The Clark Method of Phasing

- Developed in 1990 by Andy Clark
Developed in 1990 by Andy Clark

The phasing problem now is related to PCR ambiguity in diploid individuals

For example, suppose that the two chromosomes are:

- - - - - ∗ - - - - ∗ - - - -

mother

- - - - - ∗ - - - - ∗ - - - -

father

How can we resolve this ambiguity?
The Clark Method of Phasing

- Developed in 1990 by Andy Clark
- The phasing problem now is related to PCR ambiguity in diploid individuals
- For example, suppose that the two chromosomes are:

  - mother: ── ── ── ∗ ── ── ∗ ── ──
  - father: ── ── ── ∗ ── ── ∗ ── ──

  $A_0/A_1$        $B_0/B_1$

- How can we resolve this ambiguity?
If there are $k$ ambiguous sites, there are an exponential (in $k$) number of possible explanations of the ambiguity.
If there are $k$ ambiguous sites, there are an exponential (in $k$) number of possible explanations of the ambiguity.

Assume that all sites have *at most* two alleles (the infinite sites model).
If there are \( k \) ambiguous sites, there are an exponential (in \( k \)) number of possible explanations of the ambiguity.

Assume that all sites have \textit{at most} two alleles (the infinite sites model).

Denote the two alleles present by 0 and 1.

\begin{itemize}
  \item \textit{Haplotypes}:
    \begin{align*}
      &1) \quad 001010 \\
      &2) \quad 011000
    \end{align*}
  \end{itemize}

\begin{itemize}
  \item \textit{Genotype}:
    \begin{align*}
      &0(\frac{0}{1})10(\frac{0}{1})0
    \end{align*}
\end{itemize}
If there are $k$ ambiguous sites, there are an exponential (in $k$) number of possible explanations of the ambiguity.

Assume that all sites have at most two alleles (the infinite sites model).

Denote the two alleles present by 0 and 1.

*Haplotypes:*  
1) 001010  

If we were given just the genotype, then there are two possible explanations ($k = 2$)

*Genotype:* 0(0\_1)10(0\_1)0

2) 011000

001000 001010
011010 011000
The Clark Method of Phasing

- Notation: A genotype will be a sequence of 0, 1, and 2’s [ex. 012120]
The Clark Method of Phasing

- Notation: A genotype will be a sequence of 0, 1, and 2’s [ex. 012120]
- ‘2’ denotes ambiguous site
Notation: A genotype will be a sequence of 0, 1, and 2’s [ex. 012120]
‘2’ denotes ambiguous site
haplotype will be a sequences of 0 and 1’s [ex. 0101010]
Notation: A genotype will be a sequence of 0, 1, and 2’s [ex. 012120]

‘2’ denotes ambiguous site

Haplotype will be a sequence of 0 and 1’s [ex. 0101010]

An explanation of a genotype will be a *pair* of haplotypes

\[\text{Haplotypes: } \begin{align*}
1) & \quad 001010 \\
2) & \quad 011000
\end{align*}\]

\[\rightarrow \quad \text{Genotype: } 0\left(\frac{0}{1}\right)10\left(\frac{0}{1}\right)0 \\
0(2)10(2)0\]
Two easy cases:

- **Homozygous**: genotype with only 0's and 1's. The mother and the father chromosome have the same composition.
- **Single heterozygote**: genotype with a single ‘2’. The explanation will still be unique.
Two easy cases:

- **Homozygous**: genotype with only 0's and 1's. The mother and the father chromosome have the same composition.
- **Single heterozygote**: genotype with a single ‘2’. The explanation will still be unique.

**Multiple heterozygote case**: More than single ‘2’s in the genotype [ex. 01212]
Two easy cases:
- Homozygous: genotype with only 0's and 1's. The mother and the father chromosome have the same composition.
- Single heterozygote: genotype with a single ‘2’. The explanation will still be unique.

