Rose-Hulman Institute of Technology
Rose-Hulman Scholar

Mathematical Sciences Technical Reports (MSTR) | Mathematics

8-10-2010

G-lattices for an Unrooted Perfect Phylogeny

Monica Grigg

Advisors:
Allen Holder

Follow this and additional works at: http://scholar.rose-hulman.edu/math_mstr

 Part of the Applied Mathematics Commons, Bioinformatics Commons, and the Computational Biology Commons

Recommended Citation
Grigg, Monica, "G-lattices for an Unrooted Perfect Phylogeny" (2010). *Mathematical Sciences Technical Reports (MSTR)*. 28.
http://scholar.rose-hulman.edu/math_mstr/28

MSTR 10-07

This Article is brought to you for free and open access by the Mathematics at Rose-Hulman Scholar. It has been accepted for inclusion in Mathematical Sciences Technical Reports (MSTR) by an authorized administrator of Rose-Hulman Scholar. For more information, please contact weir1@rose-hulman.edu.

G-lattices for an Unrooted Perfect Phylogeny

Monica Grigg

Adviser: Allen G. Holder

Mathematical Sciences Technical Report Series
MSTR 10-07

August 2, 2010

Department of Mathematics
Rose-Hulman Institute of Technology
http://www.rose-hulman.edu/math

Fax (812)-877-8333 **Phone (812)-877-8193**

G-lattices for an Unrooted Perfect Phylogeny

Monica Grigg
Faculty Sponsor: Allen Holder

August 2, 2010

Abstract

We look at the Pure Parsimony problem and the Perfect Phylogeny Haplotyping problem. From the Pure Parsimony problem we consider structures of genotypes called g-lattices. These structures either provide solutions or give bounds to the pure parsimony problem. In particular, we investigate which of these structures supports an unrooted perfect phylogeny, a condition that adds biological interpretation. By understanding which g-lattices support an unrooted perfect phylogeny, we connect two of the standard biological inference rules used to recreate how genetic diversity propagates across generations.

1 Introduction

DNA (deoxyribonucleic acid) encodes the genetic information of an organism. In a diploid organism, such as a human, the DNA is a combination of two chromosome copies, one from each parent. This combination of the two copies is defined as a genotype and is a pair of two haplotypes. Genotype information is easier and cheaper to obtain for a population. However, it is more biologically meaningful to have haplotype information. This is because haplotype information can describe the mutations that can occur from generation to generation and diseases that occur in populations. The use of a full Haplotype Map would prove extremely useful in order to look at a population that can explain the complex genetic diseases. For example, the haplotype information can describe which genetic diseases a population is prone to. The HapMap project [2] is an international effort to understand the similarities of genes between populations, which gives information about the health, diseases, and effects of specific medications. Using the genotype information to obtain haplotypes relies on an inference rule. Two of the inference rules are: the Pure Parsimony problem and Perfect Phylogeny Haplotyping problem. Both give methods to calculate the possible haplotypes that pair together to create a population of genotypes.

2 Biological Notation

A *SNP*, single-nucleotide polymorphism, is a sequence in the genetic code that has a variation, and a *haplotype* is a collection of SNPs. A *genotype* is a pair of haplotypes. When looking at the genetic information, we look at the individual sites of the genotypes and haplotypes. The terms site, SNP, and position are to be used to describe the same genetic information provided by a haplotype; terms are used interchangeably.

The problem we consider consists of a population of m genotypes, where each genotype is represented as a vector of length n. Let H be the set of haplotypes $H = \{0,1\}^n$, and let G be the set of genotypes, $G = \{0,1,2\}^n$. If the position in the genotype vector has a 0 or 2 then the chromosome sites are the same, and the position is called *homozygous*. If a SNP is homozygous then we know which haplotype SNP is necessary and there is only one option for the pair of haplotypes. If the site has a value of 1, then it is *heterozygous*. If a SNP in the genotype is a 1, then this means that there are two possibilities for the haplotype solutions to differ, thus the data is ambiguous. The addition for haplotypes is simple and is component-wise addition. For example, let $h = (1,0,0,1),$ and $h' = (1,1,0,0)$

$$h + h' = (1,0,0,1) + (1,1,0,0) = (2,1,0,1) = g,$$

where g is the genotype that is produced by the haplotypes h and h'.

