Characterizations of Minimum Diversity Graphs

Courtney Davis
Advisor: Allen Holder
Senior Mathematics Project
April 30, 2003

Every child is an amalgamation of its parents' physical characteristics. Inheritance of sequences of deoxyribonucleic acid (DNA) from both parents gives rise to the child's physical traits (phenotype). Due to the huge amount of DNA found in the human genome (about 30,000 genes with thousands of DNA bases per gene), every individual (with the exception of identical twins) carries a unique set of physical characteristics. However, what happens if we reduce the genome size? Suddenly, it is possible for multiple parents to have the same genetic child; hence, placing restrictions on genome size can significantly affect the number of parents able to produce a group of children. We seek to mathematically describe the minimum genetic diversity needed in a parental population to yield the amount of diversity found in a set of children. First, we introduce the biological basis for the mathematical theory that we subsequently develop.

Genes are linear sequences of DNA that code the phenotype of every living organism. Each individual has a maternal and paternal form of every gene; these alternate forms are called alleles, and when paired, make three possible combinations: homozygous dominant, homozygous recessive, or heterozygous. For pedagogical simplicity, assume that the expression of one pair of alleles yields a single physical trait, such as eye color. Since both parents donate an allele for each particular gene, and since multiple genes are often under consideration when examining an individual's physical features, it is helpful to examine the linear collection of alleles that each parent contributes to a child. This linear sequence of alleles is called a haplotype. A genotype, on the other hand, is a linear collection of paired alleles in a child consisting of the two haplotypes donated by the parents. A locus is the position of an allele or pair of alleles on a haplotype or genotype, respectively.

Each allele is expressed as a binary pattern; that is, the allele is either dominant, denoted A, or recessive, denoted B. Paired alleles at the same locus have three possible combinations: homozygous dominant (AA), homozygous recessive (BB), or heterozygous (AB or BA). When describing haplotypes, alleles from a single parent are described; hence, the alphabet for haplotypes is {A, B}. The alphabet for genotypes is {A,B,X}, where A represents the pair AA, B signifies the pair BB, and X denotes pairs AB and BA. We call any genotype locus occupied by an X *ambiguous*.

Two parental haplotypes determine the genotype of an offspring. This means that at every locus, the alleles on two parental haplotypes define the form of the gene at the corresponding locus in the offspring's genotype. Mathematically, let H denote a set of haplotypes and G denote a set of genotypes. Let h_i^j be the allele in locus i on haplotype h^j and g_i^j be the allele in locus i on genotype g^j. (Note that if only one haplotype or genotype is under consideration, the index j is disregarded.) For parent haplotypes h^1, $h^2 \in H$ and offspring genotype $g \in G$, we have the following at each locus:

- 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 three rules. For example, let h^1 = AABAAB and h^2 = ABBABB. Then, $h^1 \oplus h^2 = g$ = AXBAXB.

Parental haplotypes that contribute genetic information to the same offspring's genotype are called *mates*. That is, if $h^1 \oplus h^2 = g$, then h^1 mates with h^2 to form g. Furthermore, we say that h^1 *reconciles* g if $h^1 \oplus h^2 = g$, for some h^2.

Notice that h^1 and h^2 uniquely define a child. To see this, let $h^1, h^2 \in H$, and assume that $g^1, g^2 \in G \ni h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^2 = g^2$. Then, at each locus we have the following possible cases:

Case 1: If $h_i^1 = A$ and $h_i^2 = A$, then $g_i^1 = A$ and $g_i^2 = A$.
Case 2: If $h_i^1 = B$ and $h_i^2 = B$, then $g_i^1 = B$ and $g_i^2 = B$.
Case 3: If $h_i^1 = A$ and $h_i^2 = B$, then $g_i^1 = X$ and $g_i^2 = X$.
Case 4: If $h_i^1 = B$ and $h_i^2 = A$, then $g_i^1 = X$ and $g_i^2 = X$.

So, $g_i^1 = g_i^2, \forall\, i$, and we have that $h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^2 = g^2$ implies $g^1 = g^2$. The implication of this fact is that any two haplotypes mate together to yield exactly one genotype.

Fact 1 $\oplus$ *is a binary operation.*

From this fact we see that the graph in Figure 1 cannot occur in our problem.

Figure 1: Fact 1 eliminates this graph from consideration.

To express pedigree structure between two populations, we introduce a graph that contains edges between genotypes and the haplotypes that are capable of reconciling them.

Definition 2 *A bipartite graph $D = (H, G, E)$ is a diversity graph if*

- *G is nonempty.*
- *$E \subseteq H \times G$ with 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 $h^1 \oplus h^2 = g$.*

For a diversity graph $D = (H, G, E)$, we say H *resolves* G if for all $g \in G$, there exists h^1, $h^2 \in H$ such that $h^1 \oplus h^2 = g$.

