Constraint Programming and Biology: Haplotype Inference

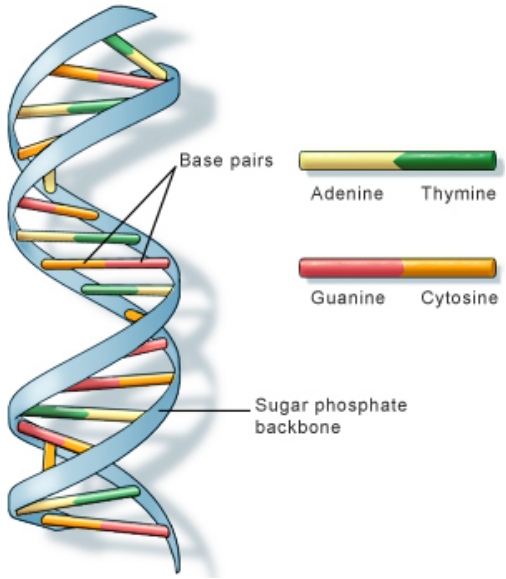
Agostino Dovier

Dept. Math and Computer Science, Univ. of Udine, Italy

ACP Summer School in Constraint Programming
Wrocław, September 2012

DNA and Genome in a nutshell

- DNA (DeoxyriboNucleic Acid) is characterized by a string of nucleotides: A, C, G, and T (Adenine, Cytosine, Guanine, Thymine)
- Given a sequence $s \in \{A, C, G, T\}^*$ the complementary sequence $\bar{s}$ is deterministically obtained by substituting $A \leftrightarrow T$ and $C \leftrightarrow G$
- s and $\bar{s}$ fold together forming the famous double helix



DNA and Genome in a nutshell

- DNA strings are huge (10^6–10^{10} nucleotides).
- Differences between the DNAs of two members of the same specie are limited (e.g., 1 on 1000 for humans)
- Some fragments of the DNA encode proteins (we'll be back on that later). Let's say for now that they are very important parts and called genes.
- In the Human DNA it is estimated that there are 23000 (maybe few) protein-coding genes.
- Differences of some nucleotides in the same gene characterize a property of an individual w.r.t. another.
- The set of all genes of an individual is called genome

Haplotype Inference

- Genes are packaged in bundles called chromosomes. (Chromosomes are therefore regions of DNA)
- In diploid organisms (like humans) we have 23 homologous chromosome pairs, one coming from the DNA of the father, another coming from the DNA of the mother.
- A haplotype is a DNA sequence that has been inherited from one parent.

Haplotype Inference

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

···	G	A	T	C	T	G	T	A	C	T	G	A	G	T	···
···	G	A	T	C	T	G	T	A	C	T	G	A	A	T	···

Haplotype Inference

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	G	T	$\cdots$
$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	A	T	$\cdots$

In some (typical) points, the bases can be different.

Haplotype Inference

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

···	G	A	T	C	T	G	T	A	C	T	G	A	G	T	···
···	G	A	T	C	T	G	T	A	C	T	G	A	A	T	···

In some (typical) points, the bases can be different.
If this is the case, we say that there is a Single Nucleotide Polymorphism (SNP).

Haplotype Inference

Each person inherits two haplotypes (from the mother and from the father) for most regions of the genome.

$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	G	T	$\cdots$
$\cdots$	G	A	T	C	T	G	T	A	C	T	G	A	A	T	$\cdots$

In some (typical) points, the bases can be different.
If this is the case, we say that there is a Single Nucleotide Polymorphism (SNP).

Changes are always $C \leftrightarrow T$ and $A \leftrightarrow G$

Haplotype Inference

- A good introduction in `http://csiflabs.cs.ucdavis.edu/~gusfield/gusfieldorzack.pdf`
- The Haplotype Inference problem(s) is(are) introduced to investigate genetic variations in a population.
- Some particular points of the DNA where typically mutations are concentrated are selected (SNPs).
- These lists of points are analyzed.

Haplotype Inference

Single Nucleotide Polymorphism (SNP)

Each person has two haplotypes (from the mother and from the father) for most regions of the genome:

G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	G	T
G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	A	T
Let us focus on the SNPs:																
	A			C								T			G	
	A			C								T			A	

Haplotype Inference

Single Nucleotide Polymorphism (SNP)

Each person has two haplotypes (from the mother and from the father) for most regions of the genome:

G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	G	T
G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	A	T
Let us focus on the SNPs:																
	A			C								T			G	
	A			C								T			A	
	0			0								1			2	

We know that at a location (site/locus) there is a SNP.

Haplotype Inference

Single Nucleotide Polymorphism (SNP)

Each person has two haplotypes (from the mother and from the father) for most regions of the genome:

G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	G	T
G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	A	T
Let us focus on the SNPs:																
	A			C								T			G	
	A			C								T			A	
	0			0								1			2	

We know that at a location (site/locus) there is a SNP.
We know whether the SNP is $C \leftrightarrow T$ and $A \leftrightarrow G$.

