Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Milan, Italy

Haplotype Inference on Pedigrees with Recombinations and Mutations

WABI 2010

Yuri Pirola, Paola Bonizzoni, Tao Jiang

pirola@disco.unimib.it

Outline

HI on Pedigrees with Recombinations and Mutations:

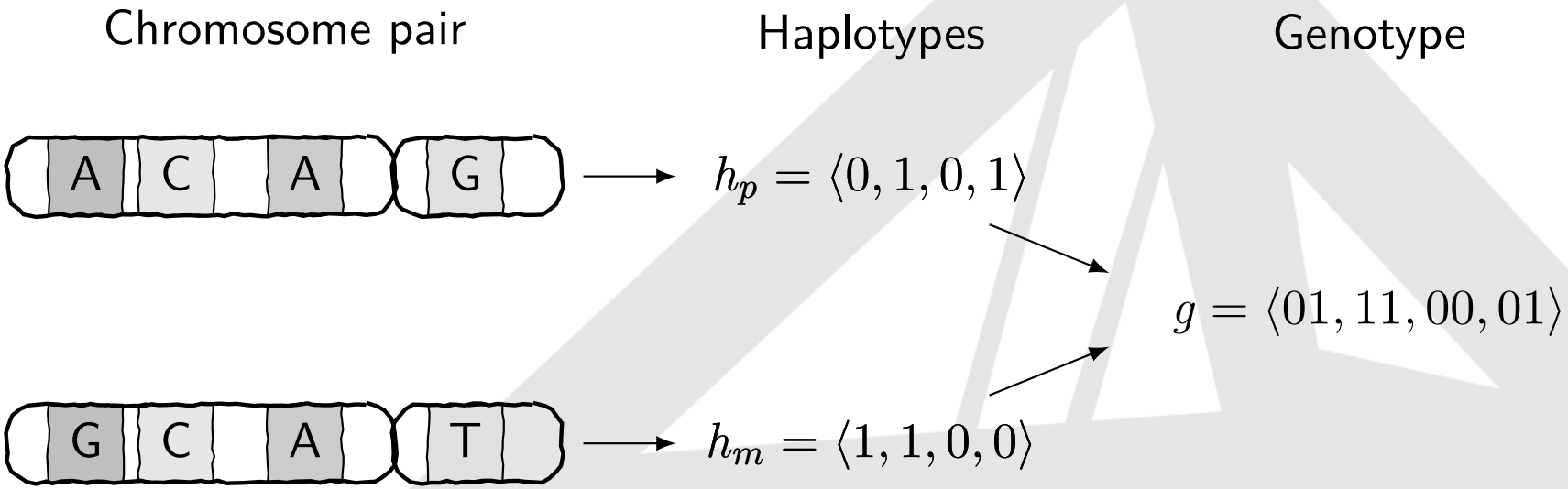
- Introduction
- Background
- *Minimum Change Haplotype Configuration* problem:
  - Heuristic algorithm
  - Experimental evaluation and comparison
- Conclusions and open problems

Our Contribution

Original Contributions:

- *Generalization* of existing models for HI to a more *realistic* setting (MCHC)
- *Efficient* and *effective* heuristic algorithm:
  - for the *new* and the *old* formulations (*MCHC*, *MRHC*, *MMCH*)
  - *well-founded* approach (based on commonly-used algorithms)