Let $N(g) = \{h : (h, g) \in E\}$; that is, $N(g)$ is the *neighborhood* of $g \in G$. We define the *degree* of g, denoted $deg(g)$, as the number of edges emanating from g. The next fact establishes an upper bound on the cardinality of H in a diversity graph $D = (H, G, E)$ by examining $deg(g)$ for each $g \in G$. We denote the set of all possible haplotypes of length n by $\mathcal{H}$. Since $D = (H, G, E)$ is a diversity graph, we know H resolves G. Let locus g_i = X, and let $h^1 \oplus h^2 = g$ for h^1, $h^2 \in H$. Then, if A is in locus h_i^1, B must be in locus h_i^2. So, for every X on every g_i there are two choices for alleles in the corresponding haplotype locus. Hence, for r_g ambiguous loci on genotype g, there are 2^{r_g} potential combinations of alleles, with each combination pertaining to different possible haplotype.

Fact 3 *Let $D = (H, G, E)$ be a diversity graph. Then, $deg(g) \leq 2^{r_g}$ for all $g \in G$, where r_g is the number of ambiguous positions on genotype g.*

It is possible to construct both diversity graphs that satisfy the proper inequality and graphs that satisfy the equality. For example, let $r = 2$. Notice that H = {AAA, ABB} resolves G = AXX; here, $deg(AXX) = 2 < 4 = 2^2$, so $deg(g) < 2^{r_g}$. Conversely, if $|G| = 1$, then for $r > 1$, it is always possible to satisfy the equality.

Note that this fact gives a loose upper bound that regulates the maximum size of H. We can use this to discuss an ordering of $\mathcal{H}$ that establishes haplotype pairs that mate to yield a wholly ambiguous genotype. First, extend the definition of $\oplus$ to mate loci in addition to mating haplotypes. That is, let i be a given locus. Then, $h_i^1 \oplus h_i^2 = g_i$ for $h^1, h^2 \in H$ and $g \in G$ if the following are true:

- 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.

Theorem 4 *Let $\mathcal{H}$ be the set of all haplotypes of length n. Order the elements of $\mathcal{H}$ lexicographically. Then $h^j \oplus h^{(2^r - j+1)} = XX...X$, where $1 \leq j \leq 2^n$.*

Proof: Let $\mathcal{H}$ be the set of all haplotypes of length n. Then, $|\mathcal{H}| = 2^n$. Let h_1^j be the right-most locus in each haplotype. Without loss of generality, let A > B, and let the haplotypes be ordered lexicographically such that $h^j >_L h^{j+1}$. (So h^1 = AA...A.) Now, exchange every A for B and B for A. Notice that it is now true that $h^j <_L h^{j+1} \, \forall \, j$. That is, this new list is the reverse of the previous list.

Let α be any position such that h_α^β = A. That is, we start at an A at h_α^1 and travel down the list β haplotypes to get an A at h_α^β. Recall that reading the list in reverse yields the same results as flipping A's and B's. Hence, if we start at a B at $h_\alpha^{2^n}$ and move up the list β haplotypes, we get to a B at $h_\alpha^{(2^n - \beta + 1)}$. So, due to the symmetry of the ordering, if h_α^β = A, then $h_\alpha^{(2^n - \beta + 1)}$ = B, and if h_α^β = B, then $h_\alpha^{(2^n - \beta + 1)}$ = A. Hence, $h_\alpha^\beta \oplus h_\alpha^{(2^n - \beta + 1)}$ = X. Since this is true for

any locus α and β, we have in general that $h^j \oplus h^{(2^n - j+1)} = \text{XX...X}$. ■

As an example, let $n = 3$, and let A > B. This gives us the list on the left below. Exchanging A's and B's then produces the list on the right.

AAA		BBB
AAB		BBA
ABA		BAB
ABB	⇒	BAA
BAA		ABB
BAB		ABA
BBA		AAB
BBB		AAA

Notice that the right list is simply the left list in reverse order. In addition, each entry on the left mates with the entry directly across from it on the right to form XXX.

To express a lower bound on the number of haplotypes needed to yield a given set of genotypes, we introduce the term H^*.

Definition 5 *H^* is any $H \subseteq \mathcal{H}$ such that $|H^*| = min\{|H| : H \text{ resolves } G\}$.*

So, H^* is a set of haplotypes with the smallest cardinality that resolves G. H^* biologically represents the smallest grouping of parental haplotypes that yields the set of children's genotypes. The primary focus of this work is to investigate properties of an H^* formed from the haplotypes in $\mathcal{H}$. First, notice that we may consider an H^* for every diversity graph.