Haplotype Inference

Single Nucleotide Polymorphism (SNP)

Each person has two haplotypes (from the mother and from the father) for most regions of the genome:

G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	G	T
G	A	A	T	C	T	T	C	G	T	A	C	T	G	A	A	T
Let us focus on the SNPs:																
	A			C								T			G	
	A			C								T			A	
	0			0								1			2	

We know that at a location (site/locus) there is a SNP.
We know whether the SNP is $C \leftrightarrow T$ and $A \leftrightarrow G$.
We assign a SNP a 0-2 value in the following way:

$C, C \mapsto 0$ $T, T \mapsto 1$ $C, T \mapsto 2$ $T, C \mapsto 2$
$A, A \mapsto 0$ $G, G \mapsto 1$ $A, G \mapsto 2$ $G, A \mapsto 2$

Haplotype Inference

- A string of $\{0, 1, 2\}^*$ is called a *genotype*
- A string of $\{0, 1\}^*$ is called a *haplotype*
- Two equal length haplotypes generate a unique genotype

Haplotype Inference

- A string of $\{0, 1, 2\}^*$ is called a *genotype*
- A string of $\{0, 1\}^*$ is called a *haplotype*
- Two equal length haplotypes generate a unique genotype
  E.g., $0010, 0101 \Rightarrow 0222$

Haplotype Inference

- A string of $\{0, 1, 2\}^*$ is called a *genotype*
- A string of $\{0, 1\}^*$ is called a *haplotype*
- Two equal length haplotypes generate a unique genotype
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture *haplotype*s that generated it

Haplotype Inference

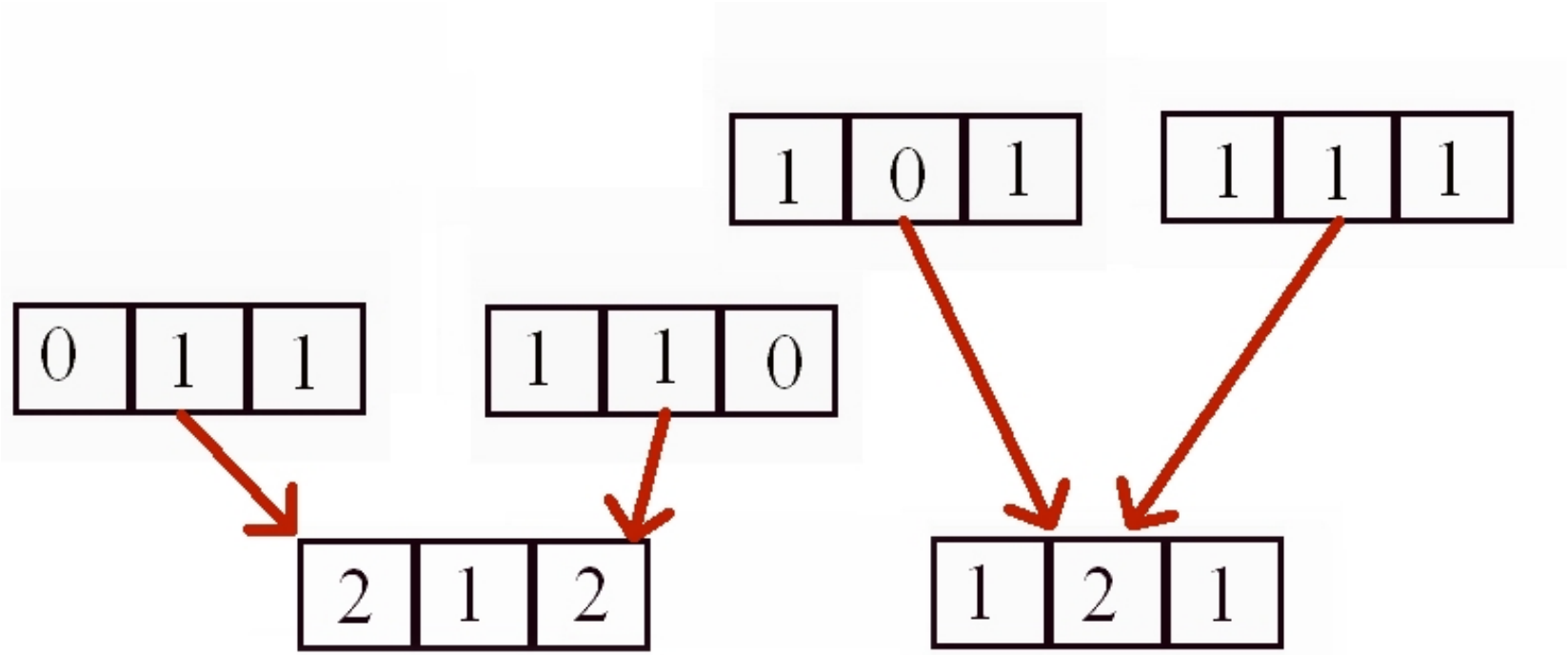
- A string of $\{0, 1, 2\}^*$ is called a *genotype*
- A string of $\{0, 1\}^*$ is called a *haplotype*
- Two equal length haplotypes generate a unique genotype
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture *haplotype*s that generated it
  (observe that, e.g., $0110, 0001 \Rightarrow 0222$)

