Optimal haplotype reconstruction in half-sib families

Aurélie Favier, Simon de Givry Jean-Michel Elsen, Andrés Legarra

Division of Applied Mathematics and Computer Science

Division of Animal Genetics

INRA, Toulouse, France

Workshop on Constraint Based Methods for Bioinformatics

1. Problem : Haplotype reconstruction
2. Suitable probabilistic Model
3. Formalism : Weighted Constraint Satisfaction Problem
4. Experimental results

Genetic notions

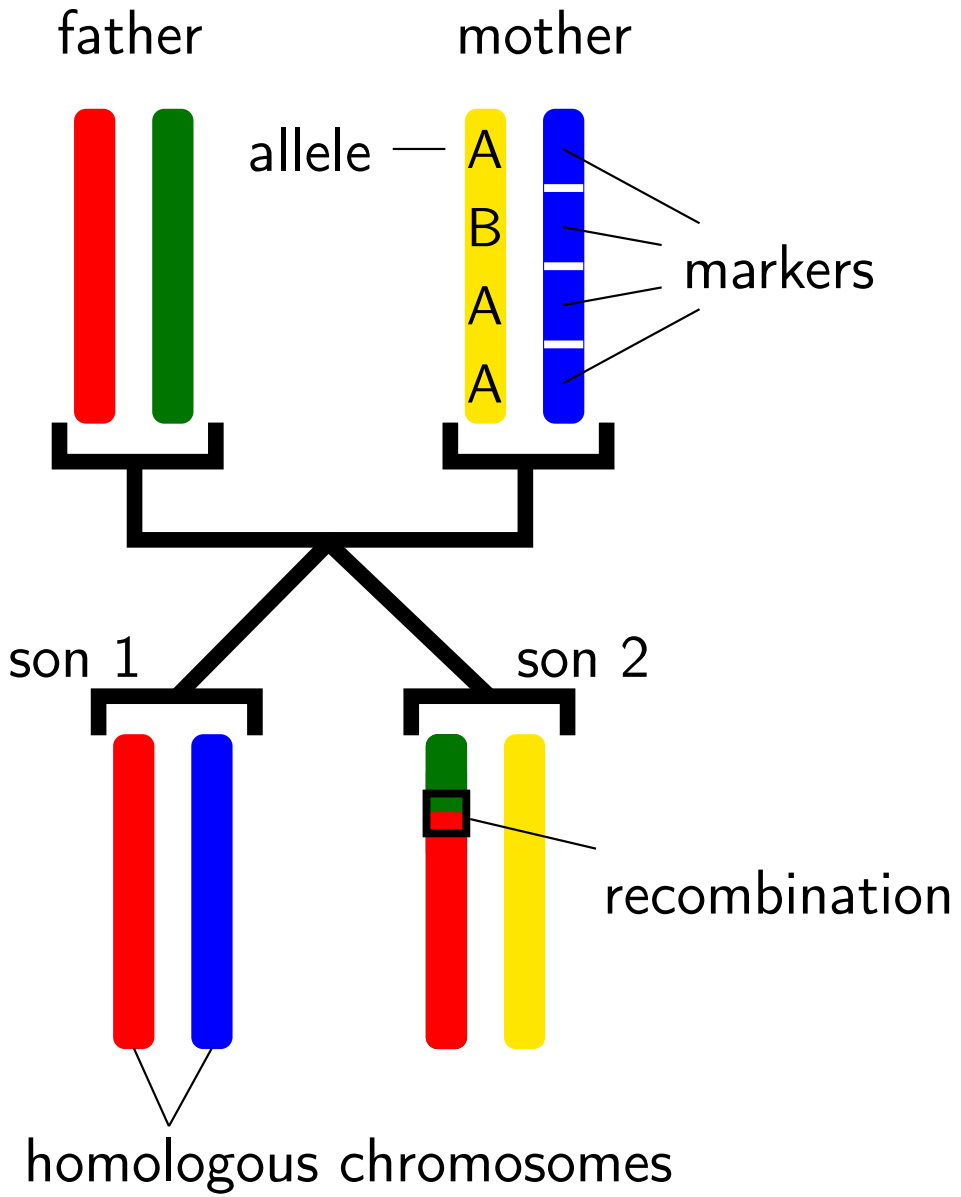
Each chromosome is associated by pair : homologous chromosomes

examples :

- Human : 23 pairs
- Ovine : 27 pairs
- Bovine : 30 pairs

Each pair is composed of :

- 1 chromosome from the father
- 1 chromosome from the mother



Genotype/Haplotype

Haplotype

Sequence of alleles present on a single chromosome

Genotype

Sequence of pairs of unordered alleles from two homologous chromosomes

(SNPs markers have 2 possible values A and B)

hap1	hap2	
A	A	{A,A}
B	A	{A,B}
A	B	{A,B}
A	A	{A,B}

Haplotype reconstruction

Why do we want to know the haplotypes ?