Fact 6 *Given any diversity graph $D = (\mathcal{H}, G, E)$, there exists an H^*.*

A few notes on H^* should be mentioned. First, H^* need not be unique. This is the primary motivation for examining $|H^*|$ instead of H^* itself. In addition, we know that $D = (\mathcal{H}, G, E)$ is a diversity graph. Let $D^* = (H^*, G, E^*)$ be a subgraph of D, where E^* is the edge set induced by H^*. Since H^* resolves G by definition, it must be true that if $(h^1, g) \in E^*$, then there exists an $(h^2, g) \in E$ such that $h^1 \oplus h^2 = g$.

Fact 7 *$D^* = (H^*, G, E^*)$ is a diversity graph.*

Diversity graphs are bipartite graphs with a specific edge structure. Note that if a genotype has at least one ambiguous locus, the degree of the genotype, denoted $deg(g)$, corresponds to the number of parental haplotypes that reconcile g. Since we know from Fact 1 that haplotypes must come in unique pairs, we would suspect that the degree of g should be even for any g. The following theorem shows that this is indeed the case.

Theorem 8 *If $D = (H, G, E)$ is a diversity graph, then $deg(g)$ is even for all $g \in G$.*

Proof: Let $D = (H, G, E)$ be a diversity graph. Let $g \in G$ be such that $deg(g) = n$. Suppose n is odd. Then, $n \geq 3$. So, $\exists\, h^1, h^2, h^3 \in H \ni h^2 \neq h^3$, $h^1 \oplus h^2 = g$, and $h^1 \oplus h^3 = g$. For each locus h_i^j and g_i we have four possible cases:

Case 1: If $h_i^1 = A$ and $g_i = A$, then $h_i^2 = A$ and $h_i^3 = A$.
Case 2: If $h_i^1 = B$ and $g_i = B$, then $h_i^2 = B$ and $h_i^3 = B$.
Case 3: If $h_i^1 = A$ and $g_i = X$, then $h_i^2 = B$ and $h_i^3 = B$.
Case 4: If $h_i^1 = B$ and $g_i = X$, then $h_i^2 = A$ and $h_i^3 = A$.

Hence $h_i^2 = h_i^3 \,\forall\, i$. So $h^2 = h^3$, which contradicts that h^1 can mate with two distinct $h \in H$ to yield g. Since this means h^1 reconciles g with a unique haplotype, there must be an even number of haplotypes that reconcile every g. However, this contradicts that n is odd. So, $deg(g)$ is even for all $g \in G$. ■

Unfortunately, the converse is not true, as is shown in the counterexample in Figure 2.

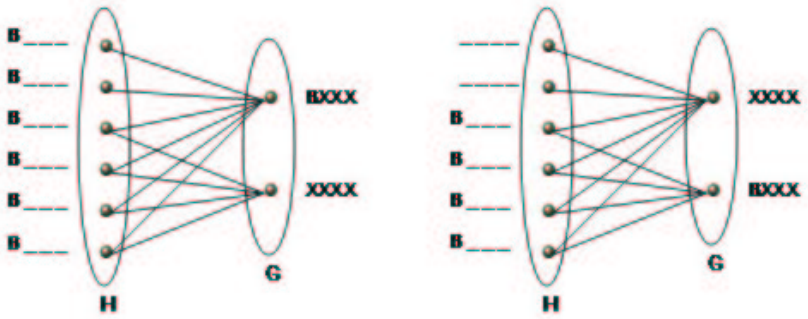

Figure 2: Counterexample. Though the graph is bipartite with $deg(g)$ even for all $g \in G$, $D = (H, G, E)$ cannot be a diversity graph.

The graph in Figure 2 is bipartite and has even degree for every $g \in G$. Since the genotypes in G must be distinct, we let one locus differ. Without loss of generality, we let the genotypes be assigned the values BXXX and XXXX. In the graph on the left, we call the top genotype BXXX and the bottom genotype XXXX. Since the top genotype is connected to every $h \in H$, the first locus of every h must contain a B. However, to satisfy the diversity graph condition, at least two h's connected to the bottom genotype must have an A in the first locus. This gives us a contradiction.

If we reverse the values for the genotype nodes, as done in the graph on the right, we come to a similar contradiction. The lower four haplotypes must contain a B in the first locus to reconcile the bottom genotype. However, in order to allocate a mate to each haplotype, we must then assign an A to the first locus of four other haplotypes. Since H has less than eight nodes, this is not possible; that is, we cannot construct four unique pairs of haplotypes that mate to form the top genotype. Thus, it is not possible to assign values to the nodes in G so that if $(h^1, g) \in E$, then there exists an $h \in H$ such that $h^1 \oplus h = g$. Therefore, every bipartite graph with an even degree for every $g \in G$ is not a diversity graph.