Haplotype Inference

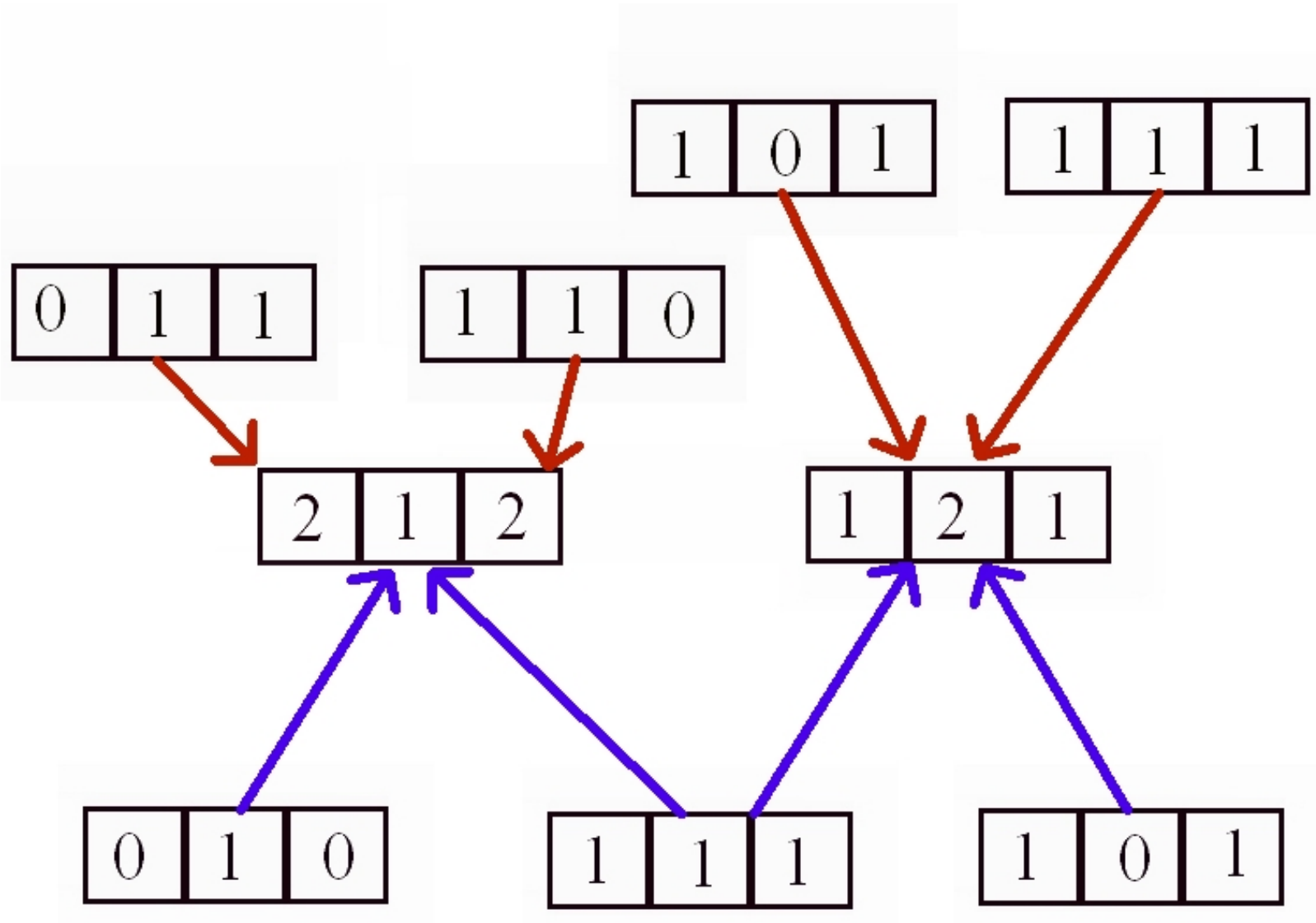
- A string of $\{0, 1, 2\}^*$ is called a *genotype*
- A string of $\{0, 1\}^*$ is called a *haplotype*
- Two equal length haplotypes generate a unique genotype
  E.g., $0010, 0101 \Rightarrow 0222$
- If we have a genotype, we can only conjecture *haplotype*s that generated it
  (observe that, e.g., $0110, 0001 \Rightarrow 0222$)
- Biological experiments allow us to know genotypes!
- Investigating sets of genotypes for a population we can understand the relationships between SNPs and physical features as well as medical information
- Since genotypes are introduced in evolution, is is reasonable to find minimal sets of haplotypes explaining the known genotypes.

Haplotype Inference

Haplotype Inference



Haplotype Inference



Haplotype Inference

- Let $\mathcal{H} = \{0,1\}^*$ be the set of *haplotypes* and
- $\mathcal{G} = \{0,1,2\}^*$ be the set of *genotypes*.
- Given $h_1, h_2 \in \mathcal{H}$ and $g \in \mathcal{G}$, $\{h_1, h_2\}$ explains g if and only if $|h_1| = |h_2| = |g|$ (let's say $n = |g|$) and $\forall i \in [1..n]$:

$$\begin{array}{lcl} g[i] \leq 1 & \longrightarrow & h_1[i] = h_2[i] = g[i] \\ g[i] = 2 & \longrightarrow & h_1[i] \neq h_2[i] \end{array}$$

- A set of haplotypes $H \subseteq \mathcal{H}$ explains a set of genotypes $G \subseteq \mathcal{G}$ if for all $g \in G$ there are $h_1, h_2 \in H$ such that $\{h_1, h_2\}$ explains g.
- Given a set of genotypes $G \subseteq \mathcal{G}$ and an integer k, the *haplotype inference problem (HIP)* by pure parsimony is the problem of finding a set $H \subseteq \mathcal{H}$ that explains G and such that $|H| = k$ (decision version).

