\}\}$ and $\mathcal{G}_n = \{\mathbf{g} = \langle g_1, g_2, \ldots, g_n \rangle : g_i \in \{A, B, X\}\} \backslash \mathcal{H}_n$ be subsets of $\mathcal{K}_n$, called the set of *haplotypes* and *genotypes* on n SNPs, respectively. Every genotype $\mathbf{g}$ is the join of some pair of haplotypes, and for each of these haplotypes $\mathbf{h}$, $\mathbf{h} \leq \mathbf{g}$. We call $\mathbf{h}$ a *parent* of $\mathbf{g}$, and denote the set of all parents of $\mathbf{g}$ by $P(\mathbf{g})$. A subset of $\mathcal{G}_n$ from which all elements are comparable is a chain of genotypes. For example, the following is a genotype chain of length 4:

$$\{BAXAB, XAXXB, XAXXX, XXXXX\}$$

We now explore solutions to such chains. Let $C = \{\mathbf{g_1}, \mathbf{g_2}, \ldots, \mathbf{g_k}\} \in \mathcal{G}_n$ be a chain of k genotypes such that $\mathbf{g_i} \leq \mathbf{g_{i+1}}$ for all i. Then we have the following:

Lemma 5.1. *Let S be an irreducible solution to the chain $C \subset \mathcal{G}_n$. If $\mathbf{g}' > \mathbf{g}_i$ for all $\mathbf{g}_i \in C$, then S does not solve g'.*

Proof. Assume to the contrary, that S is a solution to g'. Then for some $\mathbf{h}_a, \mathbf{h}_b \in S$, $\mathbf{h}_a \vee \mathbf{h}_b = \mathbf{g}'$. Because S is an irreducible solution to C, each element of S joins to form some elment of C. Thus there exist $\mathbf{h}_{a'}, \mathbf{h}_{b'} \in S$ such that $(\mathbf{h}_a \vee \mathbf{h}_{a'})$ and $(\mathbf{h}_b \vee \mathbf{h}_{b'})$ are elements of C, and without loss of generality, let $(\mathbf{h}_a \vee \mathbf{h}_{a'}) < (\mathbf{h}_b \vee \mathbf{h}_{b'})$.

Since $(\mathbf{h}_b \vee \mathbf{h}_{b'}) < \mathbf{g}'$ and $\mathbf{g}' = \mathbf{h}_a \vee \mathbf{h}_b$, by commutativity of the join operator, $\mathbf{h}_a \vee \mathbf{h}_b \vee \mathbf{h}_{b'} = \mathbf{h}_a \vee \mathbf{h}_b$. Similarly, because $(\mathbf{h}_a \vee \mathbf{h}_{a'}) < (\mathbf{h}_b \vee \mathbf{h}_{b'})$, $\mathbf{h}_a \vee \mathbf{h}_{a'} \vee \mathbf{h}_b \vee \mathbf{h}_{b'} = \mathbf{h}'_b \vee \mathbf{h}_{b'}$, and thus $\mathbf{h}_a \vee \mathbf{h}_b \vee \mathbf{h}_{b'} = \mathbf{h}_b \vee \mathbf{h}_{b'}$. It follows that $\mathbf{g}' = \mathbf{h}_a \vee \mathbf{h}_b = \mathbf{h}_b \vee \mathbf{h}_{b'}$, a contradiction, because $(\mathbf{h}_b \vee \mathbf{h}_{b'}) \in C$ and $\mathbf{g}' > \mathbf{g}_i$ for all $\mathbf{g}_i \in C$. Therefore S does not solve $\mathbf{g}'$. □

Theorem 5.1. *A minimum solution S to C has cardinality $|S| = k + 1$.*

Proof. We first construct a solution to C with cardinality $k+1$. Choose $\mathbf{h} \in \mathcal{H}_n$ so that $\mathbf{h} < \mathbf{g_1}$. Then for every $\mathbf{g}_i \in C$, $\mathbf{h} < \mathbf{g_i}$ and there exists a unique $\mathbf{h}_i \in \mathcal{H}_n$ such that $\mathbf{h} \vee \mathbf{h}_i = \mathbf{g}_i$. It is clear that $S = \{\mathbf{h}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k\}$ is a solution to C, and $|S| = k + 1$.