Fortunately, we can extend any bipartite graph (H, G, E) to a diversity graph by including additional haplotypes and edges. Recall that $N(g) = \{h : (h, g) \in E\}$. Let

$$T_g = \bigcup_{g \neq g'} [N(g) \cap N(g')]$$

and $b_g = |T_g|$. So, b_g is the number of haplotypes in the neighborhood of g that are in at least one other genotype's neighborhood. Let $C \in \mathbb{R}$, and define

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

Then, let $\hat{H}_g$ be a set of cardinality $(2b_g - |N(g)|)_+$. We extend each neighborhood with $\hat{H}_g$ such that the new neighborhood of each g is $\hat{N}(g) = N(g) \cup \hat{H}_g$. Now, let $\overline{H} = \left[\bigcup_{g \in G} \hat{N}(g)\right] \cup \left[H \backslash \bigcup_{g \in G} N(g)\right]$. So, $\overline{H}$ contains all the original haplotypes (including those with no edges to G) plus the extensions of each neighborhood. Finally, for all $h \in \hat{N}(g)$, let $(h, g) \in \overline{E}$. (See Figure 3 for a geometric depiction of these definitions.)

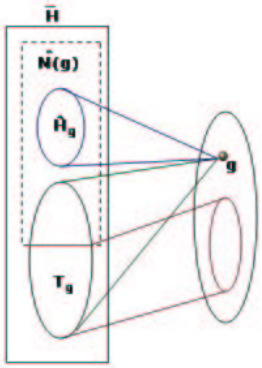

Figure 3: Geometric representation of T_g, $\hat{H}_g$, $\hat{N}(g)$, and $\overline{H}$.

Theorem 9 *Any bipartite graph (H, G, E) with $deg(g)$ even for all $g \in G$ can be extended to a diversity graph $\overline{D} = (\overline{H}, G, \overline{E})$, where*

$$|\overline{H}| - |H| \leq \sum_{g \in G} (2b_g - |N(g)|)_+.$$

Proof: Let (H, G, E) be a bipartite graph such that $|G| = 1$. Let $G = \{g\}$, and assume $deg(g) = 2p$ for some $p \in \mathbb{N}$. Obviously, every haplotype is in the neighborhood of at most one genotype. So, $T_g = \emptyset$ and $b_g = 0$. Then, $\hat{H}_g = \emptyset$, which means $\hat{N}(g) = N(g) \cup \hat{H}_g = N(g)$. Therefore, $\overline{H} = H$ and $\overline{E} = E$. Note that no additional haplotypes needed to be added in this case.

Let $r \in \mathbb{N}$ be such that $2p \leq 2^r$. Set g to be a sequence of r X's; that is, $g = XX...X$. Let the haplotypes be ordered from h^1 to h^{2p}. Set $h_i^1 =$ A for all locus i. For j odd, let h^{j+2} be the next greatest allele permutation lexicographically after h^j, where A > B. For instance, if $r = 5$, then we have the following:

$$\begin{array}{rcl} h^1 & = & \text{AAAAA} \\ h^3 & = & \text{AAAAB} \\ h^5 & = & \text{AAABA} \\ h^7 & = & \text{AAABB} \\ \vdots & & \end{array}$$

Notice that since $2p \leq 2^r$ and since we have assigned only h^j with odd j, we have labeled at most half the haplotypes. For j even, set each locus h_i^j such that $h^{j-1} \oplus h^j = g$, as demonstrated in Theorem 4. In the example above we have the following:

$$\begin{array}{rcl} h^1 \oplus h^2 = g & \Rightarrow & h^2 = \text{BBBBB} \\ h^3 \oplus h^4 = g & \Rightarrow & h^4 = \text{BBBBA} \\ h^5 \oplus h^6 = g & \Rightarrow & h^6 = \text{BBBAB} \\ h^7 \oplus h^8 = g & \Rightarrow & h^8 = \text{BBBAA} \\ \vdots & & \vdots \end{array}$$

Since $deg(g)$ is even, we establish p pairs of mates with this ordering. In addition, $\overline{E}$ has the property that if $(h^j, g) \in \overline{E}$ and j is odd, then $h^j \oplus h^{j+1} = g$ and $(h^{j+1}, g) \in \overline{E}$. Likewise, if $(h^j, g) \in \overline{E}$ and j is even, then $h^j \oplus h^{j-1} = g$ and $(h^{j-1}, g) \in \overline{E}$. Thus, $\overline{D} = (\overline{H}, G, \overline{E})$ is a diversity graph.

To show the bounding argument for this base case, notice that $(2b - |N(g)|) = (0 - 2p) < 0$, which means that $(2b - |N(g)|)_+ = 0$. Hence, $|\overline{H}| - |H| = 0 = (2b - |N(g)|)_+$.

Assume that if $|G| \leq k$, the result holds.

