



Complexity and Approximation of the Minimum Recombinant Haplotype Configuration Problem

Authors: **Lan Liu**, Xi Chen, Jing Xiao
& Tao Jiang



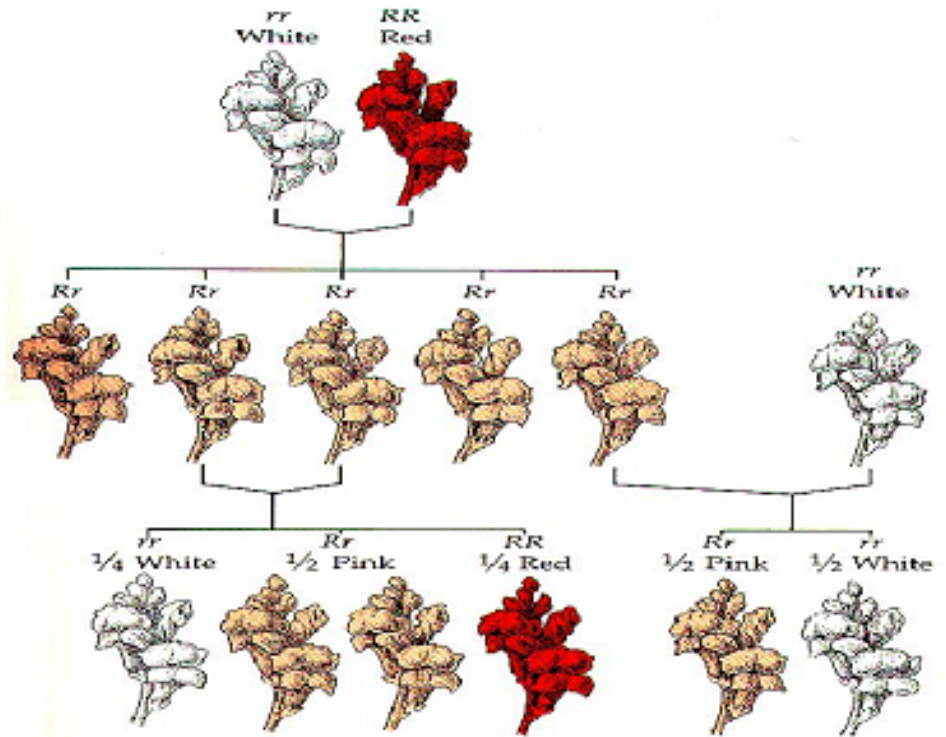
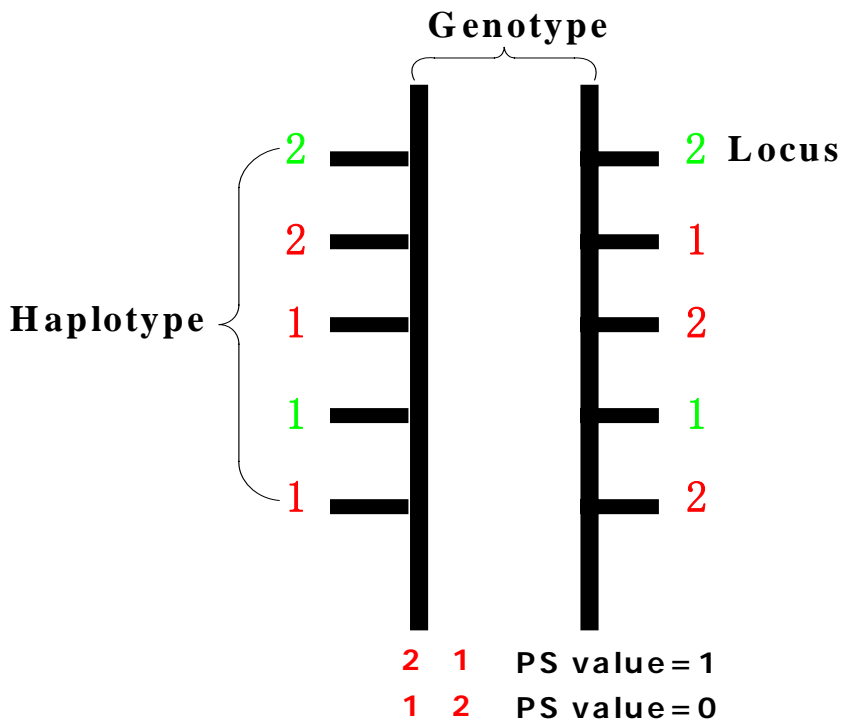
Outline

- Introduction and problem definition
- Deciding the complexity of binary-tree-MRHC
- Approximation of MRHC with missing data
- Approximation of MRHC without missing data
- Approximation of bounded MRHC
- Conclusion

Introduction

- Basic concepts

- Mendelian Law: one haplotype comes from the mother and the other comes from the father.