Multiple heterozygote case: More than single ‘2’s in the genotype [ex. 01212]

Number of explanations for a genotype with k ambiguous sites is $2^{k-1}$ [in this case 01010, 01011, 01110, 01111]
Suppose that we are given 5 sites on 5 individuals

- Individual: 01202
- Mate: 10000

- Individual: 20000
- Mate: 00000

- Individual: 00100
- Mate: 00100

- Individual: 01000
- Mate: 01000
Suppose that we are given 5 sites on 5 individuals

Start with the easy cases, the homozygotes and single heterozygotes

Example.

Individuals: 01202, 20000, 12121, 00100, and 01000

Starting with the ‘easy’ cases, we are able to resolve 3 of the 5 individuals:

<table>
<thead>
<tr>
<th>Individual</th>
<th>Inferred Mate</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000</td>
<td>10000</td>
</tr>
<tr>
<td>00100</td>
<td>00100</td>
</tr>
<tr>
<td>01000</td>
<td>01000</td>
</tr>
</tbody>
</table>
Suppose that we are given 5 sites on 5 individuals

Start with the easy cases, the homozygotes and single heterozygotes

**Example.** Individuals: 01202, 20000, 12121, 00100, and 01000
Suppose that we are given 5 sites on 5 individuals

Start with the easy cases, the homozygotes and single heterozygotes

**Example.** Individuals: 01202, 20000, 12121, 00100, and 01000

Starting with the ‘easy’ cases, we are able to resolve 3 of the 5 individuals:
Suppose that we are given 5 sites on 5 individuals

Start with the easy cases, the homozygotes and single heterozygotes

**Example.** Individuals: 01202, 20000, 12121, 00100, and 01000

Starting with the ‘easy’ cases, we are able to resolve 3 of the 5 individuals:

<table>
<thead>
<tr>
<th>Individual</th>
<th>Inferred Mate – Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000</td>
<td>10000 – 00000</td>
</tr>
<tr>
<td>00100</td>
<td>00100 – 00100</td>
</tr>
<tr>
<td>01000</td>
<td>01000 – 01000</td>
</tr>
</tbody>
</table>
The Clark Method of Phasing

For the unresolved chromosomes (01202, 12121), we have two possible explanations for each:

<table>
<thead>
<tr>
<th>Individual</th>
<th>Explanation 1</th>
<th>Explanation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>01202 :</td>
<td>01101</td>
<td>01001</td>
</tr>
<tr>
<td></td>
<td>01000</td>
<td>01100</td>
</tr>
<tr>
<td>12121 :</td>
<td>10101</td>
<td>11101</td>
</tr>
<tr>
<td></td>
<td>11111</td>
<td>10111</td>
</tr>
</tbody>
</table>
Now, if we look at the possible explanations together, we see that there is overlap between the inferred mate-pairs from the simple cases and the possible explanations for the ambiguous cases.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Inferred Mate – Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>001000 :</td>
<td>00100 – 00100</td>
</tr>
<tr>
<td>010000 :</td>
<td>01000 – 01000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individual</th>
<th>Explanation 1</th>
<th>Explanation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>01202 :</td>
<td>01101</td>
<td>01001</td>
</tr>
<tr>
<td></td>
<td><strong>01000</strong></td>
<td>01100</td>
</tr>
<tr>
<td>12121 :</td>
<td>10101</td>
<td>11101</td>
</tr>
<tr>
<td></td>
<td>11111</td>
<td>10111</td>
</tr>
</tbody>
</table>
The Clark Method of Phasing

- Notation: $G = H \oplus H'$ denotes that haplotype pair $(H, H')$ is an explanation for $G$
- Notation: $H \rightarrow^C G'$ denotes $G$ can be resolved using $H \in C$, i.e., there exists $H'$ s.t. $G = H \oplus H'$, and we call $H'$ the *inferred haplotype*
- Example: For $G = 02112$:

  Possible Mate − Pairs

  $G : 00110 \oplus 01111$
  $01110 \oplus 00111$
Clark Algorithm

1. Find all homozygotes and single heterozygotes and make a list of all of the haplotypes involved in the unique explanations.