Let (H, G, E) be a bipartite graph such that $|G| = k + 1$. Assume $deg(g) = 2p_g$ for each $g \in G$. Let G_k be a subset of G of cardinality k such that $G = G_k \cup \{\gamma\}$. Then, let H_k be the set of all haplotypes in H that reconcile some $g \in G_k$, and let E_k be such that if $(h, g) \in E$ and $g \in G_k$, then $(h, g) \in E_k$. Note that $D_k = (H_k, G_k, E_k)$ is a subgraph of (H, G, E) with γ and its edges

removed. We know by our assumption that D_k can be extended to a diversity graph $\overline{D}_k = (\overline{H}_k, G_k, \overline{E}_k)$.

Assume $|N(\gamma)| = 2p_\gamma$, and set $T_\gamma = \bigcup_{g \neq \gamma} [N(\gamma) \cap N(g)]$ and $b_\gamma = |T_\gamma|$. Let $\hat{H}_\gamma$ be a set of cardinality $(2b_\gamma - |N(\gamma)|)_+$, and extend the neighborhood of γ by setting $\hat{N}(\gamma) = N(\gamma) \cup \hat{H}_\gamma$. For all $h \in \hat{N}(\gamma)$, we let $\overline{E} = \{(h, \gamma) : h \in \hat{N}(\gamma)\} \cup E_k$. Let $\overline{H} = \hat{H}_k \cup \hat{N}(\gamma)$. We claim that $(\overline{H}, G, \overline{E})$ is a diversity graph.

The following inequalities establish that this extension is bounded appropriately:

$$\begin{aligned}
|\overline{H}| - |H| &= |\hat{H}_k \cup \hat{N}(\gamma)| - |H| \\
&\leq |\hat{H}_k| + |\hat{N}(\gamma)| - |H| \\
&= |\hat{H}_k| + |\hat{N}(\gamma)| - (|H \backslash N(\gamma)| + |N(\gamma)|) \\
&= (|\hat{H}_k| - (|H \backslash N(\gamma)|) + (|\hat{N}(\gamma)| - |N(\gamma)|) \\
&\leq \sum_{g \neq \gamma} (2b_g - |N(g)|)_+ + (|\hat{N}(\gamma)| - |N(\gamma)|) \\
&= \sum_{g \neq \gamma} (2b_g - |N(g)|)_+ + (|N(\gamma) \cup \hat{H}_\gamma| - |N(\gamma)|) \\
&\leq \sum_{g \neq \gamma} (2b_g - |N(g)|)_+ + (|N(\gamma)| + |\hat{H}_\gamma| - |N(\gamma)|) \\
&= \sum_{g \neq \gamma} (2b_g - |N(g)|)_+ + |\hat{H}_\gamma| \\
&\leq \sum_{g \neq \gamma} (2b_g - |N(g)|)_+ + (2b_\gamma - |N(\gamma)|)_+ \\
&= \sum_{g} (2b_g - |N(g)|)_+ .
\end{aligned}$$

Now, add $p_\gamma \in \mathbb{R}$ loci to the right end of every genotype in G_k, and let each of these last p_γ loci contain an A. Then, add p_γ loci to the right end of every haplotype in H_k such that each locus contains an A. Note that if $h^1 \oplus h^2 = g$ for h^1, $h^2 \in H_k$ and $g \in G_k$, then even with these additional loci, it is still true that $h^1 \oplus h^2 = g$. This means that $D_k = (H_k, G_k, E_k)$ is still a diversity graph.

Let γ be the same length as each $g \in G_k$, and let $\gamma_i = \text{X}$ for all loci i. Notice that γ is unique from every other $g \in G$. Also, note that each $h \in T_\gamma$ is already defined since h reconciles some $g^1 \in G_k$. For every $h \in T_\gamma$, set an $h' \in H_\gamma \cup (N(\gamma) \backslash T_\gamma)$ such that $h \oplus h' = \gamma$. Since γ is unique, we know that this will always yield a value for h' that differs from every other haplotype in $\overline{H}$. Note that the last p_γ loci of every h' must all contain B's by the construction.

Let $F = \{h : h$ has no assigned value and $h \in H_\gamma \cup (N(\gamma) \backslash T_\gamma)\}$; that is, F is the set of remaining unlabeled haplotypes. Since $|\hat{N}(\gamma)|$ is even by construction, and since we have assigned pairs of haplotypes, $|F|$ must be even. Let all but the last p_γ loci of $\frac{|F|-1}{2}$ $h \in F$ contain B's. Let all but the last p_γ loci of the remaining $h \in G$ contain A's. Next, let the first and second of these remaining p_γ loci contain an A and B, respectively, for the $\frac{|F|-1}{2}$ elements of F that were assigned B's in all but the last p_γ loci. Conversely, let the first and second of

these remaining p_γ loci contain a B and A, respectively, for the remaining $\frac{|F|-1}{2}$ elements of F. So, F looks like the following:

$$\begin{array}{ll} \{ & \text{BBB...BAB} - - \ldots -, \\ & \text{BBB...BAB} - - \ldots -, \\ & \vdots \\ & \text{BBB...BAB} - - \ldots -, \\ & \text{AAA...ABA} - - \ldots -, \\ & \text{AAA...ABA} - - \ldots -, \\ & \vdots \\ & \text{AAA...ABA} - - \ldots - \quad \} \end{array}$$

Assign the last $p_\gamma - 2$ loci on each element of F as demonstrated in the base case; that is, lexicographically assign the odd elements, and let the even elements mate with the odd.