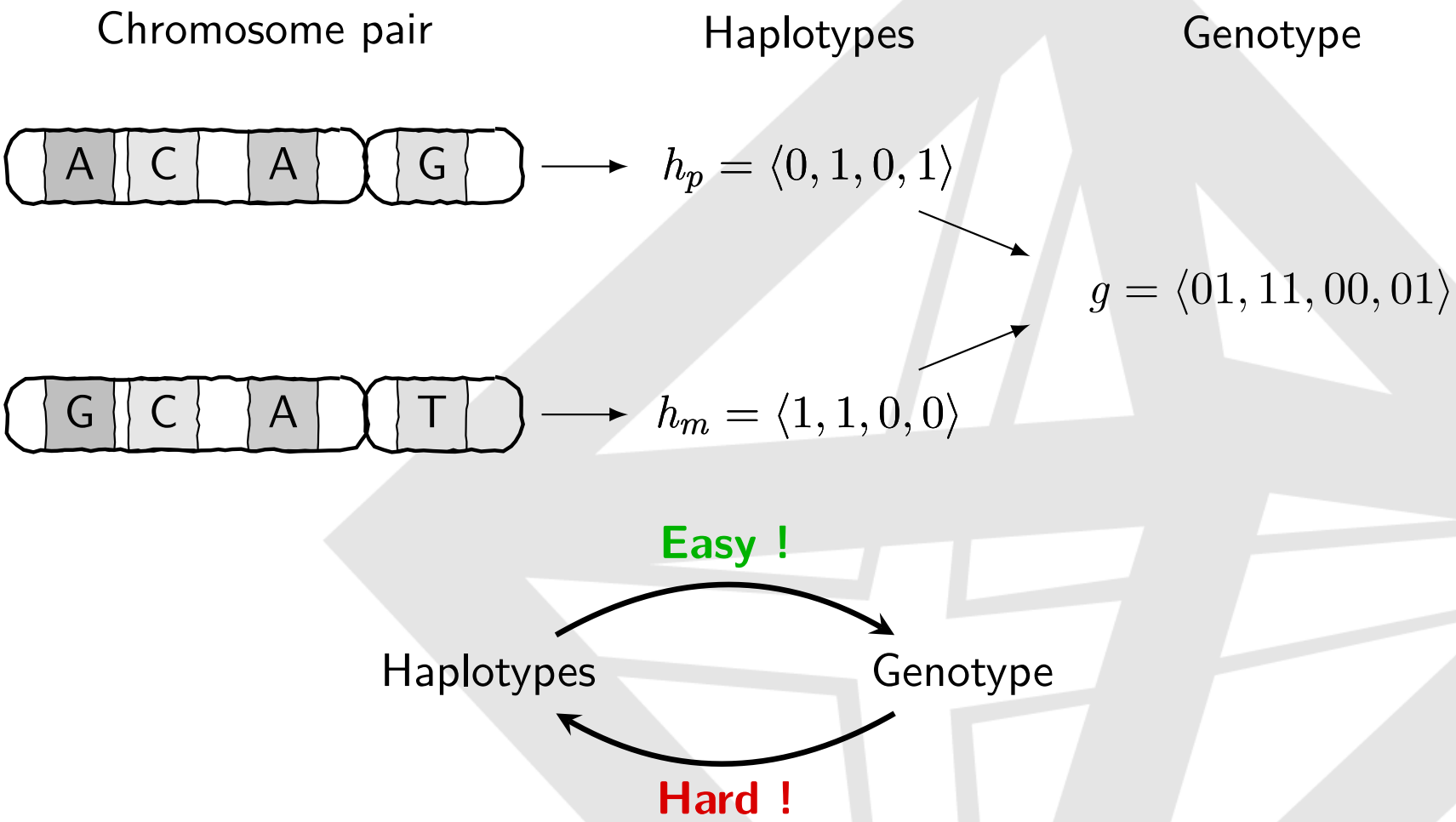
The two main "characters"



Haplotypes: **useful** (e.g., genetic mapping, association studies, . . .)

Genotypes: easy to collect

The two main "characters"



Haplotype Inference problem

Problem (Haplotype Inference)

Given the genotypes of a ***population****, to recover (=infer) the pairs of haplotypes of each individual.*

Different *kinds of populations* and *genetic models*

$\Rightarrow$

Different *computational problems*

General Overview

Haplotype Inference methods:

Approach	Population: Unstructured	Population: Structured
Statistical		
Combinatorial		

Reviews: (Gao *et al.*, Hum. Her., 2009), (Gusfield, RECOMB, 2002), and several others.