2. While there are remaining genotypes that are unresolved, attempt to find a haplotype from the list that helps resolve some unresolved genotype. If such a haplotype exists, add the corresponding mate-pair to the list of haplotypes and label the genotype as resolved.
The Clark Method of Phasing

Clark Algorithm

1. Find all homozygotes and single heterozygotes and make a list of all of the haplotypes involved in the unique explanations.

2. While there are remaining genotypes that are unresolved, attempt to find a haplotype from the list that helps resolve some unresolved genotype. If such a haplotype exists, add the corresponding mate-pair to the list of haplotypes and label the genotype as resolved.

Example: If the genotype is 12121, then 10101 is the mate-pair of 11111 in the explanation of genotype 12121.

- If we are able to find a haplotype in the list and thus explain the ambiguous haplotype, we have inferred a uniquely inferred additional haplotype, that of the mate-pair.
Drawbacks to the Clark Methods:

- Possible that a genotype can be resolved in multiple ways from the list, thus yielding multiple explanations. Which one do you choose? (Associated problem: anomalous genotypes i.e., explanations that are not correct)
- Possible that unresolved genotypes are all incompatible with haplotypes in list. Then the algorithm will stop and leave behind “orphan”, or unresolved genotypes.
- If there are no homozygotes or single heterozygotes, then the algorithm cannot start.
- The order in which you resolve the haplotypes matters. A different ordering may produce a different haplotype list.
The Clark Method of Phasing

- Drawbacks to the Clark Methods:
  - Possible that a genotype can be resolved in multiple ways from the list, thus yielding multiple explanations. Which one do you choose? (Associated problem: anomalous genotypes i.e., explanations that are not correct)
  - Possible that unresolved genotypes are all incompatible with haplotypes in list. Then the algorithm will stop and leave behind "orphan", or unresolved genotypes
  - If there are no homozygotes or single heterozygotes, then the algorithm cannot start.
  - The order in which you resolve the haplotypes matters. A different ordering may produce a different haplotype list.
  - What is the probability of the algorithm stopping prematurely or not being able to start?
The Clark Method of Phasing

- **Drawbacks to the Clark Methods:**
  - Possible that a genotype can be resolved in multiple ways from the list, thus yielding multiple explanations. Which one do you choose? (Associated problem: anomalous genotypes i.e., explanations that are not correct)
  - Possible that unresolved genotypes are all incompatible with haplotypes in list. Then the algorithm will stop and leave behind “orphan”, or unresolved genotypes
The Clark Method of Phasing

- **Drawbacks to the Clark Methods:**
  - Possible that a genotype can be resolved in multiple ways from the list, thus yielding multiple explanations. Which one do you choose? (Associated problem: anomalous genotypes i.e., explanations that are not correct)
  - Possible that unresolved genotypes are all incompatible with haplotypes in list. Then the algorithm will stop and leave behind “orphan”, or unresolved genotypes
  - If there are no homozygotes or single heterozygotes, then the algorithm cannot start.
• Drawbacks to the Clark Methods:
  • Possible that a genotype can be resolved in multiple ways from the list, thus yielding multiple explanations. Which one do you choose? (Associated problem: anomalous genotypes i.e., explanations that are not correct)
  • Possible that unresolved genotypes are all incompatible with haplotypes in list. Then the algorithm will stop and leave behind “orphan”, or unresolved genotypes
  • If there are no homozygotes or single heterozygotes, then the algorithm cannot start.
  • The order in which you resolve the haplotypes matters. A different ordering may produce a different haplotype list.
The Clark Method of Phasing

- **Drawbacks to the Clark Methods:**
  - Possible that a genotype can be resolved in multiple ways from the list, thus yielding multiple explanations. Which one do you choose? (Associated problem: anomalous genotypes i.e., explanations that are not correct)
  - Possible that unresolved genotypes are all incompatible with haplotypes in list. Then the algorithm will stop and leave behind “orphan”, or unresolved genotypes
  - If there are no homozygotes or single heterozygotes, then the algorithm cannot start.
  - The order in which you resolve the haplotypes matters. A different ordering may produce a different haplotype list.