- To describe the genomes [Gibbs *et al.*(2003)]
- To detect and map QTLs (Quantative Trait Locus) and genes [Grisart *et al.*(2004)],[Knott *et al.*(1996)]
- To help the selection in animal population [Calus *et al.*(2008)]

General Haplotype Reconstruction Problem

data : Set of genotypes for each individual

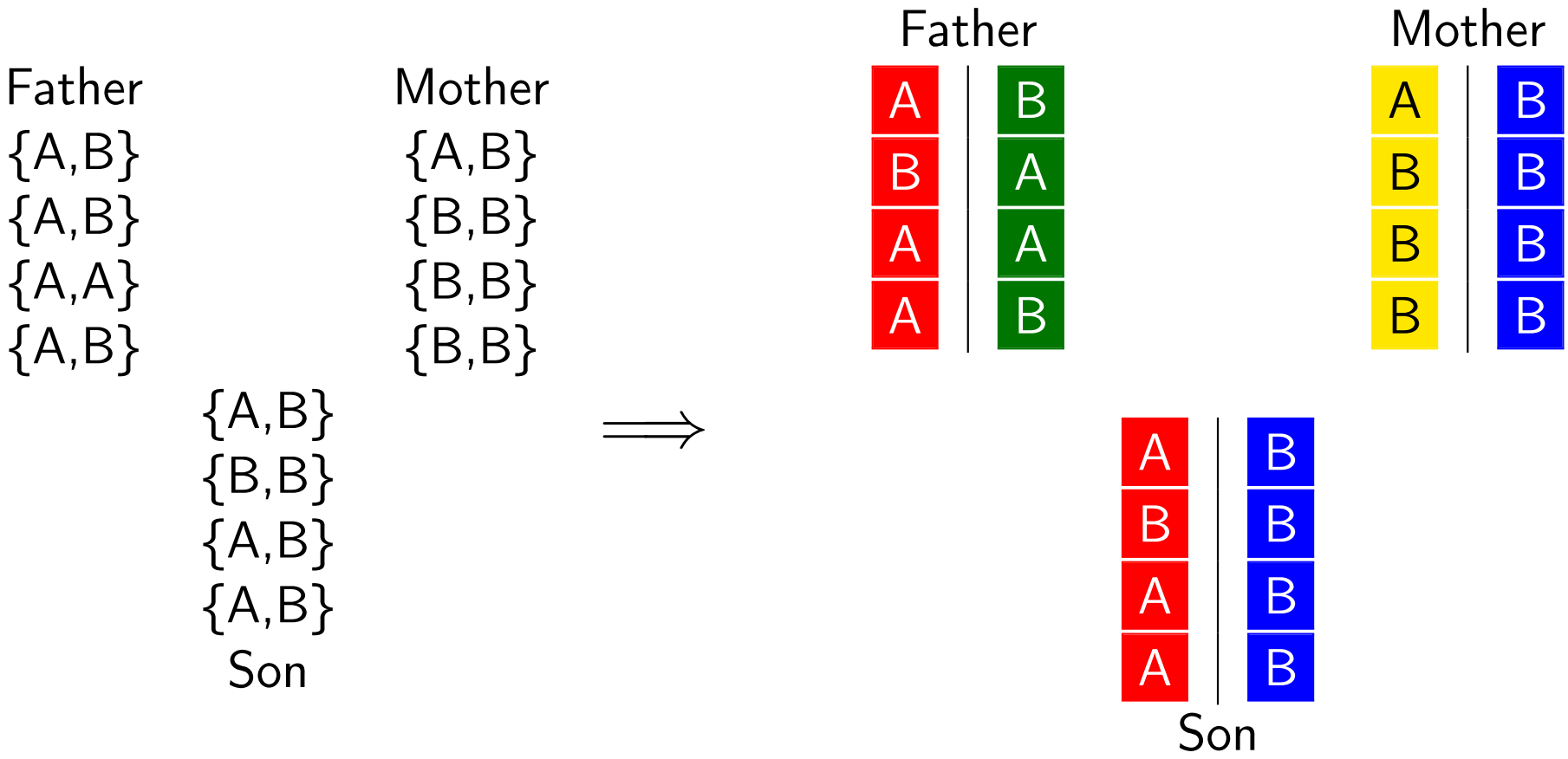
output : Two haplotypes for each individual