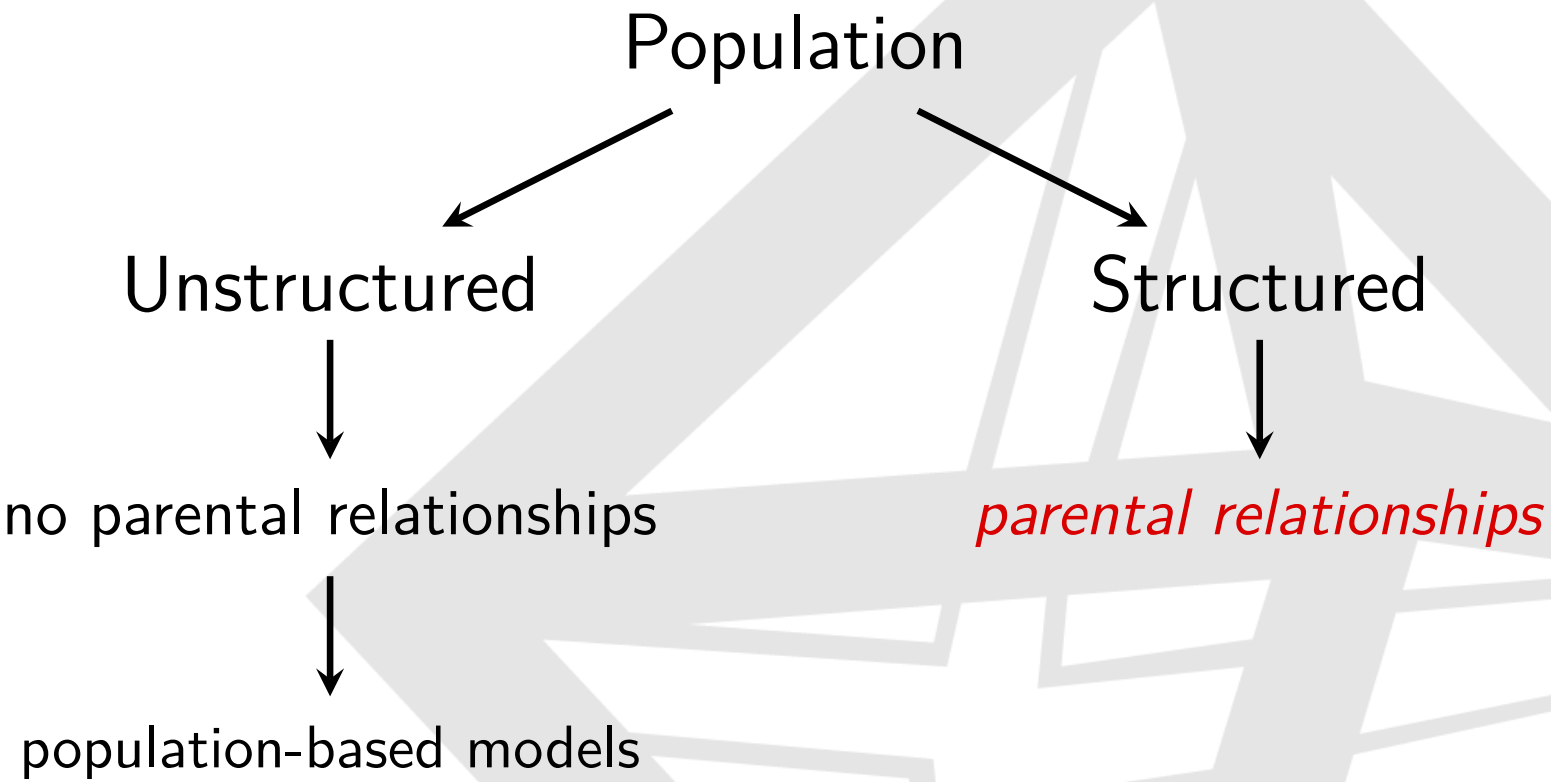
General Overview

Haplotype Inference methods:

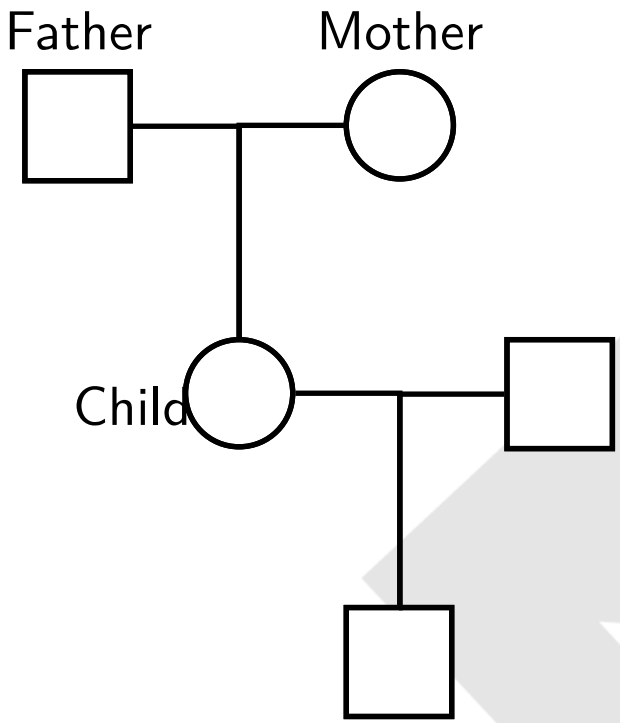
Approach	Population: Unstructured	Population: Structured
Statistical		
Combinatorial		✗