Example: Mendelian experiment

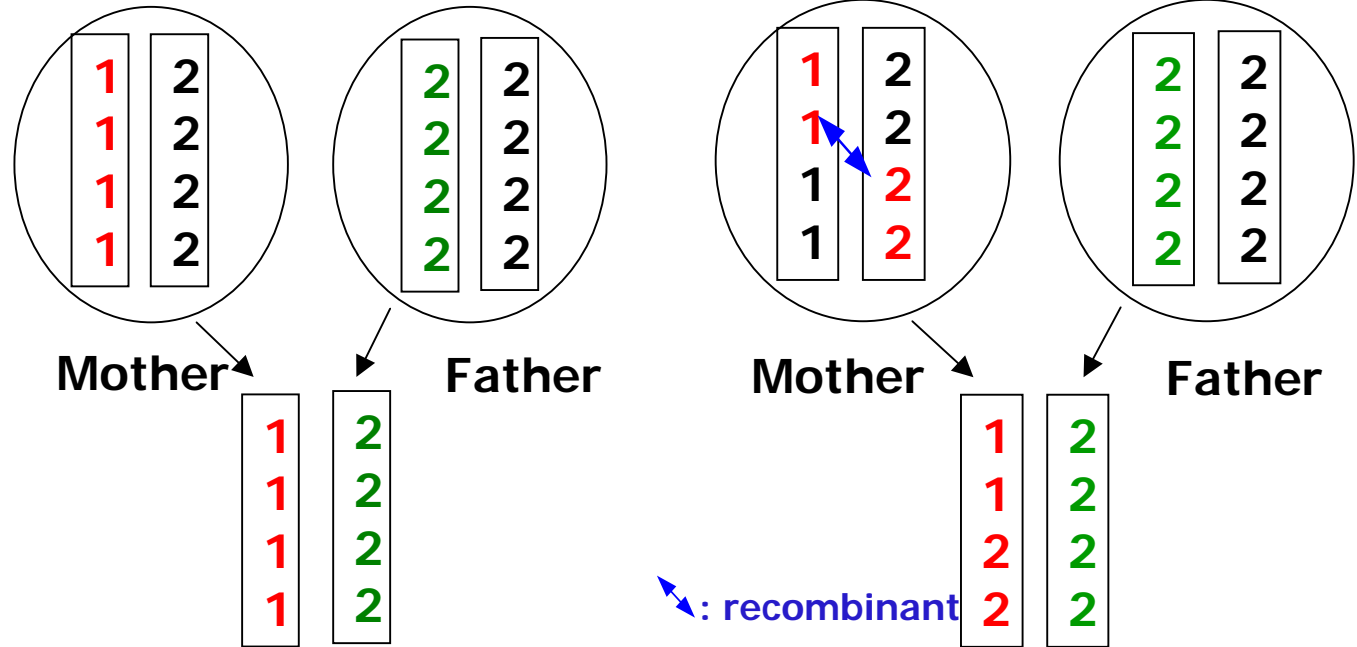
Notations and Recombinant

1 2
1 2
2 2
2 2

Genotype

1 | 2