3 Pure Parsimony Problem

3.1 Introduction

A parsimonious solution is one with few haplotypes, and the Pure Parsimony problem is to calculate the minimum number of haplotypes needed to create a genotype set. The Pure Parsimony problem, PP problem, focuses on finding an optimum solution to satisfy the genotypes in a given population. In [4] the goal was to create a polynomial bound on the PP problem and to determine conditions under which the solution is optimized. Holder and Langley explain the method of attacking an NP-hard problem that utilizes the least number of haplotypes required to satisfy the genotypes in substructures called *g-lattices*. These g-lattice structures are constructed based upon a partial order, denoted by $\preceq$. This is the component-wise comparison of genotype sites. Other methods have been developed for the PP problem, but are not discussed in this paper.

3.2 Notation

The component-wise comparison of the SNPs of the genotypes is defined by $1 \preceq 0$, $1 \preceq 2$, $0 \preceq 0$, and $2 \preceq 2$. SNP values of 0 and 2 are not comparable. Figure 1 demonstrates a population of genotypes decomposed into g-lattices.

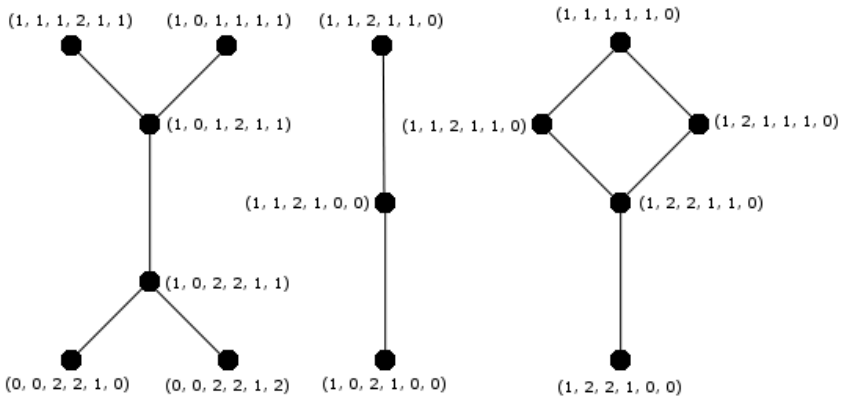

Figure 1: These g-lattice structures describe a possible population of genotypes that exist in the set of genotypes G. Note that the g-lattices do have unique genotypes present and do not reuse the genotypes in the multiple g-lattices. With the component-wise comparison, once an ambiguous site is in place, a 1, then the 1 will bubble up to the top.

Definition 1 *A g-lattice is a chain if and only if the genotype set can be component-wise compared in a single path. That is $g \preceq g'$ for all g and g' in the population.*

The PP problem can be bounded. These bounds allow us to know the minimum solution necessary to satisfy the population of genotypes if they form a chain. Further , these bounds allow us to make conclusions about the results on which g-lattices support unrooted perfect phylogenies.

Theorem 1 (Blain et al [1]) *Suppose G is a collection of m genotypes that form a chain under $\preceq$. Then a minimum solution to G has size $m + 1$ if the minimal element in the chain has at least one heterozygous SNP. Otherwise a minimum solution has size m.*

Theorem 2 (A. Holder and T. Langley [4]) *Let G be a collection of m genotypes and suppose that q minimal elements of G have at least one heterozygous SNP. Then no more than $m + q$ haplotypes are needed to resolve G.*

4 Perfect Phylogeny Haplotyping Problem

The concept of Perfect Phylogeny, also known as the *basic coalescent model*, describes a rooted tree of haplotypes. The tree describes the evolution of the haplotypes through the understanding of changes from generation to generation,

such as genetic mutations. In this model we assume no recombination, which means is that each haplotype has one ancestor in a sequence. The Perfect Phylogeny Haplotyping (PPH) problem is an inference rule that finds a set of haplotypes that satisfy the population of genotypes and that form a perfect phylogeny for some ancestral vector.