Haplotype Inference

The ILP modeling

- 0, 1, 2 are arbitrary values
- Let us swap the roles of 1 and 2 in genotypes, namely 1 is used of mismatch (and 2 stands for 1)
- Given $h_1, h_2 \in \mathcal{H}$ and $g \in \mathcal{G}$, $\{h_1, h_2\}$ explains g if and only if $|h_1| = |h_2| = |g| = n$ and $\forall i \in [1..n]$:

$$\begin{array}{lcl} g[i] = 0 & \longrightarrow & h_1[i] = h_2[i] = 0 \\ g[i] = 2 & \longrightarrow & h_1[i] = h_2[i] = 1 \\ g[i] = 1 & \longrightarrow & \{h_1[i], h_2[i]\} = \{0, 1\} \end{array}$$

Haplotype Inference

The ILP modeling

- 0, 1, 2 are arbitrary values
- Let us swap the roles of 1 and 2 in genotypes, namely 1 is used of mismatch (and 2 stands for 1)
- Given $h_1, h_2 \in \mathcal{H}$ and $g \in \mathcal{G}$, $\{h_1, h_2\}$ explains g if and only if $|h_1| = |h_2| = |g| = n$ and $\forall i \in [1..n]$:

$$\begin{array}{lcl} g[i] = 0 & \longrightarrow & h_1[i] = h_2[i] = 0 \\ g[i] = 2 & \longrightarrow & h_1[i] = h_2[i] = 1 \\ g[i] = 1 & \longrightarrow & \{h_1[i], h_2[i]\} = \{0, 1\} \end{array}$$

- Therefore $g[i] = h_1[i] + h_2[i]$
- This simplifies an ILP encoding. Experimentally it does not improve CP speed-up. Just forget it in this school.

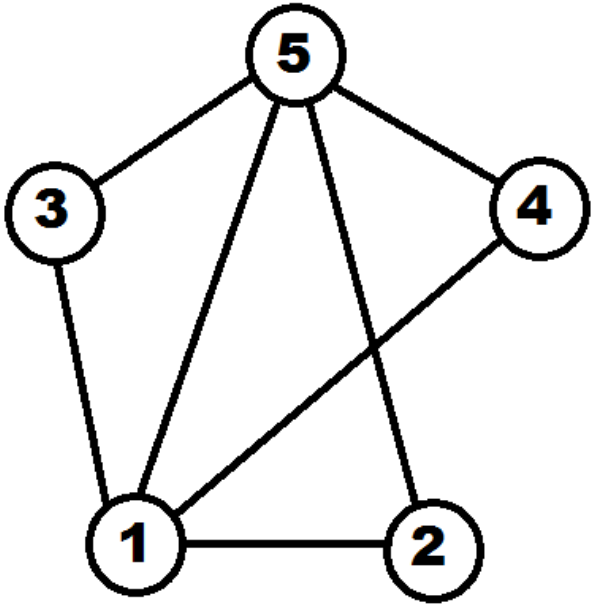
Haplotype Inference by Pure Parsimony

Use of such a parsimony criterion is consistent with the fact that the number of distinct haplotypes observed in most natural populations is vastly smaller than the number of possible haplotypes; this is expected given the plausible assumptions that the mutation rate at each site is small and recombinations rate are low.

[Gusfield and Orzack, 2006]

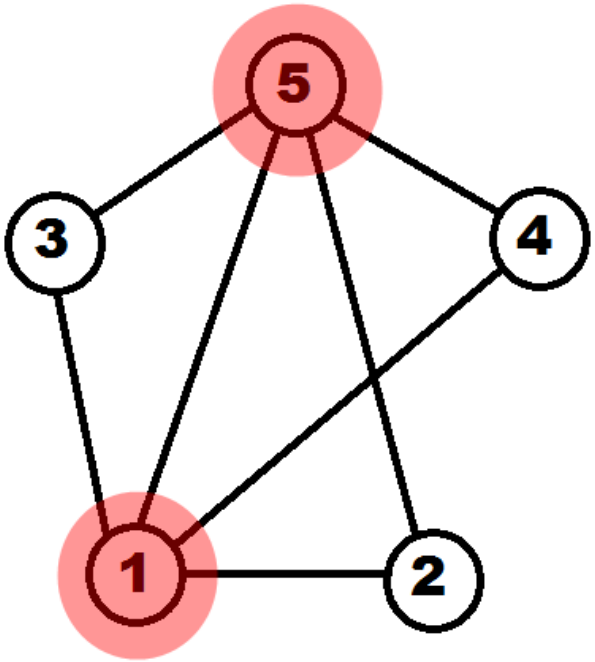
Haplotype Inference by Pure Parsimony

NP-completeness (sketch — see [LPR04])



Haplotype Inference by Pure Parsimony

NP-completeness (sketch — see [LPR04])