2 | 1
2 | 2
2 | 2

Haplotype
Configuration

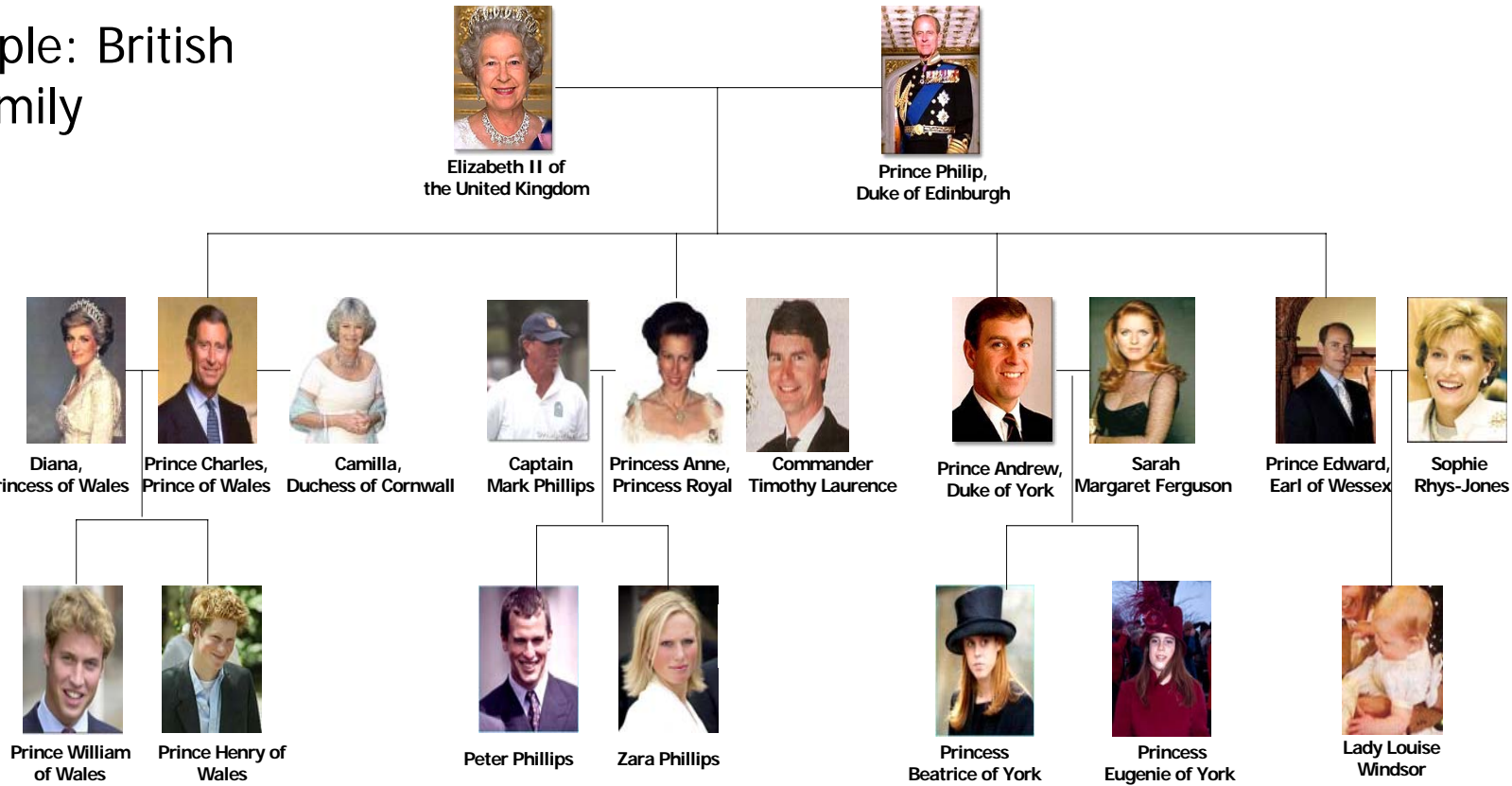
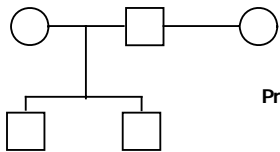