4.1 Notation

Let M' be a binary matrix (a matrix containing only zeros and ones), of dimension $2m \times n$, that contains the set of all haplotypes that combine to be the solution of genotypes from a population. For example, if a population of genotypes contains two genotypes of length five, then the matrix M' dimensions would be 4×5. Let V be a binary n-vector that is defined as the *ancestor vector*, which describes the solution to the perfect phylogeny. In the case of an *unrooted perfect phylogeny* this ancestor vector is assumed but unknown, and in the case of an *rooted perfect phylogeny* the ancestor vector is known. Let T be a matrix that is $2m \times 1$ matrix that represents the rooted tree to the perfect phylogeny created by the haplotypes in the matrix M'. Thus a rooted tree can be described as the multiplication of a matrix times a vector:

$$M'_{2m,n}V_{n,1} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,n} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{2m,1} & h_{2m,2} & \cdots & h_{2m,n} \end{bmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = T.$$

Definition 2 (Gusfield [5]) *Let M be an $2m$ by n binary matrix. Let V be an m-length binary vector, called the ancestor vector. A perfect phylogeny for M and V is a rooted tree T with exactly $2m$ leaves that obeys the following properties:*

1) *Each of the $2m$ rows labels exactly one leaf of T, and each leaf is labelled by one row.*
2) *Each of the n columns labels exactly one edge of T.*
3) *Every interior edge (one not touching a leaf) of T is labelled by at least one column.*
4) *For any row i, the value $M(i, j)$ is unequal to $V(j)$ if and only if j labels an edge on the unique path from the root to the leaf labelled i. Hence, that path is a compact representation of row i.*

These properties describe a rooted tree that support a perfect phylogeny. The tree describes the mutations that occur between the haplotypes. To better understand the definition of a perfect phylogeny, here are two examples. One is of a rooted perfect phylogeny and the other is not a rooted perfect phylogeny. Both use the same set of haplotypes but will have different ancestral vectors V to show that not all vectors will support a rooted tree for a set of haplotypes.

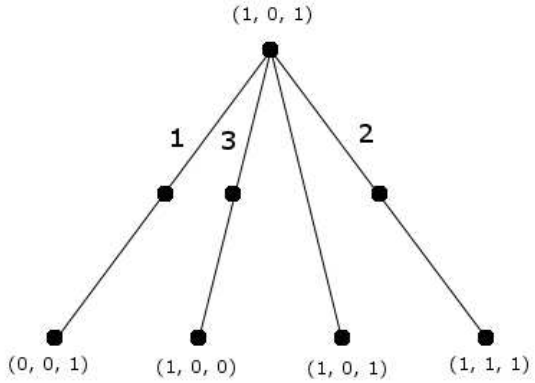

Figure 2: This is a rooted perfect phylogeny. The tree describes the mutations that have occurred from generation to generation of haplotypes.

Let

$$M' = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

and let

$$V = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

Then the set of haplotypes produce the rooted tree in Figure 2.

We see from Figure 2 that a rooted tree is created by labelling the columns of M' as 1, 2, 3. To determine which leaf should be labelled by an edge, we consider each of the sites in the haplotypes from M' compared to the corresponding sites in the ancestor vector, V. For example, the leaf (0, 0, 1) cannot be labelled by the columns of 2 or 3, and thus must be labelled by column 1. The leaf (1, 0, 0) can not be labelled by 1 or 2, which means it must be labelled by 3. (1, 0, 1) displays no mutations from the ancestor vector and does not need to be labelled by a column. While (1, 1, 1) has a mutation in site 2, thus it must be labelled by a 2.

Let M' be equal to the same matrix above, and let

$$V = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

In order to create a rooted tree of haplotypes, we first need to look at the SNPs in the which possible mutations might have occurred from the ancestor vector.