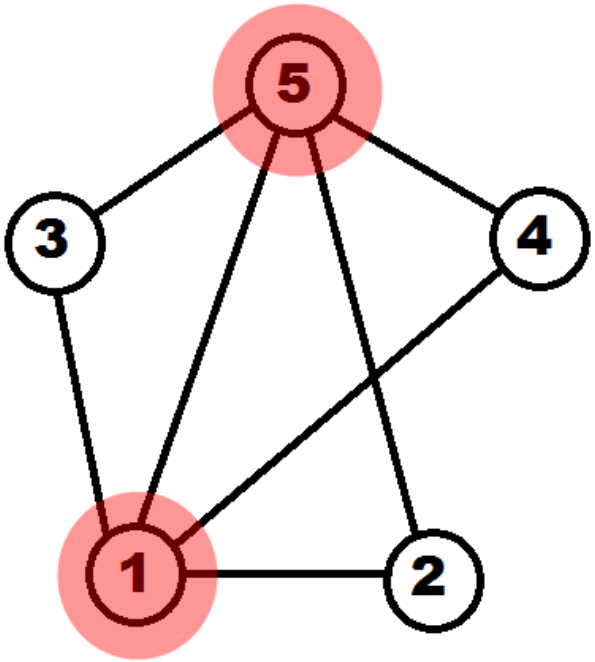


Vertex cover of cardinality 2

Haplotype Inference by Pure Parsimony

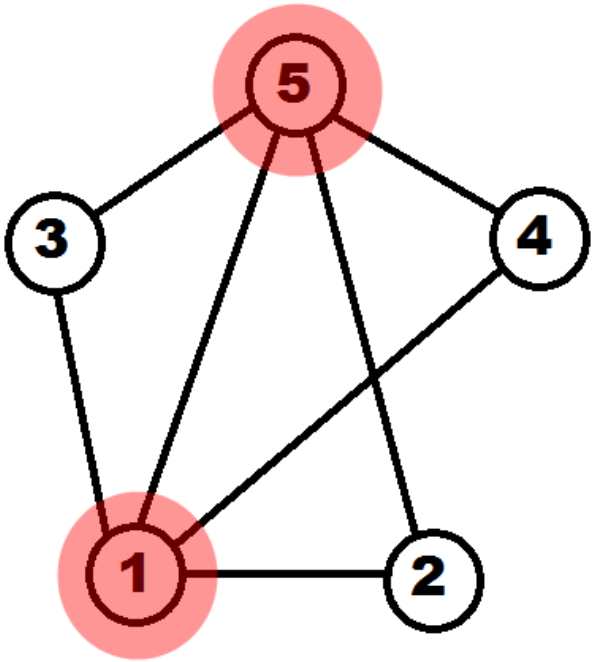
NP-completeness (sketch — see [LPR04])

	1	2	3	4	5	
1	0	1	1	1	1	0



Haplotype Inference by Pure Parsimony

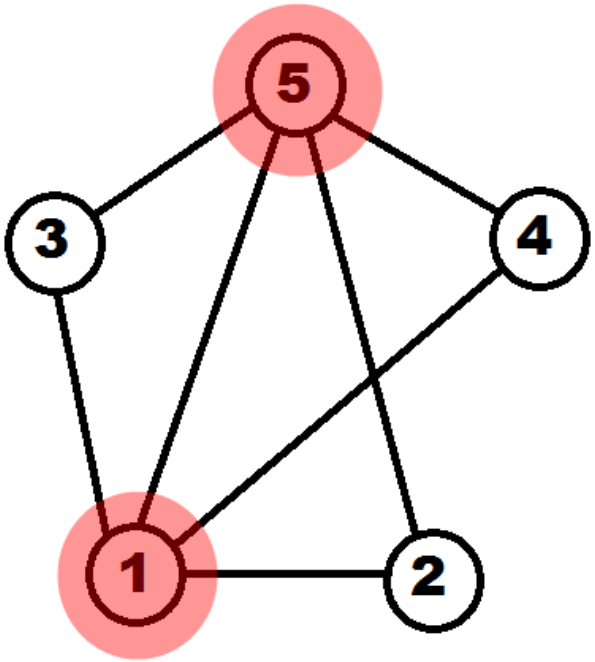
NP-completeness (sketch — see [LPR04])



	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0

Haplotype Inference by Pure Parsimony

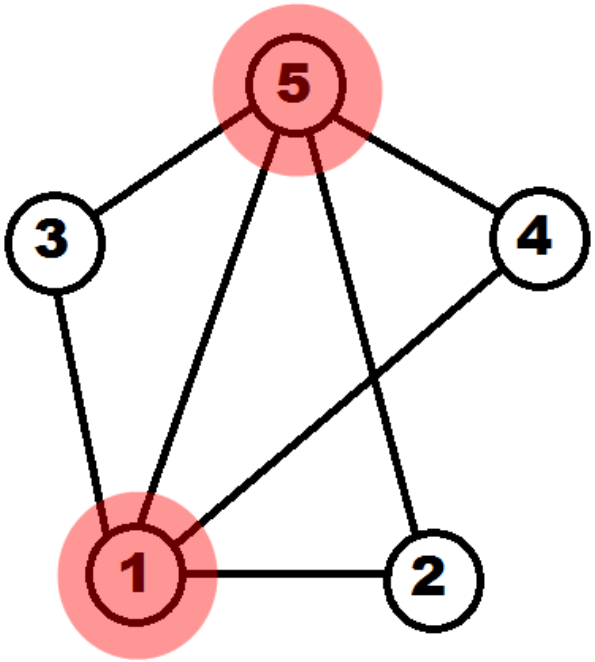
NP-completeness (sketch — see [LPR04])



	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2

Haplotype Inference by Pure Parsimony

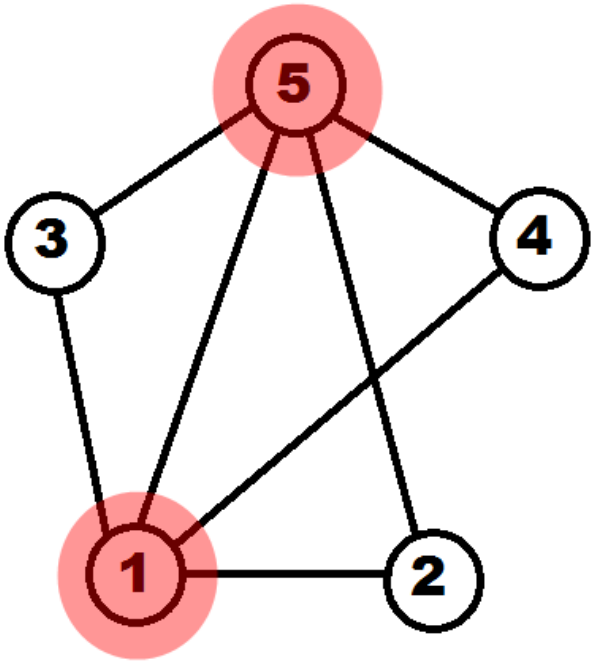
NP-completeness (sketch — see [LPR04])