Haplotype reconstruction with family information

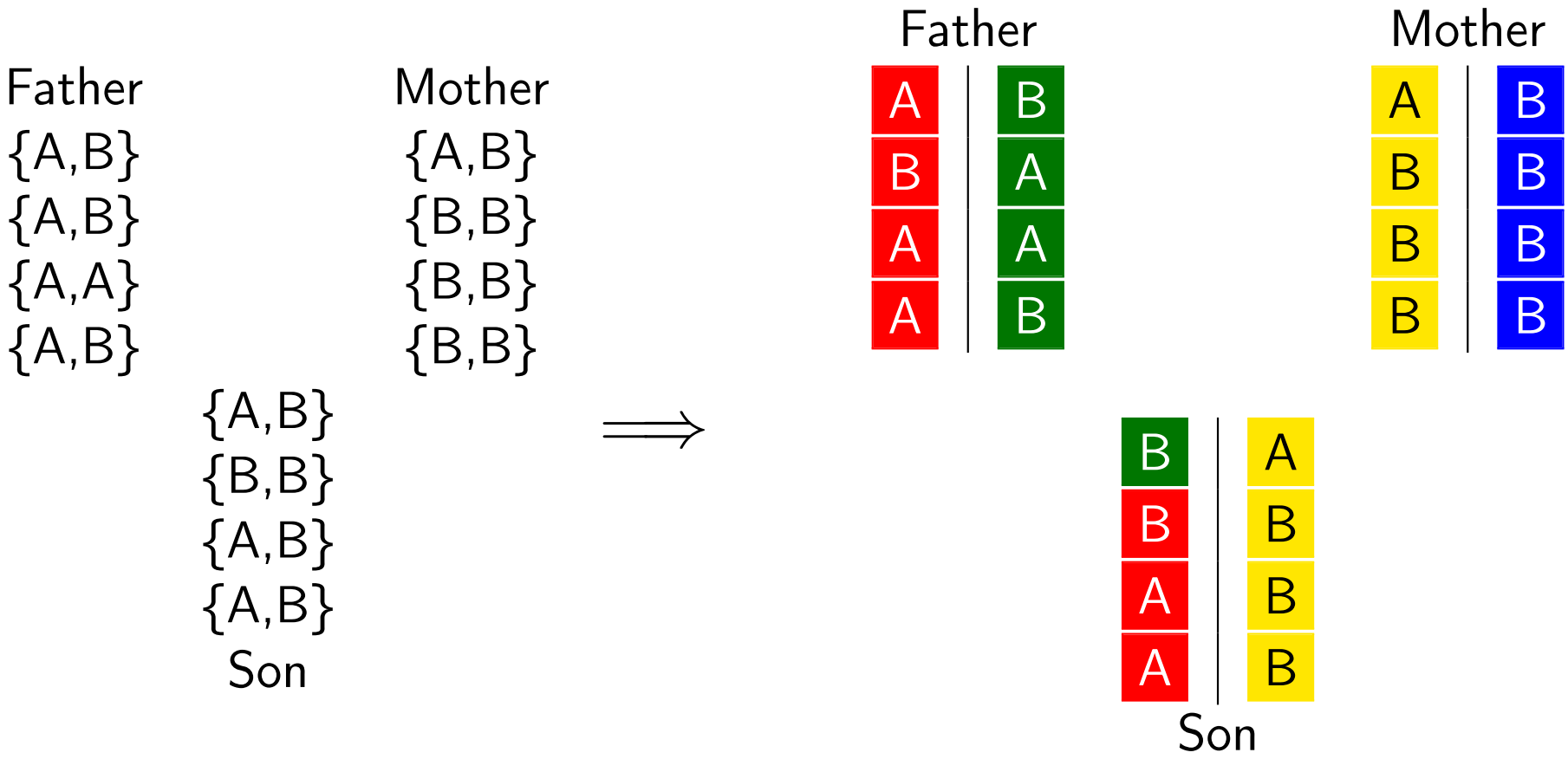
Father	Mother
{A,B}	{A,B}
{A,B}	{B,B}
{A,A}	{B,B}
{A,B}	{B,B}