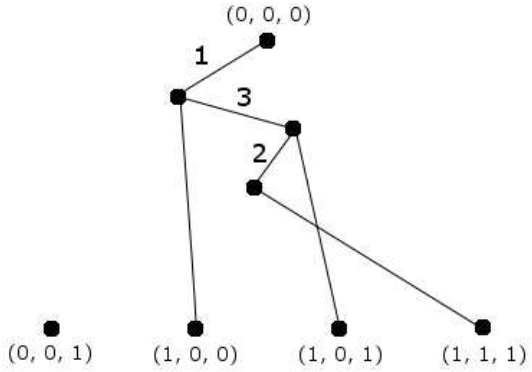

Figure 3: This is not a rooted tree that supports the definition of a perfect phylogeny.

When looking at the haplotype of (0, 0, 1) we see that there is a mutation in the third site, therefore it must be labelled by a 3. (1, 0, 0) has mutated in the first site, and must be labelled by a 1. (1, 0, 1) has mutations in the first and third sites and (1, 1, 1) must be labelled by a 1, 2, and 3. Since no tree can support the needs of the perfect phylogeny, then this particular ancestor vector is incorrect for the set of haplotypes. As you can see from Figure 3 the leaf (0, 0, 1) is not labelled by any of the columns.

Definition 3 (Gusfield [3]) *A complete-pair-matrix (CP matrix) is a matrix containing two columns with rows containing the elements of { (0, 0), (0, 1), (1, 0), (1, 1)}.*

Theorem 3 (Gusfield [3]) *A $2m \times n$ matrix M' defines an unrooted perfect phylogeny if and only if no submatrix $M'[*, (j_1, j_2)]$ formed by selecting the two columns j_1, j_2 is a complete-pair-matrix.*

This theorem shows that if the submatrix $M'[*, (j_1, j_2)]$ has the

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix},$$

then the set of haplotypes does not support an unrooted perfect phylogeny. Below is a portion of a proof of this theorem.

Proof:

$\Leftarrow$ In order to show that the set of haplotypes $H' = \{(0,0),(0,1),(1,0),(1,1)\}$ does not support a perfect phylogeny, we need to show that all of the

possible ancestor vectors cannot create a rooted tree of haplotypes. The possible ancestor vectors are

$$V = \binom{0}{0}, \binom{0}{1}, \binom{1}{0}, \binom{1}{1}.$$

When constructing a rooted tree, each haplotype is a leaf of the tree. The two columns of the matrix M' are labelled by a one and two respectively. Each leaf must be labelled by the columns in which the mutation site has occurred.

Case 1: Without loss of generality, let the ancestor vector $V = \binom{0}{0}$. (0, 0) has no mutations from the ancestor vector. (0, 1) has a mutation in the second site and must be labelled by column 1. (1, 0) has a mutation in the first site and must be labelled by column 2. (1, 1) has mutations in both the first and second sites and must be labelled by a 1 and 2. There is no rooted tree that can be created to satisfy the requirements of the labelling. Therefore this particular ancestor vector does not support a rooted tree of haplotypes.

Case 2-4: These cases have similar arguments to case 1.

Therefore we know that the set of haplotypes does not satisfy the requirements to be defined as a perfect phylogeny.

$\Rightarrow$ We know that if M' is an unrooted perfect phylogeny then there exists a rooted tree of haplotypes from the matrix M'. The tree must support the definition of a perfect phylogeny. Suppose that if M' supports a rooted perfect phylogeny then the there exists the haplotypes {(0, 0), (0, 1), (1, 0), (1, 1)}. From the arguments above, we know that there does not exist a ancestor vector V that satisfies the set of haplotypes. Thus we know that if the matrix M' supports an unrooted perfect phylogeny then there can not exist the submatrix.