Reviews: (Gao *et al.*, Hum. Her., 2009), (Gusfield, RECOMB, 2002), and several others.

Classification of Populations



Pedigrees



Parental relationships

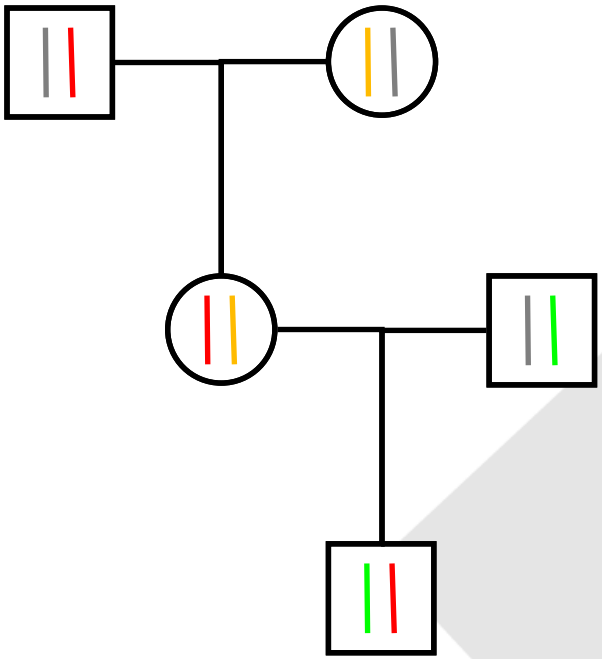
$\Longleftrightarrow$

Mendelian laws of inheritance

$\Longleftrightarrow$

Easier/More accurate HI

Pedigrees



Genotyped Pedigree:
pedigree + genotypes

Haplotype Configuration:
assignment of haplotypes consistent with genotypes

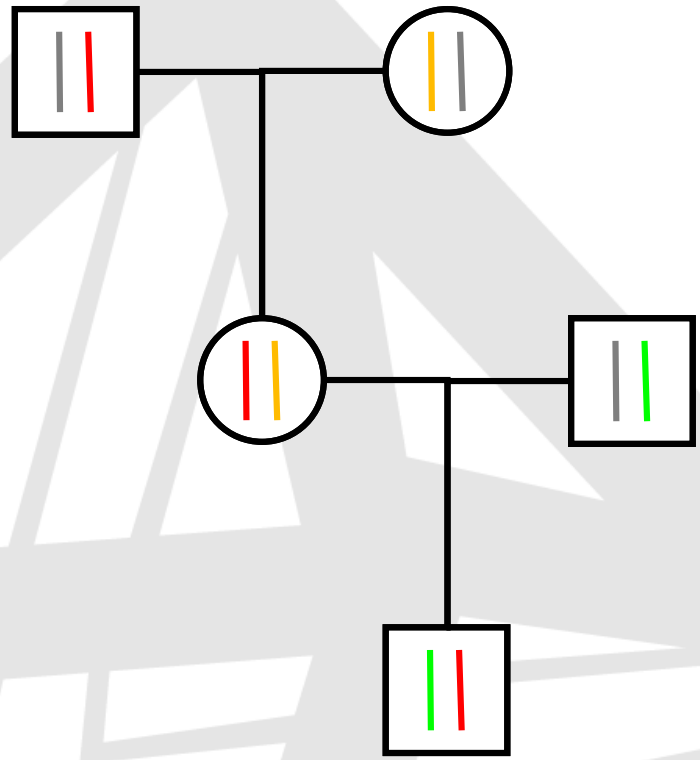
Zero-Recombinant Haplotype Configuration (ZRHC)

Main assumption:
Haplotypes are inherited **without** variations.

