The Pure Parsimony Problem

Haplotyping and Minimum Diversity Graphs

Courtney Davis - University of Utah
Allen Holder - Trinity University
The Pure Parsimony Problem

Some Genetics

Mother
- Physical Trait
  - ABABBA
  - AABBB
  - A\times ABBX
- Genotype Representation
- Paired Gene Representation
- SNP 2

Father
- Physical Trait
  - BBAABA
  - BBAAAB
  - BB\times AXX
- Genotype Representation
- Paired Gene Representation
- SNP 2

Child
- Physical Trait
  - ABABBA
  - A\times BA
  - XBABX
- Genotype Representation
- Paired Gene Representation
- SNP 2

Mother’s Donation
- ABABBA
- ABA

Father’s Donation
- BBABA
- BBA

Haplotype
- ABABBA
- ABA
- BBABA
- BBA
- XBAXB
- XBA

SNP 2

SNP 7

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

**Haplotyping**

**Definition:** The problem is to reconstruct the haplotypes donated by a previous population from the genotypes of the current population.

**Why:** Tracing genetic markers from generation to generation is needed to gauge a population’s susceptibility to disease and in the design of patient-specific drugs.

**Past Research:** Investigations were started by Clark in 1990, and recent contributions were made by Lancia and Gusfield.

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

Clark’s Rule

1. Start with an empty collection of haplotypes.
2. Choose a genotype.
3. Add as few haplotypes to the set as possible (you need to add either 1 or 2) so that the genotype can be formed from the collection of haplotypes.
4. Continue until all genotypes can be formed.

This technique mimics what happens in nature. Notice that it can be interpreted as an attempt to find the smallest collection of haplotypes, but the process is dependent on the sequence of genotypes.

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

**Parsimony Problems** Haplotyping is a situation where simple explanations appear to be biologically relevant. So, finding small collections of haplotypes that can explain the genotypic information of the current population is important.

**Pure Parsimony Problem** Finding a smallest collection of haplotypes that can reconstruct a set of genotypes is called the Pure Parsimony Problem.
The Pure Parsimony Problem

For parent haplotypes $h^1$ and $h^2$ and offspring genotype $g$, we have the following at each SNP:

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

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

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
A bipartite graph \( D = (H, G, E) \) is a diversity graph if

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

Notice that the definition is biological.
A bipartite graph is a diversity graph if the nodes can be labeled to satisfy the definition.

The definition requires that the degree of every node in the genotype set has even degree.

There are graphs with each node having an even degree but that are not diversity graphs. As an example, $K(2, 2)$ is not a diversity graph because it violates $\oplus$. 

\begin{center}
\begin{tikzpicture}
\node[shape=circle,fill=magenta] (a) at (0,1) {};
\node[shape=circle,fill=magenta] (b) at (1,1) {};
\node[shape=circle,fill=magenta] (c) at (1,0) {};
\node[shape=circle,fill=magenta] (d) at (0,0) {};
\draw (a) -- (b);
\draw (a) -- (c);
\draw (b) -- (d);
\draw (c) -- (d);
\end{tikzpicture}
\end{center}
The Pure Parsimony Problem

Some Definitions

- The set of all haplotypes of length $n$ is denoted by $\mathcal{H}$.
- The largest edge set between the collection of genotypes $G$ and $\mathcal{H}$ is $\mathcal{E}$.
- Any subgraph of $(\mathcal{H}, G, \mathcal{E})$ that is a diversity graph and has the property that the subset of $\mathcal{H}$ is as small as possible is a solution to the Pure Parsimony problem. These subgraphs are denoted by $(\mathcal{H}^*, G, \mathcal{E}^*)$.
- There are typically several optimal subgraphs, which makes solving an IP formulation of the problem difficult.
The Pure Parsimony Problem

A Simple, but Useful Result

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

The proof is simple. An example for $n = 3$ is below.

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

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
Let \((V, W, E)\) be a bipartite graph. For each \(w\), define

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

Let \(\hat{V}(w)\) and \(\hat{F}(w)\) be vertex sets such that

\[
|\hat{V}(w)| = 2|T(w) - N(w)|_1
\]

and

\[
|\hat{F}(w)| = \begin{cases} 
0, & |N(w) \cup \hat{V}(w)| \text{ is even} \\
1, & |N(w) \cup \hat{V}(w)| \text{ is odd}
\end{cases}
\]

May need to add a node if \(|N(w)|\) is odd and \(\hat{V}(w)\) is empty.

Add enough to remove conflicts in \(N(w)\).
**Lemma** The bipartite graph \((V, W, E)\) can be extended and labeled to become a diversity graph by adding no more than

\[
\sum_{w \in W} |\hat{F}(w)| + (2|T(w)| - |N(w)|)_+ + (M_V - M_W)_+
\]

nodes to \(V\), provide that there are no isolated nodes.