Definition 4 *Let there be a minimum of three genotypes g, g', g'' that compose a g-lattice. The g-lattice is described to be a tent structure if and only if $g \preceq g'$ and $g'' \preceq g'$.*

A population of genotypes is said to be a *cascading additive triple* if the genotypes are defined as a tent structure and there exists three SNP pairs $(j_1^k, j_2^k), k = 1, 2, 3$ such that

$$((g_{j_1}^{k_1}, g_{j_2}^{k_1}) + (g_{j_1}^{k_2}, g_{j_2}^{k_2})) \bmod 2 = (g_{j_1}^{k_3}, g_{j_2}^{k_3})$$

In this definition, the cascading describes a g-lattice structure that is tent like. This definition can be applied to a population of genotypes larger than or equal

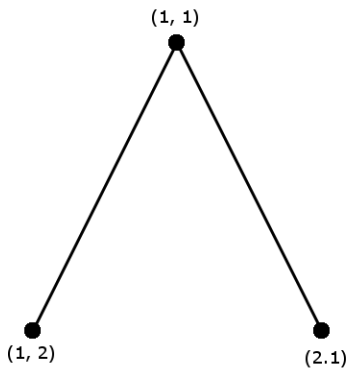

Figure 4: Tent structure comprised of three genotypes based upon a component-wise comparison.

to three, and those genotypes can be of m-length. However, it is only necessary to look at two SNPs at a time to decide if the population supports an unrooted perfect phylogeny. For example, let the population of genotypes be $G' = \{(1,1),(1,2),(2,1)\}$, then the addition under modulus two is:

$$((1,1)+(1,2)) \bmod 2 = (0,1)$$

$$((1,1)+(2,1)) \bmod 2 = (1,0)$$

$$((2,1)+(1,2)) \bmod 2 = (1,1).$$

We know in order for a g-lattice to be a cascading additive triple the genotypes must satisfy the addition modulus two and support a tent structure. From Figure 4 we see that the set does support the definition of a tent structure. Thus, we know that this population of genotypes is a cascading additive triple since it satisfies the addition and is a tent structured g-lattice.

The *obstructed set of haplotypes*, also known as the *forbidden haplotype pairs*, describes a set of haplotypes that when no submatrix of this form is found in M' implies that the set supports an unrooted perfect phylogeny. The *obstructed set of haplotypes* is

$$\{(0,0),(0,1),(1,0),(1,1)\} = \{0,1\}^n.$$

These haplotypes result in the following *obstructed set of genotypes*, which represents the possible genotypes that can be formed by $\{0,1\}^n$. This set is

$$\{(1,0),(0,1),(1,1),(1,2),(2,1)\},$$

and is referred to as the *forbidden genotype pairs*. The interchangeable language is used because when these sets are found within the population, they determine directly if a population supports an unrooted perfect phylogeny. For example

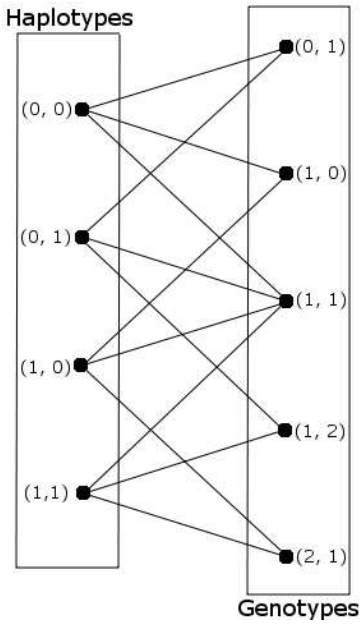

Figure 5: This image describes the pairing of the forbidden haplotype pairs to create the set of forbidden genotypes. Notice that the set of four haplotypes creates six genotypes, however only five of the genotypes are unique. That is, $(0, 0) + (1, 1) = (0, 1) + (1, 0) = (1, 1)$. This is shown above, but for simplicity, we will only list (1, 1) once with the understanding that there are two possible solutions to obtain this genotype.

consider a g-lattice structure in a tent shape with the genotypes of {(1, 0), (2, 1), (1, 1)}. These genotypes come from the ordered pair of genotypes set and the haplotypes that result in the genotype set is $\{(1,0),(1,1),(0,0)\}$. Since the set of haplotypes does not require the fourth haplotype of $(0,1)$, the population of genotypes supports an unrooted perfect phylogeny.