Computational problem:

ZRHC

Compute a haplotype configuration without recombinations



(Valid only for short genomic regions and medium-size pedigrees.)

Zero-Recombinant Haplotype Configuration (ZRHC)

Polynomial-time algorithms:

- General pedigrees: $O(mn^2 + n^3 \log^2 n \log\log n)$
  (Xiao *et al.*, SODA, '07)
- Tree pedigrees: $O(nm)$
  (Chan *et al.*, SIAM JComp, '09), (Liu and Jiang, JoCO, '10)

On General Pedigrees: (Xiao *et al.*, SODA, '07)

ZRHC $\Leftrightarrow$ (a particular) Linear System over Z_2

more on that later on...

Minimum Recombinant Haplotype Configuration (MRHC)

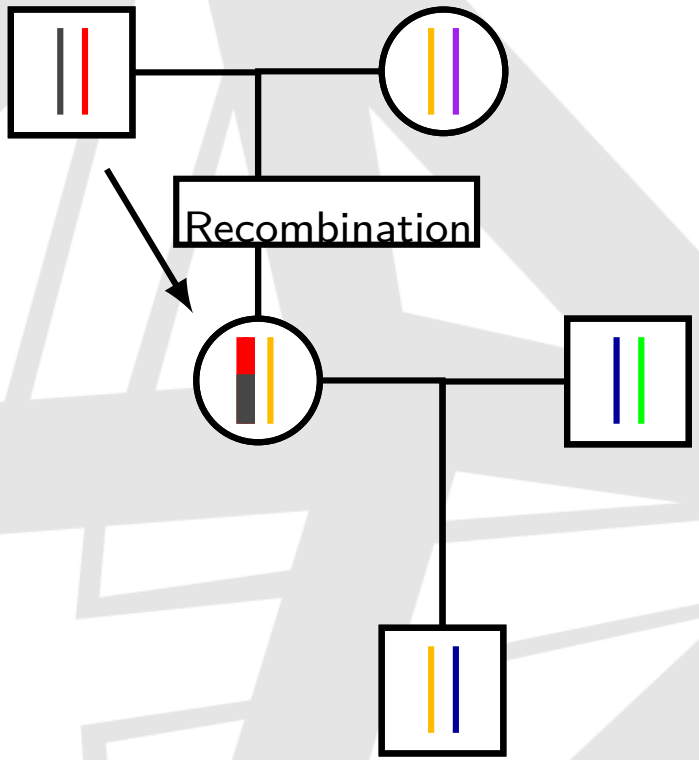
Recombinations naturally occur!

Main assumption:
the most likely solution is the one with the **minimum number of recombinations**

Computational problem:

MRHC

Compute the haplotype configuration with the minimum number of recombinations



Minimum Recombinant Haplotype Configuration (MRHC)

Computational Complexity:

MRHC $\in$ **NP_hard** even on simple pedigrees or "short" genotypes (Liu *et al.*, TCS, 2007)

Algorithms:

- ILP formulation (*PedPhase* by Li and Jiang, JCB, '05)
- Probabilistic algorithm (Xiao *et al.*, ESA, '09)
- Dynamic programming (Doi *et al.*, WABI, '03)
- ...

Minimum Mutation Haplotype Configuration (MMHC)