Notice that if $(h^1, \gamma) \in F$, we have that $\exists\, h^2 \in F \ni h^1 \oplus h^2 = \gamma$. In addition, we have now constructed values for every $h \in \hat{N}_\gamma$ so that every h can mate with an $h' \in \hat{N}_\gamma$ to form γ. Hence, we have by our construction of $\hat{N}_\gamma$ and by our assumption that D_k is a diversity graph that if $(h^1, g) \in \overline{E}$, then there exists an $h^2 \in \overline{H}$ such that $(h^2, g) \in \overline{E}$ and $h^1 \oplus h^2 = g$. Therefore, $\overline{D} = (\overline{H}, G, \overline{E})$ is a diversity graph.

■

Our investigation is broadened by exploring the effects of controlling the highest number of mates that any haplotype is allowed. We now extend our optimization problem to incorporate restrictions on the maximum number of mates that any h can have, denoted m. To this end, we define a function $\phi(m)$ that relates m to the cardinality of the H^*.

Definition 10 *Let $D = (H, G, E)$ be a diversity graph. Then, $\phi(m) = |H^*|$, where no haplotype can have more than m mates.*

At some threshold, increasing m does not change the cardinality of H^*. Intuitively, this makes sense; for instance, if a haplotype is not compatible with more than m genotypes, then allowing this haplotype to mate with $m+1$ haplotypes introduces no additional edges. Thus, $\phi(m) = \phi(m+1)$ for some m. In addition, increasing the number of possible mates that any haplotype is allowed will never cause an increase in H^*. Thus, $\phi(m) \geq \phi(m+1)$ for all m, or equivalently, $\phi(m)$ is non-increasing.

Definition 11 *m^* is the smallest m such that $\phi(m) = \phi(m+k)$ for all $k \in \mathbb{N}$.*

For any natural number $m \geq m^*$, $\phi(m) = \phi(m^*)$. Notice that haplotypes cannot reconcile more than $|G|$ genotypes. Hence, no haplotype can ever mate with more than $|G|$ other haplotypes.

Fact 12 *Since* $\phi(|G|) = \phi(|G| + k)$ *for all* $k \in \mathbb{N}$*, it must also be true that* $\phi(m^*) = \phi(m^* + k)$ *for all* $k \in \mathbb{N}$ *for some* $m^* \leq |G|$.

Thus, we have established an upper bound on m^*. Alternatively, if no haplotype reconciles more than one genotype, then $m^* = 1$. This places a lower bound on the value of m^*.

Fact 13 $1 \leq m^* \leq |G|$.

The example in Figure 4 shows that for some G, $m^* < |G|$. Note that since no other set of haplotypes could resolve G, m^* must equal two. Since $|G| = 3$, $m^* < |G|$.

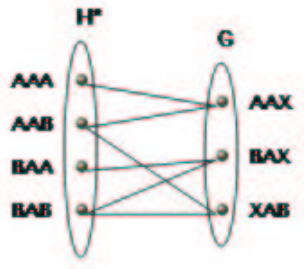

Figure 4: Example showing $m^* < |G|$. Here, $m^* = 2$, whereas $|G| = 3$.

When there is an $h \in H$ that is compatible with every $g \in G$, restricting m^* to a number less than $|G|$ often forces an additional haplotype to be included in H in order to resolve G. For example, let $G = \{\text{XAAA, AAXA, AXAX}\}$, as shown in Figure 5.

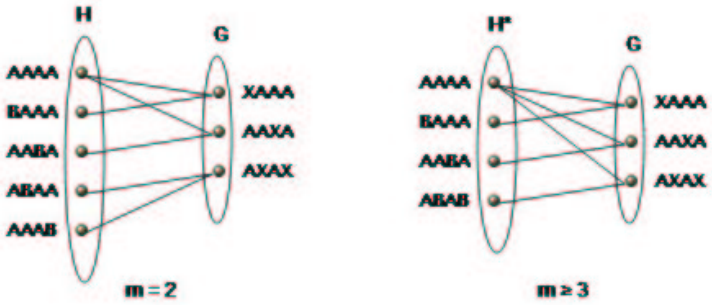

Figure 5: Example showing $m^* = |G|$. Here, $m^* = |G| = 3$.

Haplotype AAAA can reconcile every $g \in G$ if it is allowed at least three mates. However, if only two mates are allowed for any h, then $|H|$ must increase by one. Hence, $m^* = |G|$.