5 Results

The CP matrix lists the four elements that imply that the haplotypes do not describe an unrooted perfect phlyogeny. In order for a chain of genotypes to not support an unrooted perfect phylogeny there needs to be a minimum of three genotypes to allow for the pairings of the four forbidden haplotypes pair to be present as a submatrix of M'. In a chain the lower bound on the pure parsimony problem [4] requires the minimum of haplotypes is $m + 1$, m is the number of genotypes in the population set. Therefore for the set of haplotypes to not

support an unrooted perfect phylogeny, the number of genotypes required are greater than or equal to three. This means there will be at least four haplotypes that generate the genotypes.

Theorem 4 *If the genotypes form a chain under $\preceq$, then the population of genotypes supports the definition of a unrooted perfect phylogeny.*

Proof: Let G be the set of genotypes that form a chain structured g-lattice. Let M' be a binary matrix (a matrix composed of 0 and 1) of haplotypes that satisfy the pure parsimony problem of G. We show that if M' is not an unrooted perfect phylogeny then there exists

$$M'[(i_1, i_2, i_3, i_4), (j_1, j_2)] = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

The genotypes that are possible to create this submatrix in M' are $M'[i_k, (j_1, j_2)] \in \{(1,1), (1,2), (0,1), (2,1), (1,0)\}$. Since the genotypes form a chain, there can be at most two elements from the forbidden genotype pairs. This means there are at most three elements from the forbidden haplotype pairs, and hence, a chain must support a perfect phylogeny.

From this result we see that, a chain structured g-lattice always supports an unrooted perfect phylogeny, and we know that the solution to the pure parsimony problem and the Perfect Phylogeny Haplotying problem are the same solution.

Theorem 5 *If the g-lattice structure only contains a cascading additive triples, then the population of genotypes supports an unrooted perfect phylogeny.*

Proof: From the forbidden genotype pairs we can only compose four tent-structured g-lattices that satisfy the requirement of being a cascading additive triple. That is, the population sets {(1, 0), (1, 2), (1, 1)}, {(1, 0), (2, 1), (1, 1)}, {(1, 0), (0, 1), (1, 1)}, and {(1, 2), (2, 1), (1, 1)}. From Figure 6 we can see that the cascading additive triples support perfect phylogenies. The haplotypes that satisfy the solution to the genotypes when put into a matrix M' do not create a submatrix with the elements {(0, 0), (0, 1), (1, 0), (1, 1)}. Therefore, if the g-lattice structure only contains cascading additive triples, then the population supports an unrooted perfect phylogeny.

An extension of Gusfield's theorem that describes if a set of haplotypes

$$\{(0,0), (0,1), (1,0), (1,1)\}$$

is to consider the minimum number to satisfy the various pairings of haplotypes to ensure that an unrooted perfect phylogeny is not supported by the forbidden genotype pairs.

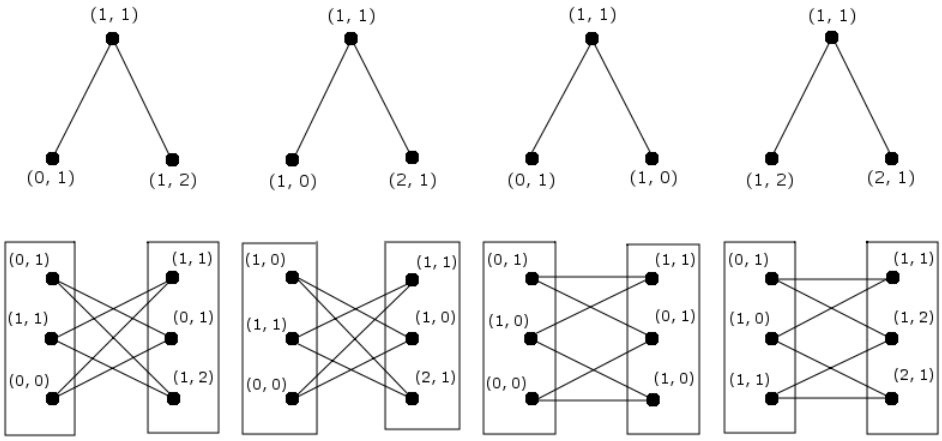

Figure 6: These are the four possible cascading additive triples that are constructed using the forbidden genotypes pairs.