- **What is the probability of the algorithm stopping prematurely or not being able to start?**
How can we estimate the number of 2’s?

Population model: Infinite Sites Model (at most one mutation can happen at any site on the chromosome)
The Clark Method of Phasing

- How can we estimate the number of 2’s?
- Population model: Infinite Sites Model (at most one mutation can happen at any site on the chromosome)
- Neutral Model of Evolution [\[ \text{expected number of mismatches of a DNA sequence: } \Theta = \frac{L \Theta_{nt}}{N} \]

\( L \) = length of the sequence
\( \Theta_{nt} = 4N\mu \)
\( N \) = the effective population size
\( \mu \) = the mutation rate per nucleotide per generation
The Clark Method of Phasing

- How can we estimate the number of 2’s?
- Population model: Infinite Sites Model (at most one mutation can happen at any site on the chromosome)
- Neutral Model of Evolution [ ]
- expected number of mismatches of a DNA sequence:

\[ \Theta = L\Theta_{nt} \]

where

- \( L \) = length of the sequence
- \( \Theta_{nt} = 4N\mu \)
- \( N \) = the effective population size
- \( \mu \) is the mutation rate per nucleotide per generation
For example, in Drosophila, $\Theta = .005L$.

$$Pr(2 \text{ sequences have } m \text{ mismatches}) = \frac{1}{\theta + 1} \left( \frac{\theta}{\theta + 1} \right)^m$$

In the infinite sites model,

$$Prob(2 \text{ genes identical}) = \frac{1}{1 + \theta} (m = 0)$$

Probability of two different genes is

$$1 - Prob(2 \text{ genes identical}) = \frac{\theta}{1 + \theta}$$
Now, if we have $N$ diploid individuals,

$$\Pr(\text{No homozygotes}) \ [\text{i.e., } \Pr(\text{algorithm won’t start})]$$

is obtained using \textit{Ewing’s Sampling Lemma}.

If $\Theta > 0.5$, Clark’s algorithm will work well.
Now, if we have $N$ diploid individuals,

$$Pr(\text{No homozygotes}) \ [\text{i.e., } Pr(\text{algorithm won't start})]$$

is obtained using *Ewing's Sampling Lemma*

$$Pr(\text{No Homozygotes}) = \frac{\theta^3 + 4\theta^2 + 2\theta}{(1 + \theta)(2 + \theta)(3 + \theta)}$$
Now, if we have $N$ diploid individuals, 

$$Pr(\text{No homozygotes}) \quad [\text{i.e., } Pr(\text{algorithm won’t start})]$$

is obtained using *Ewing’s Sampling Lemma*

$$Pr(\text{No Homozygotes}) = \frac{\theta^3 + 4\theta^2 + 2\theta}{(1 + \theta)(2 + \theta)(3 + \theta)}$$

$$Pr(\text{Single site Heterozygote}) = \frac{\theta}{(1 + \theta)^2}$$
Now, if we have \( N \) diploid individuals,

\[
Pr(\text{No homozygotes}) \quad [\text{i.e., } Pr(\text{algorithm won't start})]
\]
is obtained using *Ewing’s Sampling Lemma*

\[
Pr(\text{No Homozygotes}) = \frac{\theta^3 + 4\theta^2 + 2\theta}{(1 + \theta)(2 + \theta)(3 + \theta)}
\]

\[
Pr(\text{Single site Heterozygote}) = \frac{\theta}{(1 + \theta)^2}
\]

\[
Pr(\text{Algorithm won't start}) = \left[1 - \frac{1}{1 + \theta} - \frac{\theta}{(1 + \theta)^2}\right]^N
\]

Thus, if \( \Theta > 0.5 \), Clark’s algorithm will work well.
Now, if we have $N$ diploid individuals,

$$Pr(\text{No homozygotes}) \ [i.e., \ Pr(\text{algorithm won’t start})]$$

is obtained using Ewing’s Sampling Lemma

$$Pr(\text{No Homozygotes}) = \frac{\theta^3 + 4\theta^2 + 2\theta}{(1 + \theta)(2 + \theta)(3 + \theta)}$$

$$Pr(\text{Single site Heterozygote}) = \frac{\theta}{(1 + \theta)^2}$$

$$Pr(\text{Algorithm won’t start}) = \left[1 - \frac{1}{1 + \theta} - \frac{\theta}{(1 + \theta)^2}\right]^N$$

Thus, if $\Theta > 0.5$, Clark’s algorithm will work well.
How can we deal with the ‘orphan’ genotypes? What is the probability of not finishing?