Son
{A,B}
{B,B}
{A,B}
{A,B}

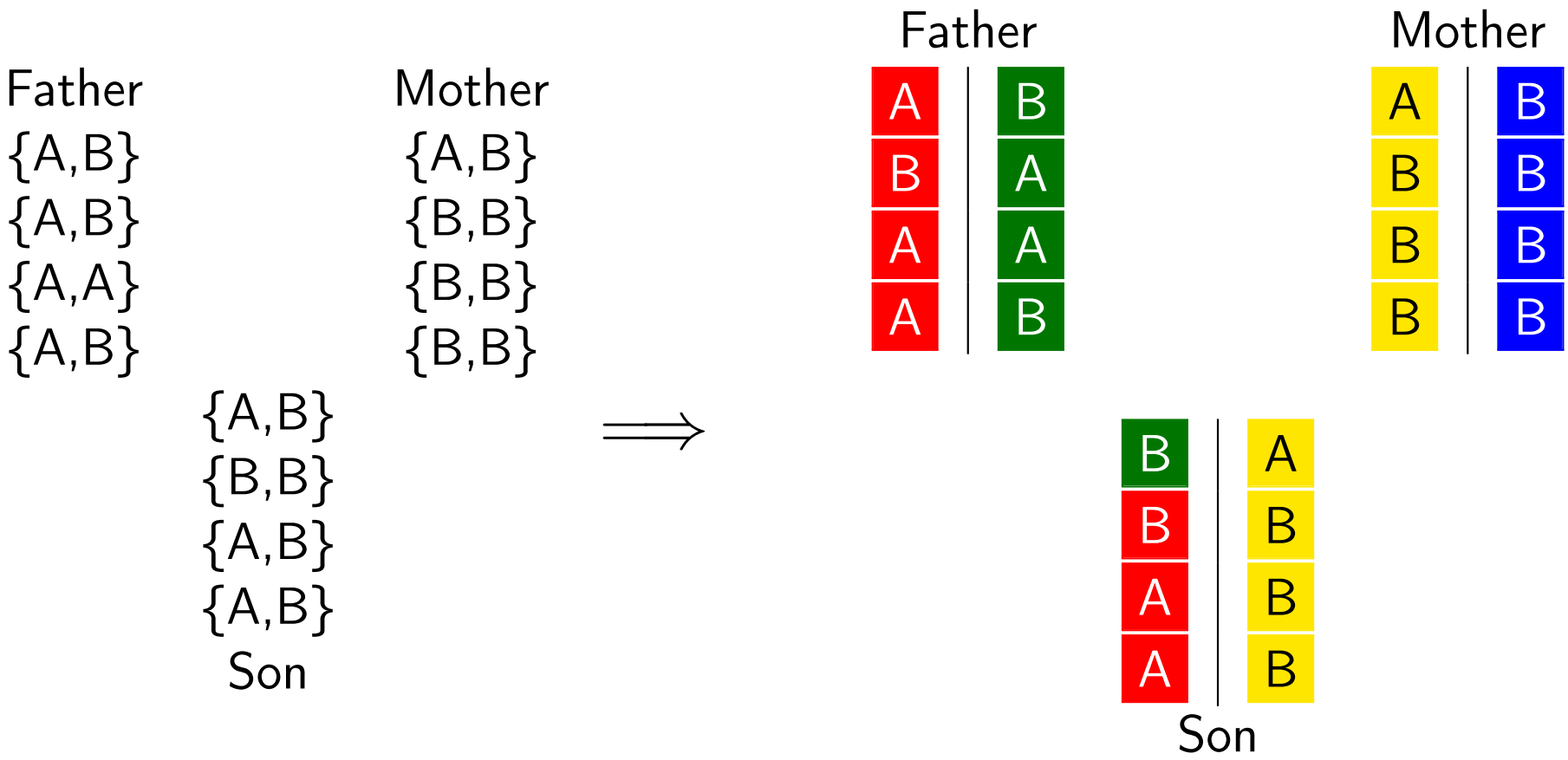
Haplotype reconstruction with family information