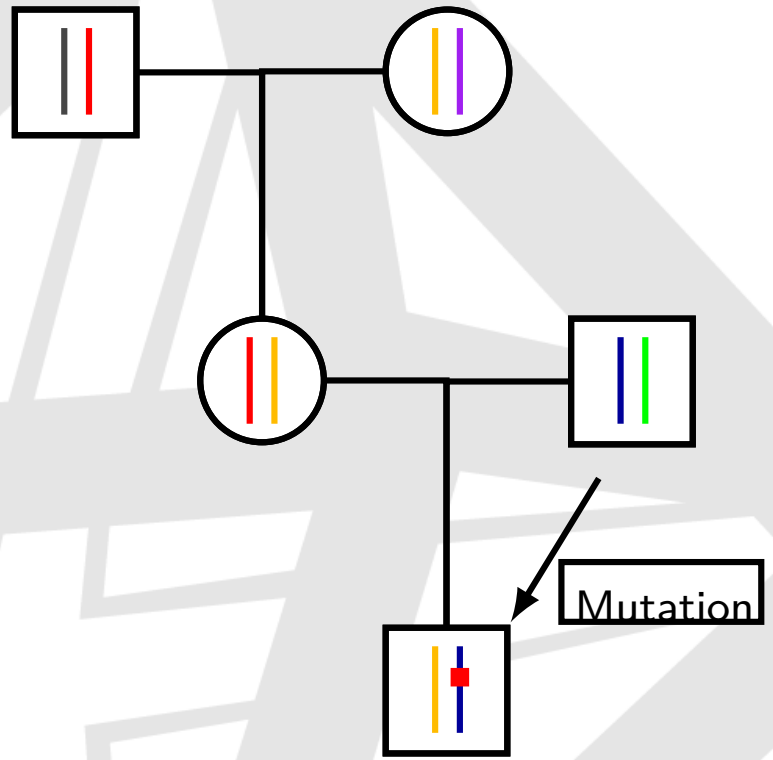
Also **mutations** naturally occur!

Main assumption:
the most likely solution is the one with the **minimum number of mutations**

Computational problem:

MMHC

Compute the haplotype configuration with the minimum number of mutations



Minimum Mutation Haplotype Configuration (MMHC)

Computational Complexity:

MMHC $\in$ NP_hard (Wang and Jiang, CPM, 2009)

Algorithm: (*MMPhase* by Wang and Jiang, CPM, 2009)

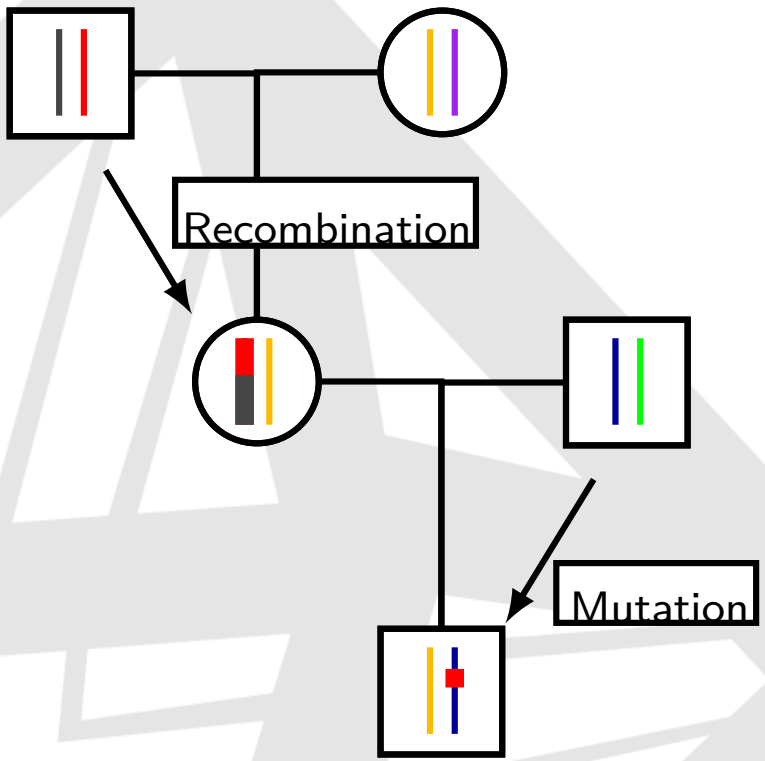
- "Incremental" ILP formulation
  - Worst-case: exponential-size formulation
  - Fast in practice
- Missing genotype imputation
- Infinite-site assumption (only one mutation per locus)

Why Recombinations or Mutations alone?

Recombinations and **mutations** may occur at the same time!

Main assumption:
the most likely solution is the one with the **minimum number of recombinations and mutations**

Our *new* computational problem:
MCHC



$$\text{MRHC} + \text{MMHC} \Rightarrow \textbf{MCHC}$$

Minimum Change Haplotype Configuration (MCHC)

Minimum Change Haplotype Configuration (MCHC) problem

Given a genotyped pedigree, to compute a haplotype configuration that induces the minimum number of recombinations and mutations.