The answer depends on the algorithm implementation and the order in which you resolve the haplotypes.
The Clark Method of Phasing

How can we deal with the ‘orphan’ genotypes? What is the probability of not finishing?
The answer depends on the algorithm implementation and the order in which you resolve the haplotypes.

**Maximum Resolution Problem:**

**Input:** A set of vectors, ambiguous (0,1,2) and resolved (0,1) genotypes

**Output:** Maximum number of ambiguous vectors that can be resolved by successive applications of Clark’s Rule.

Equivalent to minimize the number of orphan genotypes
Theorem

The Maximum Resolution Problem is NP-complete.
The Clark Method of Phasing

Theorem

The Maximum Resolution Problem is NP-complete.

The proof for this theorem is based on a reduction to the satisfiability problem (SAT).
The Clark Method of Phasing

Proof.

- Let $X_1$, $X_2$, $X_3$, and $X_4$ be variables, and let $(x_1, \overline{x_1}, x_2, \overline{x_2}, x_3, \overline{x_3}, x_4, \overline{x_4})$ be literals. The satisfiability problem is
  - **Input**: Boolean function $F$
  - **Output**: Is there a truth assignment that makes $F$ true? If so, find such an assignment.
The Clark Method of Phasing

Proof.

Let $X_1, X_2, X_3,$ and $X_4$ be variables, and let $(x_1, \bar{x}_1, x_2, \bar{x}_2, x_3, \bar{x}_3, x_4, \bar{x}_4)$ be literals. The satisfiability problem is

- **Input**: Boolean function $F$
- **Output**: Is there a truth assignment that makes $F$ true? If so, find such an assignment.

We are going to attempt to create a 1-1 correspondence between the boolean logic and haplotypes using the above function $F$. 
To the function $F$, we associate a set of genotypes (corresponding to columns in the matrix). The number of rows in the matrix will be \# variables $+ 2 \times \# variables + \# clauses + 1 = 3V + C + 1$. 
To the function $F$, we associate a set of genotypes (corresponding to columns in the matrix). The number of rows in the matrix will be $\#\text{ variables} + 2 \times \#\text{variables} + \#\text{clauses} + 1 = 3V + C + 1$.

We fill in rows $C_1, C_2, C_3$ in the following manner: if $X_i$ is absent from a clause, put ‘1’ in both $T_i$ and $F_i$. If $X_i$ is present in a clause, place ‘1’ in $T_i$ and ‘0’ in $F_i$. If $\overline{X}_i$ is present in a clause, then place a ‘0’ in $T_i$ and a ‘1’ in $F_i$. 
To the function $F$, we associate a set of genotypes (corresponding to columns in the matrix). The number of rows in the matrix will be $\# \text{ variables} + 2 \times \# \text{variables} + \# \text{clauses} + 1 = 3V + C + 1$.

We fill in rows $C_1, C_2, C_3$ in the following manner: if $X_i$ is absent from a clause, put ‘1’ in both $T_i$ and $F_i$. If $X_i$ is present in a clause, place ‘1’ in $T_i$ and ‘0’ in $F_i$. If $\overline{X_i}$ is present in a clause, then place a ‘0’ in $T_i$ and a ‘1’ in $F_i$.

Fill in columns $S_1 \rightarrow S_4$ by:

\[
\begin{cases} 
0' & \text{if 00} \\
1' & \text{if 11} \\
2' & \text{if 01 or 10}
\end{cases}
\]
To fill in the $C_1 \rightarrow C_3$ columns, we look at our function $F$. 
To fill in the $C_1 \rightarrow C_3$ columns, we look at our function $F$. We see that $x_1$ appears in both clause 2 and clause 3. So we place a ‘2’ in both $C_2$ and $C_3$. 