	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2

Haplotype Inference by Pure Parsimony

NP-completeness (sketch — see [LPR04])



	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2
(1,4)	2	1	1	2	1	2
(1,5)	2	1	1	1	2	2
(2,5)	1	2	1	1	2	2
(3,5)	1	1	2	1	2	2
(4,5)	1	1	1	2	2	2

Haplotype Inference by Pure Parsimony

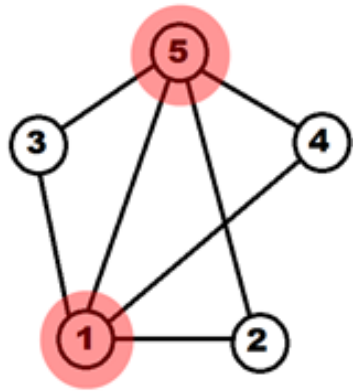
NP-completeness (sketch — see [LPR04])

	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2
(1,4)	2	1	1	2	1	2
(1,5)	2	1	1	1	2	2
(2,5)	1	2	1	1	2	2
(3,5)	1	1	2	1	2	2
(4,5)	1	1	1	2	2	2

0 1 1 1 1 0
1 0 1 1 1 0
1 1 0 1 1 0
1 1 1 0 1 0
1 1 1 1 0 0

0 1 1 1 1 1

1 1 1 1 0 1



Vertex cover for $G = \langle N, E \rangle$ of cardinality $k \Rightarrow |H| = |N| + k$.

Haplotype Inference by Pure Parsimony

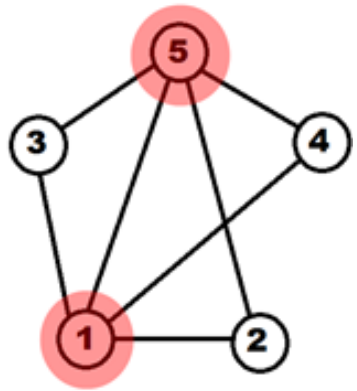
NP-completeness (sketch — see [LPR04])

	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2
(1,4)	2	1	1	2	1	2
(1,5)	2	1	1	1	2	2
(2,5)	1	2	1	1	2	2
(3,5)	1	1	2	1	2	2
(4,5)	1	1	1	2	2	2

0 1 1 1 1 0
1 0 1 1 1 0
1 1 0 1 1 0
1 1 1 0 1 0
1 1 1 1 0 0

0 1 1 1 1 1

1 1 1 1 0 1

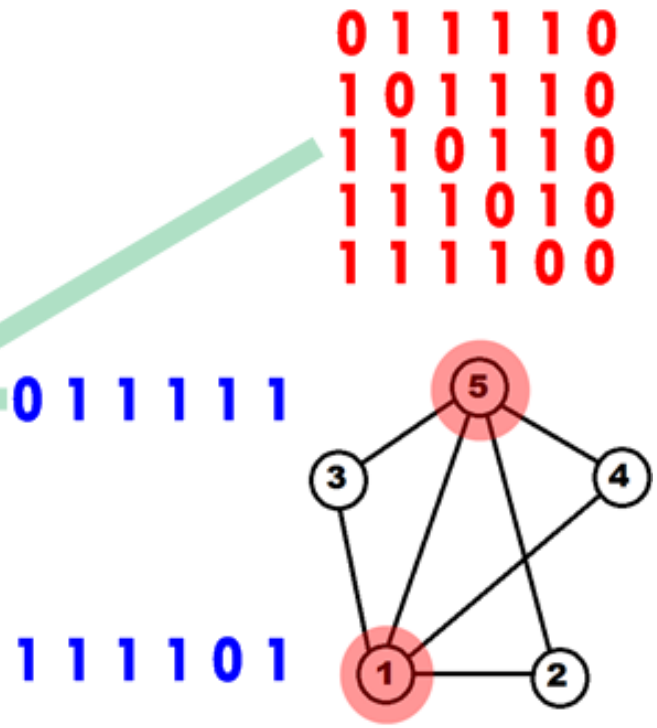


Vertex cover for $G = \langle N, E \rangle$ of cardinality $k \Rightarrow |H| = |N| + k$.