We now show that there exists no solution to C with cardinality less than $k+1$. The proof proceeds by induction. Certainly the claim is true when $|C| = 1$, and we assume the claim is true when $|C| = j$; that is, a minimum solution to a chain of length j has cardinality $j+1$.

Now let C be a chain of length $j+1$, and let S be a solution to C. Let $C' = \{\mathbf{g_1}, \mathbf{g_2}, \ldots, \mathbf{g_j}\} \subset C$. Because C contains a subchain of length j, we know $|S| \geq j+1$. Assume that $|S| = j+1$. Since $C' \subset C$, S is also a solution to C' and by the inductive assumption, we know that any solution to C' of cardinality $j+1$ is minumum, and thus irreducible. Consider $\mathbf{g}_{j+1} \in C$. For every $\mathbf{g}_i \in C'$, $\mathbf{g}_{j+1} > \mathbf{g}_i$; hence by Lemma 5.1, S does not solve $\mathbf{g}_{j+1}$. Therefore S is not a solution to C. It follows that $|S| \geq j+2$. □

Now assume that $C = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k\}$ is a chain of length k where $\mathbf{g}_2$ does not have exactly two zero components. Then looking at the other end of the spectrum, we have:

Theorem 5.2. *There exists an irreducible solution to C with cardinality $2k$.*

Proof. The proof proceeds by induction. The theorem clearly holds when $k = 1$, and we assume the theorem holds when $k = j$, that is, there exists a solution to C with cardinality $2j$. Now let $|C| = j+1$, and let $S' = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{2j}\}$ be an irreducible solution to $C' = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_j\} \subset C$. Note that by Lemma 5.1, S' is not a solution to $\mathbf{g}_{j+1}$. We need to show that there exists an irreducible solution S to C such that $S' \subset S$ and $\mathbf{g}_{j+1} = \mathbf{h}_{2j+1} \vee \mathbf{h}_{2j+2}$ where $\mathbf{h}_{2j+1}, \mathbf{h}_{2j+2} \notin S'$.

Because $\mathbf{g}_{j+1}$ is the $j+1^{th}$ element in a chain it has at least $j+1$ ambiguous SNPs and thus 2^{j+1} parent haplotypes, i.e. $2^{j+1} \leq |P(\mathbf{g}_{j+1})|$. Since we've made the restriction that $\mathbf{g}_2$ does not have exactly two zero components, $\mathbf{g}_{j+1}$ has more than j+1 ambiguous SNPs, thus $2^{j+1} < |P(\mathbf{g}_{j+1})|$, which is equivalent to $2^j < |P(\mathbf{g}_{j+1})|/2$. For $j \geq 1$, $2j \leq 2^j$, and thus $2j < |P(\mathbf{g}_{j+1})|/2$.

Since $|P(\mathbf{g}_{j+1})|/2$ is the number of pairs of parents of $\mathbf{g}_{j+1}$ and $2j = |S'|$, there is some pair of haplotypes $\mathbf{h}_{2j+1}, \mathbf{h}_{2j+2} \in P(\mathbf{g}_{j+1})$ that is disjoint from S' such that $\mathbf{g}_{j+1} = \mathbf{h}_{2j+1} \vee \mathbf{h}_{2j+2}$. Then $S = S' \cup \{\mathbf{h}_{2j+1}, \mathbf{h}_{2j+2}\}$ is an irreducible solution to C with cardinality $2(j+1)$. □

Corollary 5.1. *There exists an irreducible solution to C with cardinality i for each $k+1 \leq i \leq 2k$.*

Proof. Let i be fixed and let j be such that $i = k+j$ and let $C_1 = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{k-j+1}\}$ and $C_2 = \{\mathbf{g}_{k-j+2}, \mathbf{g}_{k-j+3}, \ldots, \mathbf{g}_k\}$ be subchains of C. By Theorem 5.1 we can find a solution S_1 to C_1 of cardinality $k-j+2$ and by Theorem 5.2 we can find a solution S_2 to C_2 of cardinality $2j-2$. We show by induction that we can, in fact, find disjoint solutions S_1, S_2.