Earlier, we described our problem as finding an H^* for various G's. In certain cases, we can explicitly identify the structure of a diversity graph in relation to the cardinality of a minimum haplotype set. Recall that $N(g)$ is the set of $h \in H$ that reconcile $g \in G$ and $T_g = \bigcup_{g \neq g'} [N(g) \cap N(g')]$.

Lemma 14 *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 G we must select two elements from each $N(g) \backslash \bigcup_{g \in G} T_g$ to resolve $\{g\}$ provided that g has at least one ambiguous locus. If g contains no X's, we select one element from $N(g) \backslash \bigcup_{g \in G} T_g$ to resolve $\{g\}$. This implies that $|H^*| = 2|G| - u$, where u is the number of g with no ambiguous loci. However, we know that T_g is nonempty for some g, which means $\exists\, g^1, g^2 \in G$ such that $h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^3 = g^2$ for some $h^1, h^2, h^3 \in H$. If we include h^1, h^2, h^3 in H^*, then $|H^*| = 2|G| - u - 1$, which contradicts the previously established size of H^*. Hence, H^* contains an element of $\bigcup_{g \in G} T_g$. ■

Theorem 15 *Let all $g \in G$ have one or more ambiguous loci. Then, $|H^*| = 2|G|$ if, and only if, $\{N(g) : g \in G\} \cup \{\{h : deg(h) = 0\}\}$ partitions $\mathcal{H}$.*

Proof: ($\Leftarrow$) Let all $g \in G$ have at least one ambiguous locus, and let $\{N(g) : g \in G\} \cup \{\{h : deg(h) = 0\}\}$ partition $\mathcal{H}$. Then, there does not exist an $h \in H$ such that h reconciles both g_1 and g_2 in G, with $g_1 \neq g_2$. Since all g are ambiguous in some locus, there does not exist an $\overline{h} \in \mathcal{H}$ such that $\overline{h} \oplus \overline{h} = g$ for any $g \in G$. So, two distinct h must mate to form every g. Since $\mathcal{H}$ is partitioned, H^* has exactly two h from each $N(g)$. Therefore, $|H^*| = 2|G|$.

($\Rightarrow$) Let $|H^*| = 2|G|$, and suppose that $T_g \neq \emptyset$ for some $g \in G$. Then by Lemma 14, H^* contains an element from $\bigcup_{g \in G} T_g$. Let $g^1, g^2 \in G$ be such that $h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^3 = g^2$ for some $h^1, h^2, h^3 \in H$. Let $G' = G \backslash \{g^1, g^2\}$, and let $H' = \bigcup_{g \in G'} N(g)$. Let $(H')^*$ be such that

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

. So, we find $(H')^*$ on $(\mathcal{H}, G', E')$. It is obvious from the definition of H^* that $|(H')^*| \leq 2|G'|$. We know that we can resolve G' by adding h^1, h^2, and h^3 back into $(H')^*$. Since all three haplotypes might not be required to resolve G', we know that $2|G| = |H^*| \leq |(H')^*| + 3$. This means the following:

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

Since this is a contradiction, we have that $T_g = \emptyset \, \forall \, g \in G$. Hence, $N(g^i) \cap N(g^j) = \emptyset \, \forall \, i \neq j$. Obviously, $\{\{h : deg(h) = 0\}\} \cap N(g) = \emptyset \, \forall \, g$. Notice that $\{N(g) : g \in G\} \cup \{\{h : deg(g) = 0\}\} = \mathcal{H}$. Hence, $\{N(g) : g \in G\} \cup \{\{h : deg(g) = 0\}\}$ partitions $\mathcal{H}$. ■

One advantage of identifying m^* is that we may use it to elucidate the size of H^* either independently (as with the previous theorem) or through $\phi(m^*)$.

Theorem 16 *Let $m^* = |G|$. Then,*

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

Proof: Let $m^* = |G|$. Then, $\exists \, h' \in H^* \ni h'$ is compatible with every $g \in G$. Since H^* resolves G by definition, there exists $h^i \in H^* \ni h' \oplus h^i = g \, \forall \, g \in G$. We have two cases:

Case 1: Suppose that $h' \oplus h' \notin G$. Then, Fact 1 gives us that h' mates with a unique $h^i \in H$ to reconcile each $g^i \in G$. Hence, $\phi(m^*) = |H^*| = |G| + 1$.