Haplotype Inference by Pure Parsimony

NP-completeness (sketch — see [LPR04])

	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2
(1,4)	2	1	1	2	1	2
(1,5)	2	1	1	1	2	2
(2,5)	1	2	1	1	2	2
(3,5)	1	1	2	1	2	2
(4,5)	1	1	1	2	2	2

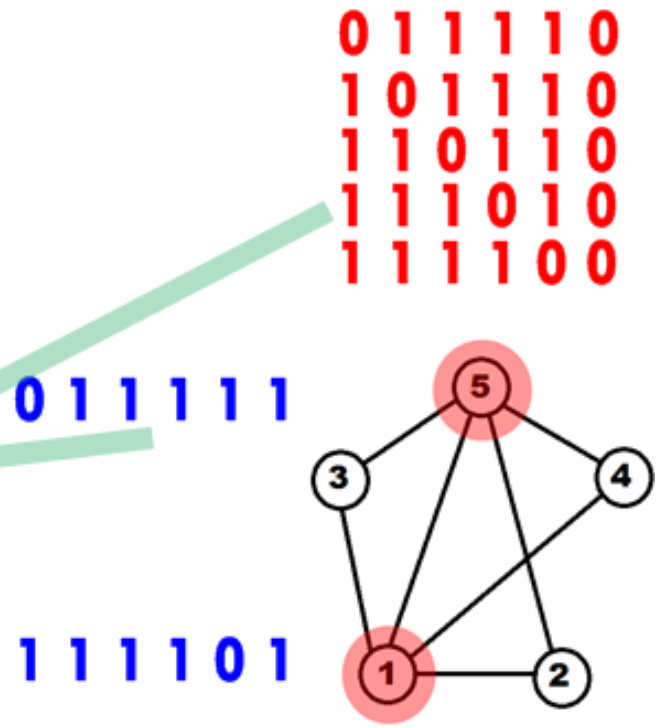


Vertex cover for $G = \langle N, E \rangle$ of cardinality $k \Rightarrow |H| = |N| + k$.

Haplotype Inference by Pure Parsimony

NP-completeness (sketch — see [LPR04])

	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2
(1,4)	2	1	1	2	1	2
(1,5)	2	1	1	1	2	2
(2,5)	1	2	1	1	2	2
(3,5)	1	1	2	1	2	2
(4,5)	1	1	1	2	2	2

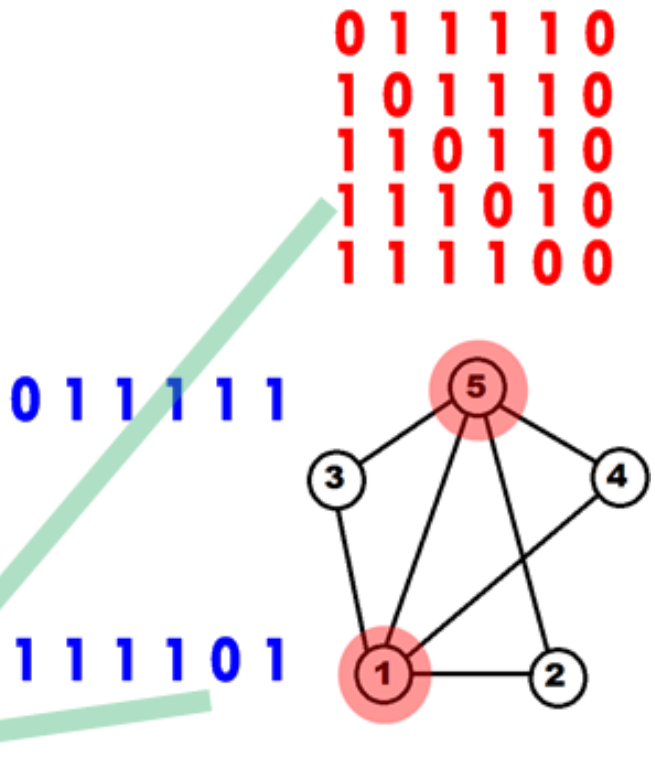


Vertex cover for $G = \langle N, E \rangle$ of cardinality $k \Rightarrow |H| = |N| + k$.

Haplotype Inference by Pure Parsimony

NP-completeness (sketch — see [LPR04])

	1	2	3	4	5	
1	0	1	1	1	1	0
2	1	0	1	1	1	0
3	1	1	0	1	1	0
4	1	1	1	0	1	0
5	1	1	1	1	0	0
(1,2)	2	2	1	1	1	2
(1,3)	2	1	2	1	1	2
(1,4)	2	1	1	2	1	2
(1,5)	2	1	1	1	2	2
(2,5)	1	2	1	1	2	2
(3,5)	1	1	2	1	2	2
(4,5)	1	1	1	2	2	2



Vertex cover for $G = \langle N, E \rangle$ of cardinality $k \Rightarrow |H| = |N| + k$.

Haplotype Inference

1st CP encoding

- Let us focus on the decisional version: Is there an explanation for G with k haplotypes?
- Generate k vectors of 0-1 FD variables $H_1, \ldots, H_k$ of length n
- Add a lexicographical constraint on $H_1, \ldots, H_k$.
- Build a constraint of the form:

$$\forall G_i \in G \, \exists H_{i_1} \exists H_{i_2} \text{ s.t. } \langle H_{i_1}, H_{i_2} \rangle \text{ explain } G_i$$

- Basically, for each i, i_1, i_2 we have a flag $F^i_{i_1,i_2}$ true iff

$$\bigwedge_{j=1}^{n} \left(\begin{array}{l} G_i[j] \leq 1 \rightarrow (H_{i_1}[j] = H_{i_2}[j] = G_i[j]) \wedge \\ G_i[j] = 2 \rightarrow (H_{i_1}[j] \neq H_{i_2}[j]) \end{array} \right)$$

- Then forall $i \in [1..|G|]$: $\sum_{i_1,i_2} F^i_{i_1,i_2} \geq 1$

Haplotype Inference

2nd CP encoding

- Let us focus on the decisional version: Is there an explanation for G with k haplotypes?
- Generate $m = 2|G|$ vectors of 0-1 FD variables $H_1, \ldots, H_m$ of length n
- Add a lexicographical constraint on pairs $(H_1, H_2), (H_3, H4), \ldots, (H_{m-1}, H_m)$ (we can have repetitions now!)
- Build a constraint of the form:

$$(\forall G_i \in G)\,(\langle H_{2i-1}, H_{2i}\rangle \text{ explain } G)$$

- Namely, again,

$$\bigwedge_{j=1}^{n} \left(\begin{array}{l} G_i[j] \leq 1 \rightarrow (H_{2i_1}[j] = H_{i_2}[j] = G_{2i}[j]) \wedge \\ G_i[j] = 2 \rightarrow (H_{2i_1}[j] \neq H_{2i}[j]) \end{array} \right)$$

- We need to state (using constraints!) that $|\{H_1, \ldots, H_m\}| = k$.