Let $\mathbf{g}_t \in C_2$ and assume that S' is a solution to C' where $C' = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_t\}$ and $|S'| = 2t + j - k$. Now let $C'' = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{t+1}\}$. Since $1 \leq j \leq k$, we know that $2t+j-k \leq 2t \leq 2^t$. As in the proof of Theorem 5.2, $2^t < |P(g_{t+1})|/2$, and thus $|S'| = 2t + j - k < |P(g_{t+1})|/2$, and there is some pair of haplotypes

$\mathbf{h}_a, \mathbf{h}_b \in P(\mathbf{g}_{t+1})$ that is disjoint from S' such that $\mathbf{g}_{t+1} = \mathbf{h}_a \vee \mathbf{h}_b$. Then $S'' = S' \cup \{\mathbf{h}_a, \mathbf{h}_b\}$ is a solution to C'', and it is irreducible by Lemma 5.1.

□

6 Branch and Bound Approach

Despite the insight gained from research on diversity graphs, the problem of solving the Pure Parsimony problem remains beyond our reach for realistic problems. One approach that may be fruitful is a branch and bound algorithm. A formulation of such an algorithm may help in reducing the computational difficulties of finding the fewest necessary haplotypes by omitting portions of the solution space that are known to contain more haplotypes than an already known bound. There is much to be researched in constructing an approach with regards to the optimization problem, but a well structured algorithm accompanied by a tight upper bound on the number of haplotypes necessary to resolve a given genotype set will be beneficial.

Assume $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$ has an arbitrary ordering of $\mathcal{G}$. Let T be a tree and $x_{i,j}$ be decision variables for the i^{th} level of T and the j^{th} node at that level. We begin building T with $x_{0,0}$ as the root node of T and set $x_{0,0} = 0$. Let all nodes $x_{1,1}, x_{1,2}, \ldots, x_{1,j}$ adjacent to $x_{0,0}$ represent all of the possible j haplotype pairings that resolve an initial $g_1 \in \mathcal{G}$. Furthermore, let the nodes adjacent to $x_{1,1}$ be all possible haplotype pairings that resolve $g_2 \in \mathcal{G}$, likewise for all remaining nodes at level 1 of T and for all remaining $n-1$ levels of T. Each path from $x_{0,0}$ to $x_{1,1}, x_{1,2}, \ldots, x_{1,j}$ introduces two haplotypes into the Pure Parsimony problem. As we construct the tree, each new node represents the number of haplotypes introduced that are not in the path prior to that node. For example, in Figure 6, haplotypes AAAA and ABBA are introduced at node $x_{2,1}$, but ABBA is already in the path, therefore $x_{2,1} = 1$. The decision variables take on values as follows:

$$x_{i,j} = \begin{cases} 0 & \text{if 0 haplotypes are introduced;} \\ 1 & \text{if 1 haplotype is introduced;} \\ 2 & \text{if 2 haplotypes are intoduced.} \end{cases}$$

All trees formed from the same $\mathcal{G}$, regardless of genotype order, have s end nodes, $x_{n,j}$, where $1 \leq j \leq s$. In general, $s = \prod_{i=1}^{n} k(g_i)$, where $k(g_i)$ is the number of ambiguous SNPs in g_i. In Figure 6, AXBA, AXXA, ABXX, and AXBX have 1, 2, 2, and 2 ambiguous SNPs respectively. The resulting tree contains $(1)(2)(2)(2) = 8$ end nodes.