0 recombinant

1 recombinant

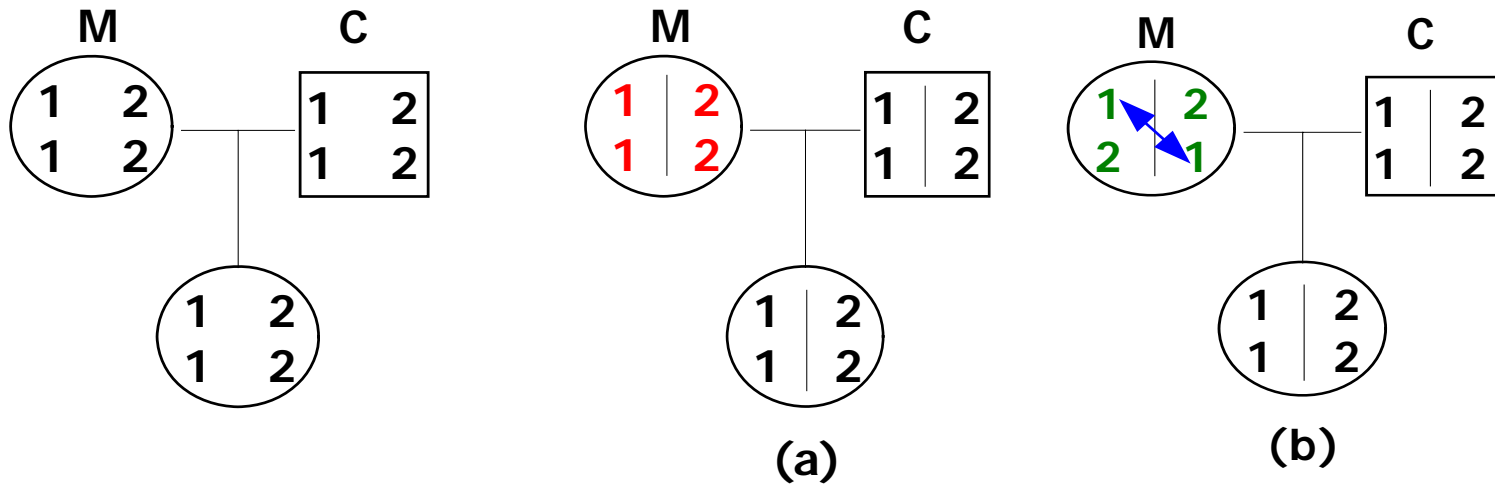
Pedigree

- An example: British Royal Family



Haplotype Reconstruction

- Haplotype: useful, expensive
- Genotype: cheaper
- Reconstruct haplotypes from genotypes





Problem Definition

- MRHC problem

Given a **pedigree** and the **genotype** information for each member, find a **haplotype configuration** for each member which obeys Mendelian law, *s.t.* the number of **recombinants** are minimized.



Problem Definition

- Variants of MRHC
 - Tree-MRHC: no mating loop
 - Binary-tree-MRHC: 1 mate, 1 child
 - 2-locus-MRHC: 2 loci
 - 2-locus-MRHC^{*}: 2 loci with missing data



Previous Work

- The known hardness results for Mendelian law checking

Loop?	Multi-allelic?	Hardness
Yes	Yes	NP-hard [AHI +03]
	No	P [AHI +03]
No		P [AHI +03]

- The known hardness results for MRHC

	Hardness
2-locus-MRHC	NP-hard [LJ03]
Tree-MRHC with bounded #members	P [LJ03]
Tree-MRHC with bounded #loci	P [DLJ03]
Tree-MRHC	NP-hard [DLJ03]



Our hardness and approximation results

	Hardness	Lower bound of approx. ratio	Assumption	The lower bound holds for	Upper bound of approx. ratio
Binary-tree-MRHC	NP				
2-locus-MRHC*		Any $f(n)$	$P \neq NP$	2-locus-MRHC* (4,1)	
Binary-tree-MRHC*		Any $f(n)$	$P \neq NP$	Binary-tree-MRHC* (1,1)	
2-locus-MRHC		Any constant	$P \neq NP$ the Unique Games Conjecture [Khot02]	2-locus-MRHC (16,15)	$O(\sqrt{\log(n)})$
Tree-MRHC		Any constant	$P \neq NP$ the Unique Games Conjecture	Tree-MRHC(1,u) Tree-MRHC(u,1)	



Our hardness and approximation results

	Hardness	Lower bound of approx. ratio	Assumption	The lower bound holds for	Upper bound of approx. ratio
Binary-tree-MRHC	NP				
2-locus-MRHC*		Any $f(n)$	$P \neq NP$	2-locus-MRHC* (4,1)	
Binary-tree-MRHC*		Any $f(n)$	$P \neq NP$	Binary-tree-MRHC* (1,1)	
2-locus-MRHC		Any constant	$P \neq NP$ the Unique Games Conjecture [Khot02]	2-locus-MRHC (16,15)	$O(\sqrt{\log(n)})$
Tree-MRHC		Any constant	$P \neq NP$ the Unique Games Conjecture	Tree-MRHC(1,u) Tree-MRHC(u,1)	



Outline

- Introduction and problem definition
- Deciding the complexity of binary-tree-MRHC
- Approximation of MRHC with missing data
- Approximation of MRHC without missing data
- Approximation of bounded MRHC
- Conclusion



A verifier for $\neq 3$ SAT (1)

