

# Haplotyping problems with “complex data”

- Missing entries
- Data generated by complex biology, such as recombination or recurrent mutation
- Genotype (conflated) sequences, rather than simpler haplotype sequences

Most of these problems are NP-hard, although some elegant poly-time solutions exist (and are well-known) for simpler data.

# Problem M1: Missing data

Given ternary sequences (0s, 1s, ?s), change the ?s to 0s and 1s in order to **minimize** the resulting number of incompatible pairs of sites.

(Special case) Perfect Phylogeny with Missing Data: Determine if the ?s can be set so that there are **no** resulting incompatibilities. NP-hard in general, but if the root of the required perfect phylogeny is specified, then the problem has an elegant poly-time solution (Pe'er, Sharan, Shamir).

# ILP for the missing data problem

Create a binary variable  $Y(i,p)$  for a ? in cell  $(i,p)$ ,  
indicating whether the cell will be set to 0 or to 1.

For each pair of sites  $p, q$  that **could be made**  
incompatible, let  $D(p,q)$  be the set of missing or  
**deficient** gametes in site pair  $p,q$ .

For each gamete  $a,b$  in  $D(p,q)$ , create the binary  
variable  $B(p,q,a,b)$ ,  
and create inequalities to set it to 1 **if** the  $Y$  variables  
for cells for sites  $p,q$  are set so that gamete  $a,b$  is  
created in **some** row for sites  $p,q$ .

## Example

p q	
0 0	$D(p,q) = \{1,1; 0,1\}$
? 1	
1 0	
? ?	
? 0	
0 ?	
---	
0 0	

To set the B variables, the ILP will have inequalities for each a,b in  $D(p,q)$ , one for each row where a,b could be created at site p,q.

For example, for a,b = 1,1 the ILP has:

$$Y(2,p) \leq B(p,q,1,1) \quad \text{for row 2}$$

$$Y(4,p) + Y(4,q) - B(p,q,1,1) \leq 1 \quad \text{for row 4}$$

## Example continued

p q  
----  
0 0  
? 1  
1 0  
? ?  
? 0  
0 ?

$D(p,q) = \{1,1; 0,1\}$

For  $a,b = 0,1$  the ILP has:

$$Y(2,p) + B(p,q,0,1) \Rightarrow 1 \quad \text{for row 2}$$

$$Y(4,q) - Y(4,p) - B(p,q,0,1) \leq 0 \quad \text{for row 4}$$

$$Y(6,q) - B(p,q,0,1) \leq 0 \quad \text{for row 6}$$

The ILP also has a variable  $C(p,q)$  which is set to 1 if **every** gamete in  $D(p,q)$  is created at site-pair  $p,q$ .

In the example:

$$B(p, q, 1, 1) + B(p, q, 0, 1) - C(p,q) \leq 1$$

So,  $C(p,q)$  is set to 1 **if** (but not only if) the  $Y$  variables for sites  $p, q$  (missing entries in columns  $p, q$ ) are set so that sites  $p$  and  $q$  become incompatible.

If  $M$  is an  $n$  by  $m$  matrix, then we have at most  $nm$   $Y$  variables;  $2m^2$   $B$  variables;  $m^2/2$   $C$  variables; and  $O(nm^2)$  inequalities in worst-case.

Finally, we have the objective function:

$$\text{Minimize } \sum_{(p,q) \text{ in } P} C(p, q)$$

Where  $P$  is the set of site-pairs that could be made to be incompatible.

# Problems related to M1

- Site-Removal Problem for **complete** data:  
Remove the minimum number of sites from the data, so that **no** incompatibilities remain. This is a common approach to incompatible data in phylogenetics. NP-hard.
- Site-Removal Problem with **missing** data (S1):  
Impute values for the missing entries to minimize the solution to the resulting Site-Removal Problem for complete data.