**proof:** This is a long constructive proof that uses the Lexicographic Theorem.

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

\[
\left[ \sum_{w \in W} |\hat{F}(w)| + (2|T(w)| - |N(w)|)_+ \right] + (M_V - M_W)_+
\]

nodes to \(V\), where \(M_V\) and \(M_W\) are the number of isolated nodes in \(V\) and \(W\), respectively.
The Pure Parsimony Problem

**Locating Favorable Haplotypes**

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

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

Characterizing when $|H^*| = 2|G|$ 

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

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

($\Rightarrow$) Let $|H^*| = 2|G|$, and suppose for the sake of obtaining a contradiction that $T(g) \neq \emptyset$ for some $g \in G$. Then by Lemma 13, $H^*$
The Pure Parsimony Problem

contains an element in $\bigcup_{g \in G} T(g)$. Let $g^1$ and $g^2$ be such that $h^1 \oplus h^2 = g^1$ and $h^1 \oplus h^3 = g^2$, for some $h^1$, $h^2$, and $h^3$. Let $G' = G \setminus \{g^1, g^2\}$, and let $H' = \bigcup_{g \in G'} N(g)$. Furthermore, let $(H')^*$ be such that

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

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

$$2|G| = |H^*| \leq |(H')^*| + 3$$
$$\leq 2|G'| + 3$$
$$= 2(|G| - 2|) + 3$$
$$= 2|G| - 1.$$  

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

Allen Holder  
agholder@trinity.edu  
http://www.trinity.edu/aholder
The Pure Parsimony Problem

Restricting Mating Structure

- We now constrain our optimization problem so that the maximum number of mates that any haplotype can have is $m$.
- A smallest haplotype set that resolves $G$ with this restriction is denoted by $H_m^*$, and we let $\phi(m) = |H_m^*|$.

If $m = 1$, each haplotype can mate with at most one other haplotype. Biologically this means each parent can donate one of two haplotypes to a unique child, so this haplotype cannot be used to form another child. So, for $m = 1$ the neighborhoods of the genotypes in an optimal subgraph are disjoint, and the smallest number of haplotypes that can resolve $G$ is $\phi(1) = 2|G| - u$, where $u$ is the number of genotypes with no ambiguous
The Pure Parsimony Problem

SNPs.

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

### Properties of $\phi(m)$

- Calculating $\phi(m^*)$ solves the Pure Parsimony problem and indicates the least amount of mating needed.
- At some threshold, increasing $m$ does not change the cardinality of $H_m^*$. Hence, for some $m$, $\phi(m) = \phi(m + k)$ for every natural number $k$.
- Increasing the number of possible mates that any haplotype is allowed never causes an increase in $H_m^*$. Thus, $\phi(m) \geq \phi(m + 1)$ for all $m$, and $\phi$ is non-increasing.
- The smallest $m$ such that $\phi(m) = \phi(m + k)$, for all $k \in \mathbb{N}$, is denoted by $m^*$. So, if $m \geq m^*$, we have that $\phi(m) = \phi(m^*)$.
- No haplotype can mate with more than $|G|$ haplotypes, and hence, $m^* \leq |G|$.
- If no haplotype reconciles more than one genotype, $m^* = 1$.  

---

Allen Holder  
agholder@trinity.edu  
http://www.trinity.edu/aholder
What if $m^*$ is at its Upper Bound

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

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

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

**Case 1:** Suppose that $h' \oplus h' \notin G$. Then, $h'$ mates with a unique $h^i \in H_{m^*}^* \setminus \{h'\}$ to resolve each $g^i \in G$. Hence,

$$\phi(m^*) = |H^*| = |G| + 1.$$  

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

Allen Holder  
agholder@trinity.edu  
http://www.trinity.edu/aholder
The Pure Parsimony Problem

Calculating $\phi(2)$

<table>
<thead>
<tr>
<th>Step 1:</th>
<th>Set $v = 0$ and $(H_v, G_v) = (H, G)$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2:</td>
<td>Find the longest path in $(H_v, G_v)$, say $P_v$. If no path exists, set $P_v = \emptyset$.</td>
</tr>
<tr>
<td>Step 3:</td>
<td>If $P_v = \emptyset$, stop.</td>
</tr>
<tr>
<td>Step 4:</td>
<td>Index $v$ by 1.</td>
</tr>
<tr>
<td>Step 5:</td>
<td>Set $(H_{v+1}, G_{v+1}) = (H_v, G_v) \backslash P_v$.</td>
</tr>
<tr>
<td>Step 6:</td>
<td>Index $v$ by 1.</td>
</tr>
<tr>
<td>Step 7:</td>
<td>Go to Step 2.</td>
</tr>
</tbody>
</table>

This greedy algorithm iteratively removes the longest paths in a diversity graph.

Allen Holder
aghoulder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

The Greedy Algorithm Works

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

**proof:** The proof follows by induction on \(|G|\).
The Pure Parsimony Problem

An Example

The paths through the genotypes

\[ g^1 = AXBBBB \quad g^2 = XAXXBB \quad g^3 = BXAXBX \]
\[ g^4 = BXXAXB \quad g^5 = BBBXAB \quad g^6 = BBXBBA \]

must pass through these genotypes as indicated below.

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

Path Decompositions

<table>
<thead>
<tr>
<th>First Path’s Genotype Progression</th>
<th>Second Path’s Genotype Progression</th>
</tr>
</thead>
<tbody>
<tr>
<td>((g^1, g^2, g^3, g^4, g^5))</td>
<td>((g^6))</td>
</tr>
<tr>
<td>((g^1, g^2, g^4, g^5))</td>
<td>((g^3, g^6))</td>
</tr>
<tr>
<td>((g^1, g^2, g^3, g^6))</td>
<td>((g^4, g^5))</td>
</tr>
<tr>
<td>((g^6, g^3, g^4, g^5))</td>
<td>((g^1, g^2))</td>
</tr>
</tbody>
</table>

The greedy algorithm finds the first solution in the Table, as the first path is as long as possible. None of the other paths have this property, and so the algorithm is not capable of finding these solutions.

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

Future Directions

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

Allen Holder
agholder@trinity.edu
http://www.trinity.edu/aholder
The Pure Parsimony Problem

Thank you for your time, please ask questions