Given a truth assignment for literals in a 3CNF formula

- Consistency checking for each variable
- Satisfiability checking for each clause

Binary-tree-MRHC is NP-hard

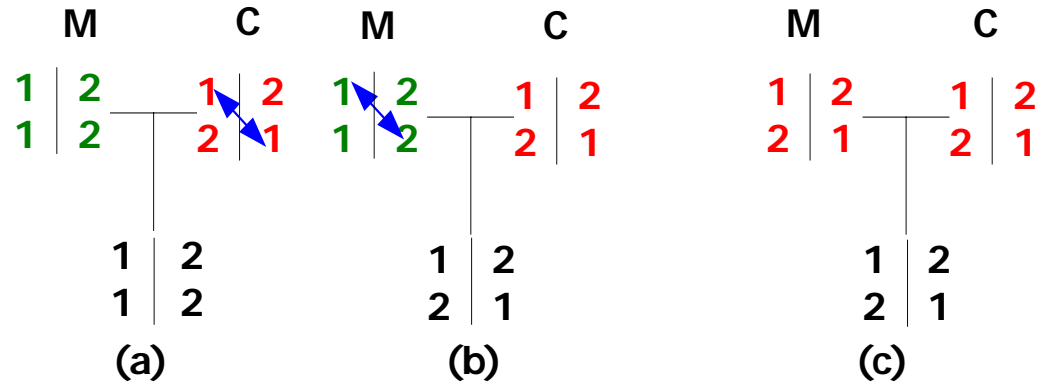
(A) C's genotype

1 2
1 2

(B) Two haplotype configurations

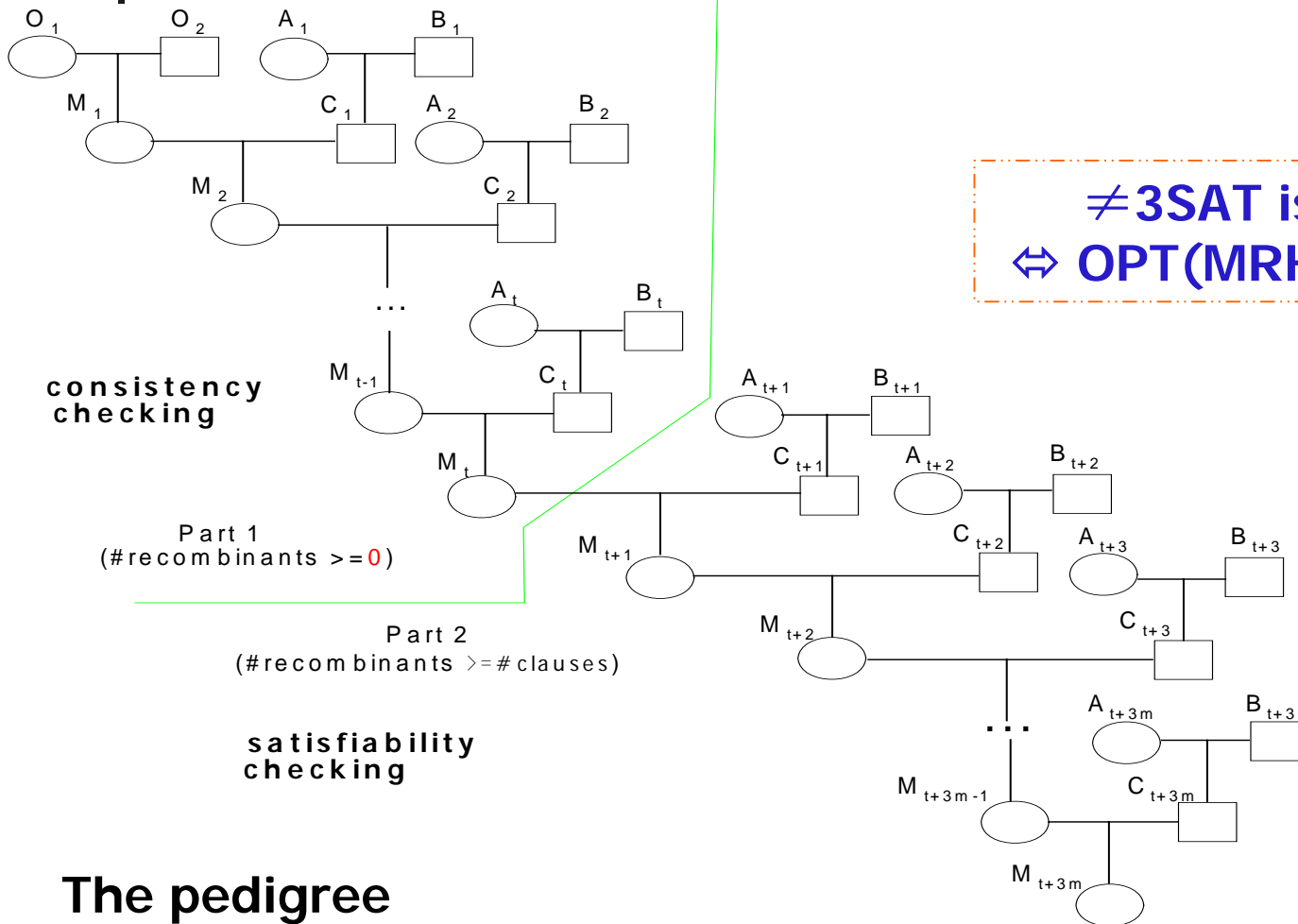
1 | 2
2 | 1

1 | 2
1 | 2



C can check if M have certain haplotype configuration!!

Binary-tree-MRHC is NP-hard



$\neq 3SAT$ is satisfiable
 $\Leftrightarrow OPT(MRHC) = \#clauses$



Outline

- Introduction and problem definition
- Deciding the complexity of binary-tree-MRHC
- Approximation of MRHC with missing data
- Approximation of MRHC without missing data
- Approximation of bounded MRHC
- Conclusion



Inapproximability of 2-locus -MRHC*

- Definition: *A minimization problem R cannot be approximated*
 - There is not an approximation algorithm with ratio $f(n)$ unless $P=NP$.
 - $f(n)$ is any polynomial-time computable function

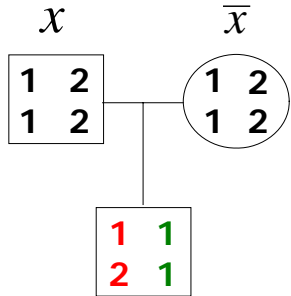
- Fact: *If it is NP-hard to decide whether $OPT(R)=0$, R cannot be approximated unless $P=NP$.*