Haplotype Inference

2nd CP encoding

- Let us focus on the decisional version: Is there an explanation for G with k haplotypes?
- Generate $m = 2|G|$ vectors of 0-1 FD variables $H_1, \ldots, H_m$ of length n
- Add a lexicographical constraint on pairs $(H_1, H_2), (H_3, H4), \ldots, (H_{m-1}, H_m)$ (we can have repetitions now!)
- Build a constraint of the form:

$$(\forall G_i \in G)\,(\langle H_{2i-1}, H_{2i}\rangle \text{ explain } G)$$

- Namely, again,

$$\bigwedge_{j=1}^{n} \left(\begin{array}{l} G_i[j] \leq 1 \rightarrow (H_{2i_1}[j] = H_{i_2}[j] = G_{2i}[j]) \wedge \\ G_i[j] = 2 \rightarrow (H_{2i_1}[j] \neq H_{2i}[j]) \end{array} \right)$$

- We need to state (using constraints!) that $|\{H_1, \ldots, H_m\}| = k$.
- This is a good constraint exercise.

Haplotype Inference

2nd CP encoding

- For $a, b \in [1..m]$ we set $F_{a,b} \leftrightarrow \bigwedge_{i=1}^{n} H_a[i] = H_b[i]$.
- Namely $F_{a,b}$ is a Boolean variable that is true iff H_a and H_b will be equal in the solution

Haplotype Inference

2nd CP encoding

- For $a, b \in [1..m]$ we set $F_{a,b} \leftrightarrow \bigwedge_{i=1}^{n} H_a[i] = H_b[i]$.
- Namely $F_{a,b}$ is a Boolean variable that is true iff H_a and H_b will be equal in the solution
- Then define $M_a \leftrightarrow \bigvee_{b=a+1}^{m} F_{a,b}$
- M_a is again a Boolean variable that is true if and only if there is another vector in $H_{a+1}, H_{a+2}, \ldots, H_m$ equal to H_a

Haplotype Inference

2nd CP encoding

- For $a, b \in [1..m]$ we set $F_{a,b} \leftrightarrow \bigwedge_{i=1}^{n} H_a[i] = H_b[i]$.
- Namely $F_{a,b}$ is a Boolean variable that is true iff H_a and H_b will be equal in the solution
- Then define $M_a \leftrightarrow \bigvee_{b=a+1}^{m} F_{a,b}$
- M_a is again a Boolean variable that is true if and only if there is another vector in $H_{a+1}, H_{a+2}, \ldots, H_m$ equal to H_a
- The size of H can be therefore expressed as $\sum_{a=1}^{n}(1 - M_a)$ (viewing Boolean truth values as 0/1)

Haplotype Inference

2nd CP encoding

- For $a, b \in [1..m]$ we set $F_{a,b} \leftrightarrow \bigwedge_{i=1}^{n} H_a[i] = H_b[i]$.
- Namely $F_{a,b}$ is a Boolean variable that is true iff H_a and H_b will be equal in the solution
- Then define $M_a \leftrightarrow \bigvee_{b=a+1}^{m} F_{a,b}$
- M_a is again a Boolean variable that is true if and only if there is another vector in $H_{a+1}, H_{a+2}, \ldots, H_m$ equal to H_a
- The size of H can be therefore expressed as $\sum_{a=1}^{n}(1 - M_a)$ (viewing Boolean truth values as 0/1)
- What is the best encoding? Try codes `clp_direct.pl` and `clp_second.pl` (0, 1, and 2 have the original meaning here).
- Search space: $O(2^{nk})$, with $k < 2|G|$ (1), $O(2^{2n|G|})$ (2).
- Call goals like `:- input(1,Gs),haplo_decision(Gs,5).` or `:- haplo_decision([[0,1,2],[0,2,1]],3).`

Haplotype Inference

Some References

- Gusfield and Orzack. Haplotype Inference (Survey, and ILP formulations) In CRC Handbook on Bioinformatics, 2006
- Lancia, Pinotti, Rizzi. [LPR04] Haplotyping Populations by Pure Parsimony: Complexity of Exact and Approximation Algorithms. INFORMS Journal on Computing 16(4):348–359, 2004.
- Graça, Marques-Silva, Lynce, Oliveira. Several works on SAT-based and specialized 0-1 ILP for Haplotype Inference. (e.g. WCB 08, WCB 09)
- Di Gaspero, Roli. Stochastic local search for large-scale instances of the haplotype inference problem by pure parsimony. J. Algorithms 63(1-3): 55-69 (2008) (also in WCB08).
- Erdem, Erdem, Türe. HAPLO-ASP: Haplotype Inference Using Answer Set Programming. LPNMR 2009: 573–578
- James Cussens Maximum likelihood pedigree reconstruction using integer programming. WCB 10.