Case 2: Suppose $h' \oplus h' \in G$. From Fact 1 we know that h' mates with a unique $h^i \in H$ to reconcile each $g^i \in G$. So, h' mates with $|G| - 1$ haplotypes in addition to itself to yield the remaining genotypes. Hence, $\phi(m^*) = |H^*| = 1 + (|G| - 1) = |G|$. ■

We would like to identify m^* efficiently from a given diversity graph. Recall from the discussion of Figure 5 that if there is an $h \in H$ that is compatible with every $g \in G$, restricting m^* to less that the number of g that h reconciles often forces an additional haplotype to be included in H in order to resolve G. So, it might make sense for m^* to be the largest number of mates held by any haplotype. Unfortunately, this supposition does not hold true, and a counterexample is supplied in Figure 6. We first develop a function, denoted $f(h)$, that measures the size of the largest subset of G that any $h \in H$ reconciles. Let $D = (H, G, E)$ be a diversity graph such that H is the set of all possible haplotypes that can resolve G. For all $h \in H$, let $S_h = \{L \subseteq G : h \oplus q \in L \text{ for some } q \in H\}$. For all h, let $f(h) = max\{|L| : L \in S_h\}$.

Proposition 17 $m^* = max_h\{f(h)\}$.

The proposition is false, as seen in the following example. Let $G = \{\text{AXAAX}, \text{AXXAA}, \text{AAXXA}, \text{AXAXA}, \text{XAAXA}\}$, as shown in Figure 6. Then, we can find $H_1 \subset H$ such that $|H_1| = 6$ and $max_{h \in H_1} f(h) = f(\text{AAAAA}) = 5$. However, we also can find $H_2 \subset H$ such that $|H_2| = 5 < |H_1|$ and $max_{h \in H_2} f(h) = f(\text{ABAAA}) = f(\text{AAABA}) = 3 < max_{h \in H_1} f(h)$. Hence, for this G, $m^* \leq 3$, while $max_{h \in H} f(h) \geq 5$. Thus, the maximum number of mates needed to resolve G does not directly relate to the maximum quantity of mates with which any given haplotype has the possibility of pairing.

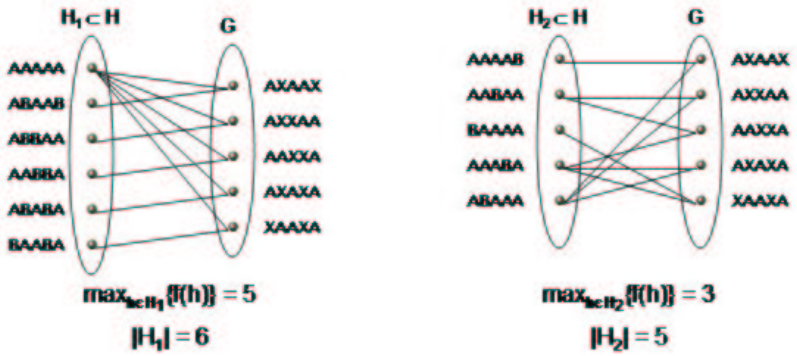

Figure 6: Counterexample demonstrating that $m^* \neq max_{h \in H}\{f(h)\}$. In the left graph, $|H_1| = 6$ and $max_{h \in H_1}\{f(h)\} = 5$. In the graph on the right, $|H_2| = 5$ and $max_{h \in H_2}\{f(h)\} = 3$. Overall, $m^* \leq 3$, while $max_{h \in H}\{f(h)\} \geq 5$.

Despite this counterexample, we have found valuable information about the role of m^* in determining $|H^*|$. Depending on the diversity graph from which it is derived, $|H^*|$ can take many forms; therefore, we have not established a general bound on the minimum number of haplotypes that resolves any G. Nevertheless, using the constructs $N(g)$, m^*, and $\phi(m)$, we have placed bounds on certain H's in relation to a diversity graph $D = (H, G, E)$.

Since diversity graphs are essentially mathematical translations of biological ideas, we can relate many of our results back to the biological basis from whence they arose. Finding an H^* is the mathematical equivalent of finding the minimum genetic diversity necessary in an ancestral population of living organisms to achieve the genetic diversity present in a later population. In the study of genetics, establishing such genetic inheritance relationships aids the development of pedigrees (family trees that contain genetic information).

Pedigrees are useful tools for studying transmission of genetic traits through families. For instance, they provide great insight into the genetic basis for diseases and aid gene therapists in identifying potentially life-threatening disorders before children become symptomatic. In addition, pedigrees can give information about the spread of historic epidemics, such as the black plague, and can give insight into the genetic basis for disease resistance. Thus, any significant

insight into pedigree structures, whether biological or mathematical, can potentially have a cascading effect in the level of biological understanding of genetic events.

Acknowledgements

Much credit and thanks is extended to advisor Allen Holder for his dedication to this project. The problem itself arose from ideas presented in the following paper:

H Greenberg, W Hart, G Lancia. *Opportunities for Combinatorial Optimization in Computational Biology.* Online publication at: http://carbon.cudenver.edu/~hgreenbe/aboutme/pubrec.html.

Further thanks is extended to my committee: Allen Holder, Jeffrey Lawson, and Vadim Ponomarenko.