Each path from $x_{0,0}$ to $x_{n,q}$ for $1 \leq q \leq s$ leads to an end node and represents an $\mathcal{H}$ that solves $\mathcal{G}$. Consider the sum of all haplotypes along a path to $x_{n,q}$, $\mathcal{S}_q = \sum_{i=0}^{n} x_{i,j}$. Therefore, the program for finding a smallest haplotype set is,

$$z = Min\,(\mathcal{S}_q) \tag{3}$$

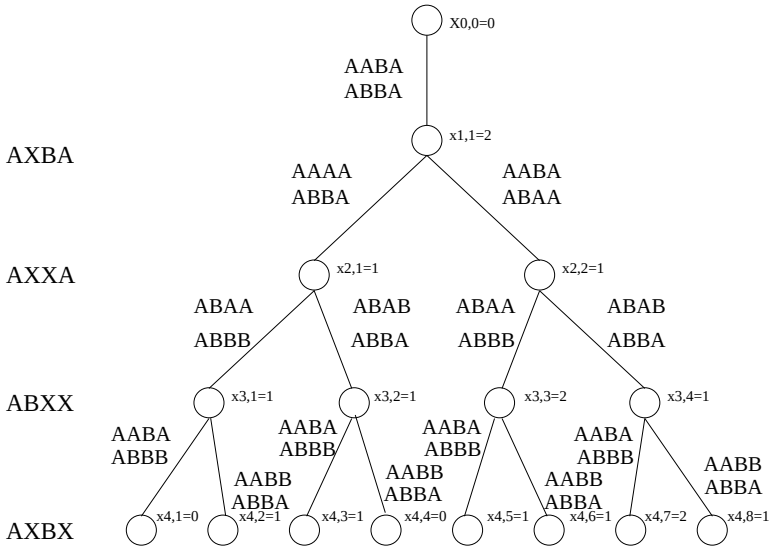


Figure 4: Branch and Bound

In addition we add the following constraint,

$$z \leq u \tag{4}$$

which restricts searching a path in the tree any further once the sum of the $\mathcal{S}_q$ reaches an imposed bound, u. This bound omits large portions of the solution space, reduce computation time; Theorem 5.1 may help provide such a bound. Clark's Rule also provides such a bound u; see [3] for reference. The ability to search all possible solutions to a set of genotypes combined with a tight bound, u, provides an intelligent search of all solutions to the Pure Parsimony problem.

7 Conclusion

We have described various mathematical methods of approaching the Pure Parsimony problem. Diversity graphs, by consisting of both a graph and a labeling of the nodes, provide insight into the problem by clearly distinguishing between the mating structure and the labeling of the haplotypes and genotypes. We have shown that any set of genotypes together with a solution of haplotypes corresponds to a diversity graph, and vice versa. Thus the ability to classify all diversity graphs is important in solving the problem. We have given two equivalent conditions for the existence of a diversity graph, as well as classified one type of structure that cannot be labeled to become a diversity graph.

We have also described the Pure Parsimony problem in terms of partially ordered set, and have determined the cardinality of all irreducible solutions to a

set of genotypes that form a chain. If a similar result can be found for genotypes that form an antichain, it may be possible to find a complete solution to the problem by decomposing the genotype set into chains and antichains. Even if this solution is not minimum, it may provide a tight upper bound for a branch and bound algorithm.

8 Acknowledgements

We would like to thank Dustin Stewart for his suggestion to investigate line graphs and paired matchings, Trinity University for hosting us while conducting our research, and the National Science Foundation for supporting our research.

References

[1] C. Davis and A. Holder. Haplotyping and minimum diversity graphs. Technical Report 73, Trinity University Mathematics, 2003.

[2] H. Greenberg, W. Hart, and G. Lancia. Opportunities for combinatorial optimization in computational biology. 2003.

[3] D. Gusfield. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *J. Computational Biology*, 8(3), August 2001.