# ILP for S1 - a simple extension to M1

- For each site  $i$ , let  $D(i)$  be a variable set to 1 if and only if site  $i$  is removed.
- For each site-pair  $p, q$  in  $P$ , add the inequality  $D(p) + D(q) - C(p, q) \Rightarrow 0$  to the M1 formulation.

The objective function is now

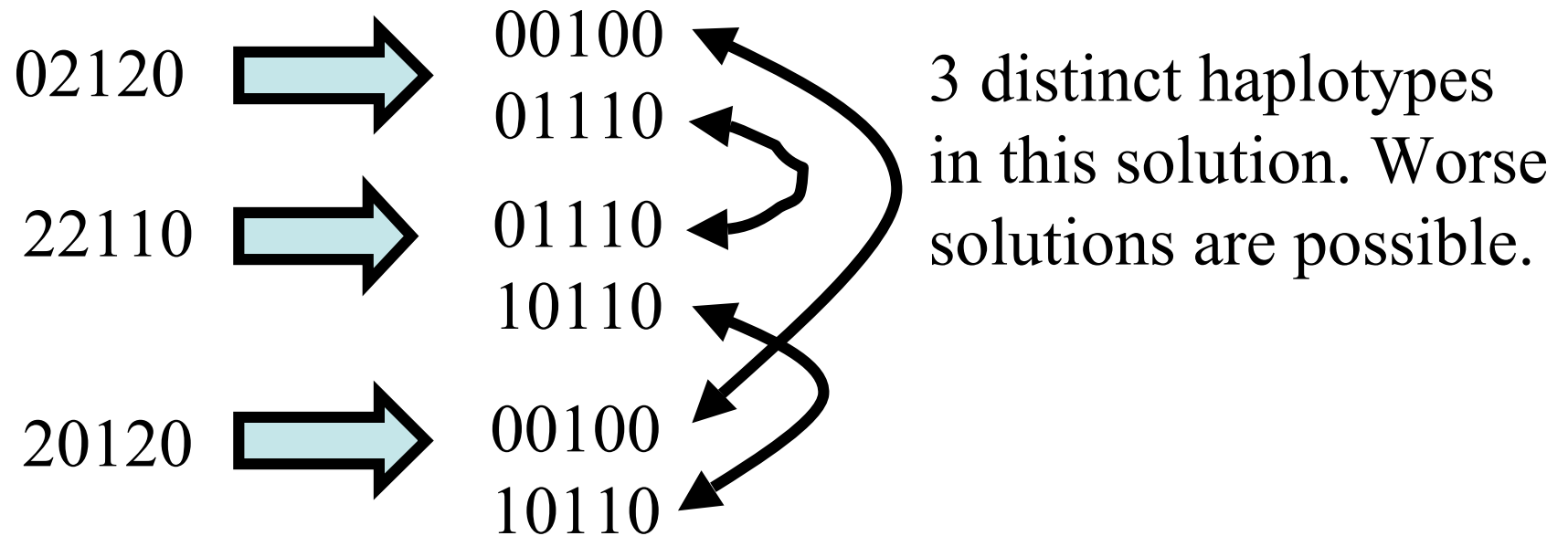
Minimize  $\sum D(i)$

# Another extension: Min PPH

Problem: If there is a PPH solution, find one **minimizing** the number of **distinct** haplotypes used.

Justified by empirical and theoretical grounds: very few distinct haplotypes observed in real populations

# Example of pure parsimony



There is a naïve ILP for Parsimony that works for moderate sized instances, and very clever (worst-case small) ILPs that, unfortunately only solve very small instances (BH and others)

# Practical solution of MinPPH

- MinPPH problem is NP-hard (Bafna, Gusfield, Hannenhali, Yooseph)
- But, it can be solved very efficiently in practice by ILP. The ILP just combines the ILP for the PPH problem and the ILP for Maximum Parsimony Haplotyping.

The interesting point is that the ILP for Maximum Parsimony can only solve tiny problem instances, but the addition of the PPH ILP inequalities makes it solve very efficiently on large problem instances.