Computational Complexity: MCHC $\in$ NP_hard

proof based on (Liu et al., TCS, '07), omitted due to lack of space

Extending the Linear System of ZRHC

A step back...

$$\text{ZRHC} \Leftrightarrow \text{A Linear System over } Z_2$$

Our Idea: Extending the linear system to accommodate recombinations and mutations

Aim: Describing *each* haplotype configuration of the genotyped pedigree (with recombinations and mutations) as a *particular* solution of a *new* linear system

Extending the Linear System of ZRHC

Original equations: for each locus l, individual i, and p parent of i

$$h_p[l] + s_{p,i} \cdot w_p[l] = h_i[l] + d_{p,i}[l]$$

New equations: for each locus l, individual i, and p parent of i

$$h_p[l] + \left(s_{p,i} + \sum_{j=1}^{l} \boldsymbol{\delta}_{p,i}[j] \right) \cdot w_p[l] = h_i[l] + d_{p,i}[l] + \boldsymbol{\mu}_{p,i}[l]$$

where:

$$\delta_{p,i}[j] = \mathbf{1} \quad \Leftrightarrow \quad \text{a recombination has occurred}$$
$$\mu_{p,i}[l] = \mathbf{1} \quad \Leftrightarrow \quad \text{a mutation has occurred}$$

Reducing MCHC to NCP

New Linear System: *(in matricial form)*

$$A_{h,s} \cdot x_{h,s} + \mathbf{A}_{\delta,\mu} \cdot \mathbf{x}_{\delta,\mu} = b$$

MCHC $\Leftrightarrow$ finding the solution with the minimum number of δ- and μ-variables equal to 1

L-reduces to

Nearest Codeword Problem (NCP)

see e.g. MS3 in (Ausiello *et al.*, 1999)

Nearest Codeword Problem (NCP)

Nearest Codeword Problem (NCP):

- Basic problem in coding theory
- **Theory:** inapproximable within $O(2^{\log^{0.5-\varepsilon} n})$

  (Arora *et al.*, JCSS, 1997)
- **Practice:** solved *extremely well* by *Belief Propagation* (or *Sum Product*) algorithm

  (Gallager, 1963), (Pearl, 1982)

The Heuristic Algorithm

Basic Idea:
first locate recombinations and mutations, then reconstruct haplotypes

Outline:

1. Compute the instance of NCP associated to the instance of MCHC
2. Compute a (minimum?) set of recombinations and mutations with the BP algorithm (on the NCP instance)
3. Compute a corresponding haplotype configuration

The Heuristic Algorithm

Running Time: $O(k \cdot n^3 m^3)$

k no. of events, n pedigree size, m genotype length

Remarks:

- Can also be used for *MRHC* and *MMHC*
- Can include prior knowledge. *E.g.*:
  - recombination hotspots
  - different mutation and recombination rates

Experimental Evaluation

Does it work?

- On **MCHC**: evaluation on *540* random-generated instances
- On **MRHC**: comparison with *PedPhase* and *SimWalk2* on *750* simulated instances
- On **MMHC**: comparison with *MMPhase* on *300* random-generated instances

Test instances: different pedigree "topology", pedigree size, genotype length, recombination and mutation rate.

Experimental Results: MCHC

MCHC	
"Success Rate":	**535/540** ($> 99\%$)
Avg. Phase Error:	**2% – 7%**
Avg. Time:	**3 min.**

Experimental Results: MRHC and MMHC

On **MRHC**:

- faster than PedPhase (*30x*) and SimWalk2 (*>1000x*)
- as accurate as PedPhase and SimWalk2 (avg. phase error: *3% – 4%*)
- optimal solution for *99%* of the instances

On **MMHC**:

- slower than MMPhase (3x) but *no infinite-site assumption!*
- same accuracy (avg. phase error: *3%*)
- optimal solution for *87%* of the instances

Conclusions

Conclusions:

- **MCHC**: new "realistic" formulation of HI
- **Heuristic algorithm**:
  - General
  - Competitive with existing algorithms
  - Can include prior knowledge