$\bar{x}_1$ appears in $C_3$, but no other clause. Place a ‘2’ in $C_3$. etc. (Note: the rows $x_1$ and $\bar{x}_1$ should sum to the $X_1$ row in the $C$ columns.)
To fill in the $C_1 \rightarrow C_3$ columns, we look at our function $F$.

We see that $x_1$ appears in both clause 2 and clause 3. So we place a ‘2’ in both $C_2$ and $C_3$.

For the rows containing the literals $(x_1 \overline{x_1},$ etc.) $x_1$ appears in $C_2$ but no other clause.
To fill in the $C_1 \rightarrow C_3$ columns, we look at our function $F$. We see that $x_1$ appears in both clause 2 and clause 3. So we place a ‘2’ in both $C_2$ and $C_3$.

For the rows containing the literals ($x_1 \overline{x}_1$, etc.) $x_1$ appears in $C_2$ but no other clause. Place a ‘2’ in $C_2$. $\overline{x}_1$ appears in $C_3$, but no other clause. Place a ‘2’ in $C_3$. etc. (Note: the rows $x_1$ and $\overline{x}_1$ should sum to the $X_1$ row in the C columns.)
How do we fill in the bottom right corner of the table?

- We have a clause set $C_1... C_c$, one for each clause. All will be ambiguous vectors.
- For each $K=1,2,...,c$ the first $v$ positions of the vector $C_k$ are zero except for position $i$ such that either $x_i$ or $\overline{x_i}$ appears in $C_k$. We are blind to the actual truth value.
- For the next $2v$ positions [the literals], place a zero except for any position $v+2i-1$ where $x_i$ appears in clause $C_k$ or position $v+2i$ where $\overline{x_i}$ appears in clause $C_k$. These positions are set to 2.
- For each $r$ from 1 to $c$, position $3v+r$ of $C_k = 0$ if and only if $r=k$ [the diagonal].
- For $r \neq k$ position $3v+r = 2$ if and only if clause $k$ and $r$ contain a variable in common [not necessarily a literal in common].
- Otherwise, position $3v+r = 1$. [This assignment captures the ambiguity related to the literals that are contained in multiple clauses. You may have both $x_i$ and $\overline{x_i}$ present in $F$].
We reduce SAT to MR. Start with a generic boolean formula $F$ with $C$ clauses and $V$ variables $X_1 \ldots X_V$. $F$ takes a set of vectors $V(F)$ that are ambiguous and resolved as input to the MR problem [the columns of out table drawn above]. We want to show that $F$ has a satisfying truth assignment iff $V(F)$ has the maximum number of ambiguous vectors explained by a series of Clark Rules. Recall that $F = (x_2 \lor x_3 \lor \overline{x}_4) \land (x_1 \lor \overline{x}_2 \lor x_4) \land (\overline{x}_1 \lor \overline{x}_3)$. 
How can we interpret resolution by the Clark Method?

- Suppose that you pick column $T_1$ to resolve one of the other columns; this is interpreted as setting the literal $X_1$ to true.
How can we interpret resolution by the Clark Method?

- Suppose that you pick column $T_1$ to resolve one of the other columns; this is interpreted as setting the literal $X_1$ to true.
- Suppose that we try to resolve column [haplotype] $S_1$. $S_1$ can be resolved by using either the $T_1$ or $F_1$ columns. By picking one of the columns, we are fixing $X_1$ as true or false.
- Using the haplotypes obtained in the resolution of the $S$ columns along with the $T$ and $F$ columns, can we resolve the columns $C_1 \rightarrow C_3$? (NO)
### The Clark Method of Phasing