Inapproximability of 2-locus -MRHC*

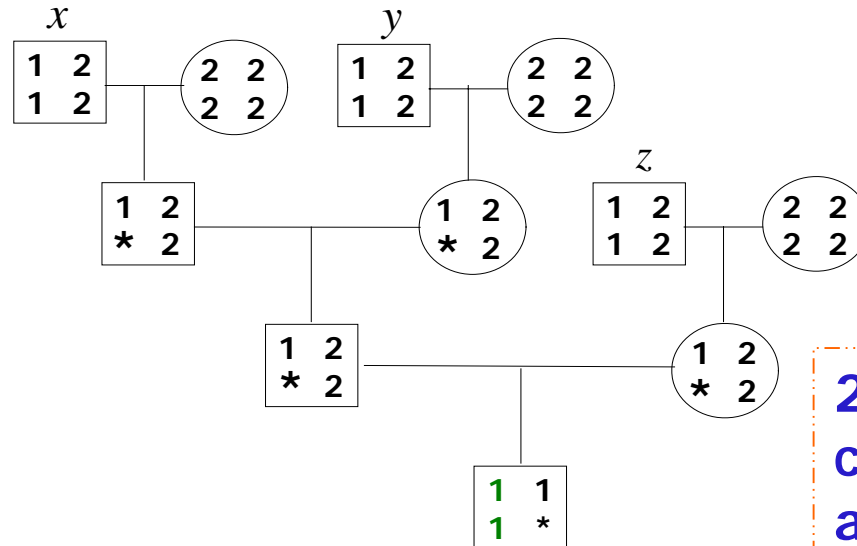
- Reduce 3SAT to 2-locus-MRHC*

$\begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$ True

$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ False



(A) gadget for variable x



(B) gadget for clause $x \vee y \vee z$

**2-locus-MRHC*
cannot be
approximated
unless P=NP!!**

- 3SAT is satisfiable \Leftrightarrow OPT(2-locus-MRHC*) = 0



Outline

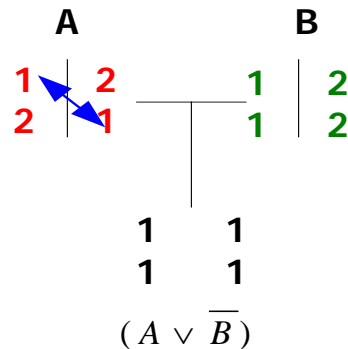
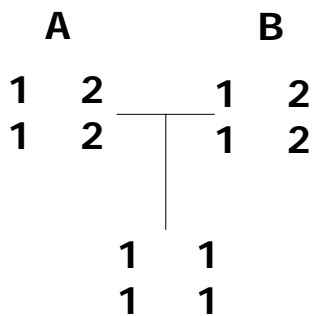
- Introduction and problem definition
- Deciding the complexity of binary-tree-MRHC
- Approximation of MRHC with missing data
- Approximation of MRHC without missing data
- Approximation of bounded MRHC
- Conclusion

Upper Bound of 2-locus-MRHC

- Main idea: use a Boolean **variable** to capture the **configuration**; use **clauses** to capture the **recombinants**.

$$\text{True} \begin{bmatrix} 1 & | & 2 \\ 1 & | & 2 \end{bmatrix} \quad \text{False} \begin{bmatrix} 1 & | & 2 \\ 2 & | & 1 \end{bmatrix}$$

- An example



Upper Bound of 2-locus-MRHC

- The reduction from 2-locus-MRHC to Min 2CNF Deletion

Genotype of the Mother (<i>A</i>)	Genotype of the Father (<i>B</i>)	Genotype of the Child (<i>C</i>)	2CNF Constraint	
1 2 1 2	1 2 1 2	1 1 2 2	$2(A \vee B) \quad (A \vee \bar{B}) \quad (\bar{A} \vee B)$	
		1 1 2 2		
		2 2 1 1	$2(\bar{A} \vee \bar{B}) \quad (A \vee \bar{B}) \quad (\bar{A} \vee B)$	
		1 1 2 2		
2 2 1 1 1 2 1 2	$(\bar{A} \vee \bar{B}) \quad (A \vee B)$			
1 2 1 2 2 2 1 1				
		1 2	$(\bar{A} \vee C) \quad (A \vee \bar{C}) \quad (\bar{B} \vee C) \quad (B \vee \bar{C})$	
		1 2		
1 2 1 2	XX YX YX XX	1 1 2 2	A	
		1 1 2 2		
		2 2 1 1	\bar{A}	
	1 1 2 2			
			1 2	$(\bar{A} \vee C) \quad (A \vee \bar{C})$
			1 2	
	XX YX	XX YX	YX	A
			XX	
	YX XX	YX XX	YX	\bar{A}
			YY	
XX			A	
XY				
		YY	\bar{A}	
		XY		



Upper Bound of 2-locus-MRHC

- Recently, Agarwal *et al.* [STOC05] presented an $O(\sqrt{\log(n)})$ randomized approximation algorithm for Min 2CNF Deletion.
- 2-locus-MRHC has $O(\sqrt{\log(n)})$ approximation algorithm.



Outline

- Introduction and problem definition
- Deciding the complexity of binary-tree-MRHC
- Approximation of MRHC with missing data
- Approximation of MRHC without missing data
- Approximation of bounded MRHC
- Conclusion



Approximation Hardness of bounded MRHC

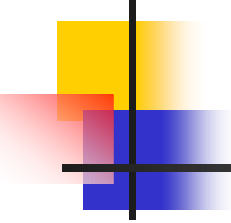
- Bound #mates and #children
 - 2-locus-MRHC: (16,15)
 - **2-locus-MRHC* : (4,1)**
 - tree-MRHC: (u,1) or (1,u)



Conclusion

- Our hardness and approximation results

	Hardness	Lower bound of approx. ratio	Assumption	The lower bound holds for	Upper bound of approx. ratio
Binary-tree-MRHC	NP-hard				
2-locus-MRHC*		Any $f(n)$	$P \neq NP$	2-locus-MRHC* (4,1)	
Binary-tree-MRHC*		Any $f(n)$	$P \neq NP$	Binary-tree-MRHC* (1,1)	
2-locus-MRHC		Any constant	$P \neq NP$ the Unique Games Conjecture	2-locus-MRHC (16,15)	$O(\sqrt{\log(n)})$
Tree-MRHC		Any constant	$P \neq NP$ the Unique Games Conjecture	Tree-MRHC(1,u) Tree-MRHC(u,1)	



Thanks for your time
and attention!