Theorem 6 *There are four or more forbidden genotype pairs in the same two column SNPs if and only if the population of genotypes does not support an unrooted perfect phylogeny.*

Proof:

$\Rightarrow$ If there are four or more forbidden genotype pairs in the same two column SNPs then the population of genotypes does not support an unrooted perfect phylogeny.

The minimum number of haplotypes required to generate four forbidden genotype pairs is four. Hence the four haplotypes create the forbidden haplotype pairs. These four haplotypes make up a submatrix of M' and therefore the set of haplotypes do not support an unrooted perfect phylogeny.

$\Leftarrow$ If the population of genotypes does not support an unrooted perfect phylogeny, then there are four or more forbidden genotype pairs in the same two column SNPs.

We know that if the population of genotypes does not support an unrooted perfect phylogeny then there exists a submatrix of $M'[*, (j_1, j_2)]$, where (j_1, j_2) are the complete-pair-matrix. The minimum number of forbidden genotype pairs created by the four forbidden haplotype pairs is four. Therefore an unrooted perfect phylogeny implies that there are four or more forbidden genotype pairs in the population.

Theorem 7 *If no tent structures are present in the g-lattice substructures, then the population of genotypes supports an unrooted perfect phylogeny.*

Proof: If there are no tent structures present, we know that there does not exist g, g', g'' such that $g \preceq g'$ and $g'' \preceq g'$. With this information, we know that there cannot be four or more forbidden genotypes within the population since there is no tent structures. We also know that there cannot be four or more forbidden genotypes since the branching up structure acts like a chain. This implies that there can only be two forbidden genotype pairs. Since there are only two forbidden genotype pairs are contained in the population of genotypes, we know that the population supports an unrooted perfect phylogeny since the complete-pair-matrix will not be able to be found as a submatrix of M'. Therefore, if there are no tent structures present in the g-lattice, then the population of genotypes supports an unrooted perfect phylogeny.

This leads to the result that if a g-lattice is upward branching, then we can construct an unrooted perfect phylogeny with no more then $m + 1$ haplotypes.

6 Conclusions and Future Work

The forbidden genotype pairs generated from the forbidden haplotype pairs allow us to search easily which pairings will allow the population of genotypes to support an unrooted perfect phylogeny. By understanding which haplotype pairings do not result in supporting an unrooted perfect phylogeny we can detect errors that may have occurred when generating the genotypes. For the Haplotype Map Project, the data has numerous misreads. With a better understanding of which haplotypes support an unrooted perfect phylogeny, a method of error detecting can be implemented. This is useful since less information will be thrown out as a result of misreads.

A topic left to consider when looking at which g-lattice structures support an unrooted perfect phylogeny is to consider the bounds placed on the Pure Parsimony problem when a g-lattice has no tent structures present, that is when it represents an branching up structure. The Perfect Phylogeny Haplotyping bound on haplotype pairings for a branching up g-lattice is the same as the bound in place from the Pure Parsimony problem bound for chains.

References

[1] P. Blaine, C Davis, A. Holder, J. Silva, and C. Vinzant. Diversity graphs. In S. Butenko, W. Chaovalitwongse, and P. Pardalos, editors, *Clustering Challenges in Biological Networks*, pages 129–150. World Scientific, Singapore, 2008.

[2] Consortium and T.I.H. Integrating ethics and science in the international hapmap project. In *Nature Reviews Genetics*, pages 467–475. 5 edition, 2004.

[3] D. Gusfield. Efficient algorithms for inferring evolutionary history. In *Networks*, pages 21:19–28. 1991.

[4] A. Holder and T. Langley. A decomposition of the pure parsimony problem. In *Lecture Notes in Computer Science*, volume 5542, pages 198–208. 2009.

[5] S. Istrail, M. W aterman, and A. Clark. An overview of combinatorial methods for haplotype inference. In *Computational Methods for SNPs and Haplotype Inference*, volume 2983 of *Lecture Notes in Computer Science*, pages 9–25. 2004.