<table>
<thead>
<tr>
<th></th>
<th>( T_1 )</th>
<th>( F_1 )</th>
<th>( T_2 )</th>
<th>( F_2 )</th>
<th>( T_3 )</th>
<th>( F_3 )</th>
<th>( T_4 )</th>
<th>( F_4 )</th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( X_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( x_1 )</th>
<th>( \bar{x}_1 )</th>
<th>( x_2 )</th>
<th>( \bar{x}_2 )</th>
<th>( x_3 )</th>
<th>( \bar{x}_3 )</th>
<th>( x_4 )</th>
<th>( \bar{x}_4 )</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \bar{x}_1 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \bar{x}_2 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \bar{x}_3 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \bar{x}_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

| \( C_1 \) | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 2 | 2 | 0 | 2 | 2 |
| \( C_2 \) | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 1 | 2 | 2 | 0 | 2 |
| \( C_3 \) | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 0 |
| \( C_B \) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
Input: A set of unresolved genotypes
Output: Maximum number of ambiguous vectors that can be resolved by successive applications of Clark’s Rule.
**Input:** A set of unresolved genotypes

**Output:** Maximum number of ambiguous vectors that can be resolved by successive applications of Clark’s Rule.

- Recall the notation $R \rightarrow^C A$ where $R = \text{resolved haplotype}$ and $A = \text{ambiguous haplotype}$. 
Input: A set of unresolved genotypes
Output: Maximum number of ambiguous vectors that can be resolved by successive applications of Clark’s Rule.

- Recall the notation $R \rightarrow^C A$ where $R = \text{resolved haplotype}$ and $A = \text{ambiguous haplotype}$.
- If we call $A[i]$ the $i$-th site of $A$ and $R[j]$ the $j$-th site in $R$ then the notation means that if $A[i] = 0$ or $1$ then $R[i] = 0$ or $1$ and $R[i] = A[i]$.
Maximum Resolution Problem

Further explanation of the NP-completeness proof: For a more full presentation, see Dan Gusfield’s paper

*Inference of Haplotypes from Samples of Diploid Populations: Complexity and Algorithms, J. Computational Biology August 2001.*
**Remarks** (Table Setup)

- For every variable, there are two columns T and F
- A column of selectors exists for each random variable
- A column exists for each clause
- The first V rows are associated with the random variables
- The next set of rows are associated with the literals
- The final set of rows corresponds to clauses, including the mysterious (for a little while longer anyways) $C_b$
- $T_1, F_1, \ldots, T_4, F_4$ are all *resolved* columns while the rest are *unresolved*
Remarks (General Properties)

- $T_i \rightarrow^C S_i$ or $F_i \rightarrow^C S_i$ but $T_i \not\rightarrow^C S_j$ and $F_i \not\rightarrow^C S_j$. i.e. Column $T_i$ can be applied by Clark Rule to column $S_i$ but to no other selector column.

- At most one T or F can be applied to any S. i.e. set $X_1$ to either T or F, but not both!

- We interpret as follows: if $T_i \rightarrow^C S_i$ as ‘$X_i = true$’; if $F_i \rightarrow^C S_i$ as ‘$X_i = false$’. Suppose $T_1 \rightarrow^C S_1$, $F_2 \rightarrow^C S_2$, $F_3 \rightarrow^C S_3$, $T_4 \rightarrow^C S_4$, so the inferred vectors are $R_1 = S_1 \oplus T_1$, $R_2 = S_2 \oplus F_2$, $R_3 = S_3 \oplus F_3$, $R_4 = S_4 \oplus T_4$

- $R_1, R_2, R_3, R_4$ can be applied only to $C_1, C_2$ and $C_3$. [Consider the Blocking Clause $C_b$] The last entry of $R_1$ to $R_4$ will be a 1

- No T or F can be applied to the C vectors [Because of the Blocking Clause, $C_b$]

- $T_i \rightarrow^C S_i$, then $R_k \rightarrow^C C_k$ iff the literal $x_i$ appears in $C_k$. Similarly, $F_i \rightarrow^C S_i$ then $R_k \rightarrow^C C_k$ iff the literal $\bar{x}_i$ appears in $C_k$. 

Sorin Istrail  Clark Method of Phasing