Future Work: *(in progress)*

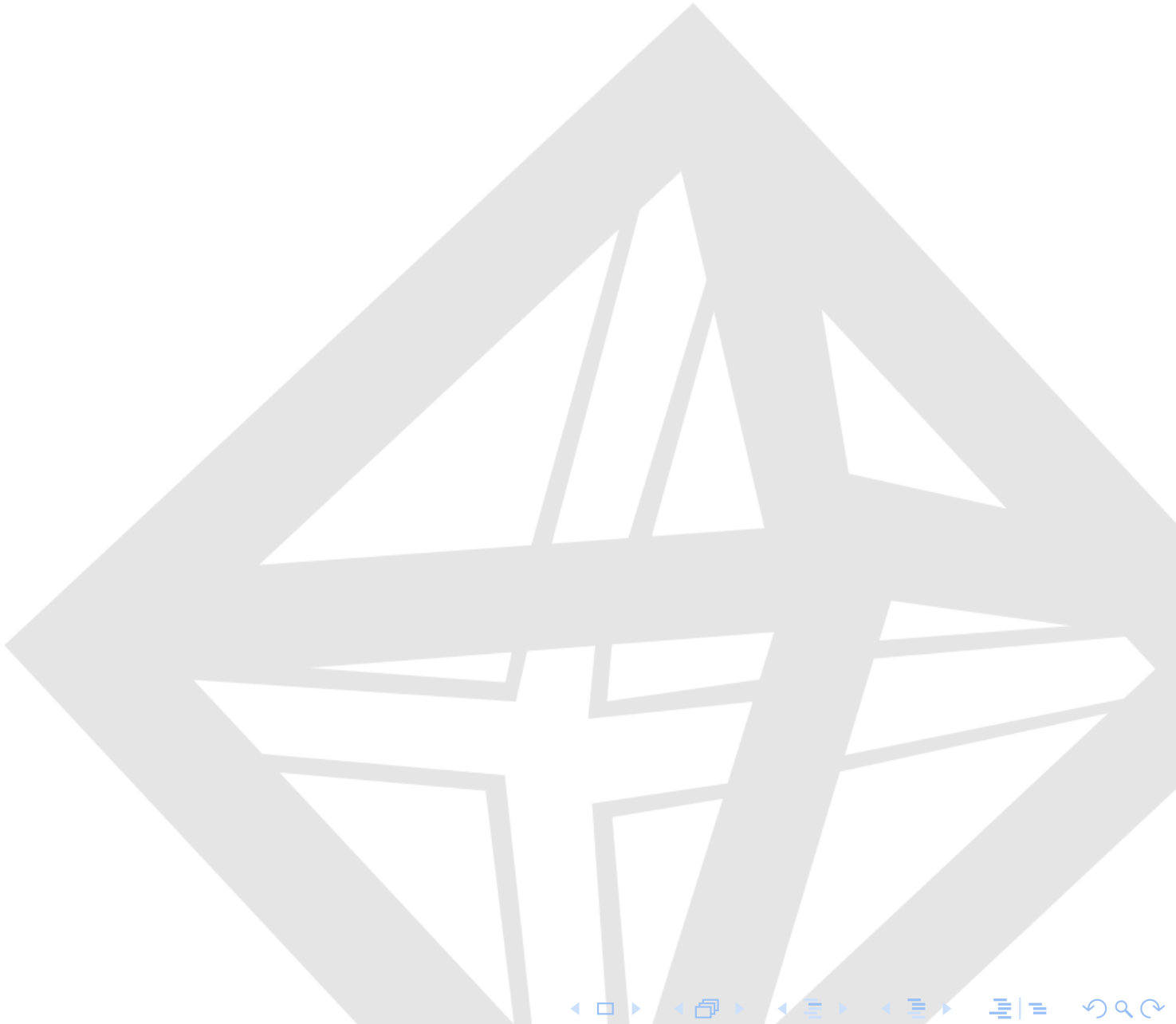
- Missing genotype imputation
- Genotyping error discovery

Haplotype Inference on Pedigrees with Recombinations and Mutations

Yuri Pirola, Paola Bonizzoni, Tao Jiang

pirola@disco.unimib.it

Thank you for your attention!



A Glimpse of Statistical Approaches

Pros:

- Quite accurate
- Numerical assessment of results
- (often) Missing genotype imputation

Cons:

- Computationally intensive (time/accuracy trade-off)

Example: *SimWalk2* (Sobel *et al.*, AJHG, 2002)