Haplotype reconstruction with family information



Haplotype reconstruction with family information



Minimizing the number of recombinations is NP-hard

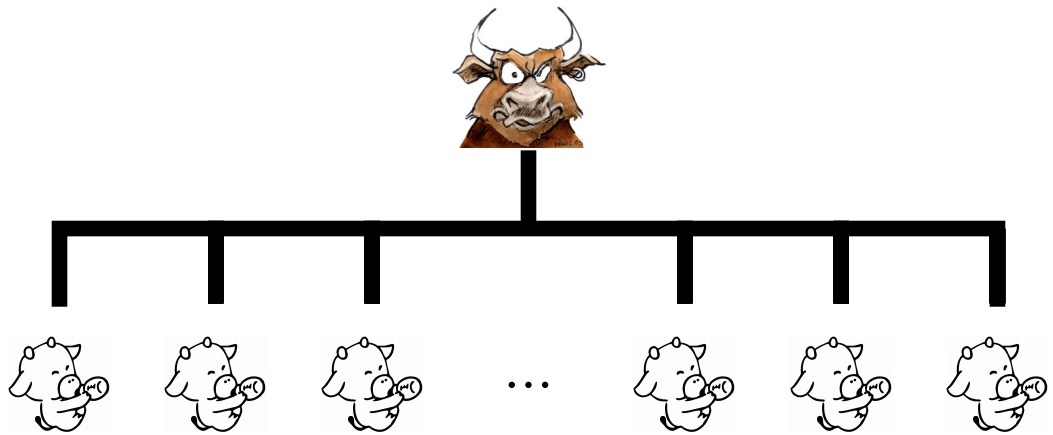
Sire haplotype reconstruction in half-sib families

Why half-sib families ?

`livestock pedigrees are often composed of a few sires and many dams.`

Why only the sire ?

`the selection is mainly on the sires`



Input : sire and offspring genotypes and the genetic map (recombination probabilities r_{ij} between markers i and j)

Minimizing the number of recombinations in half-sib families is NP-hard [Doi *et al.*(2003)]

Decision Variables

Vector of allele order **h**

The indicator of allele order for the sire :

$$h_i = \begin{cases} -1 & \text{if the order of alleles is permuted} \\ 1 & \text{otherwise} \end{cases}$$

genotypes		**h**	haplotypes	
A	**B**	1	**A**	**B**
A	A	1	A	A
A	**B**	-1	**B**	**A**

From genotype to useful information

sire	son i	T^i
AB	BB	2
AB	AA	1
AB	AB	$\star$
BB	AB	$\star$

Transmission vectors

For the marker k :

$$T_k^i = \begin{cases} 1 & \text{if } 1^{st} \text{ allele is transmitted} \\ 2 & \text{if } 2^{nd} \text{ allele is transmitted} \\ \star & \text{otherwise (unknow transmission)} \end{cases}$$

(iff sire is heterozygous and son i is homozygous)

and $\mathbf{T}=(\mathbf{T}^1 \dots \mathbf{T}^n)$.

Likelihood

- The probability of $\mathbf{h}$ is :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h}) \cdot p(\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}') \cdot p(\mathbf{h}')}$$

Likelihood

- The probability of $\mathbf{h}$ is :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h}) \cdot p(\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}') \cdot p(\mathbf{h}')}$$

- No *a priori* on the haplotypes (assume linkage equilibrium) :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}')}$$

Likelihood

- The probability of $\mathbf{h}$ is :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h}) \cdot p(\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}') \cdot p(\mathbf{h}')}$$

- No *a priori* on the haplotypes (assume linkage equilibrium) :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}')}$$

- $p(\mathbf{h}|\mathbf{T}) \propto p(\mathbf{T}|\mathbf{h}) \Longrightarrow$ maximize the likelihood p($\mathbf{T}|\mathbf{h}$)

Likelihood

- The probability of $\mathbf{h}$ is :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h}) \cdot p(\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}') \cdot p(\mathbf{h}')}$$

- No *a priori* on the haplotypes (assume linkage equilibrium) :

$$p(\mathbf{h}|\mathbf{T}) = \frac{p(\mathbf{T}|\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}')}$$

- $p(\mathbf{h}|\mathbf{T}) \propto p(\mathbf{T}|\mathbf{h}) \Longrightarrow$ maximize the likelihood $p(\mathbf{T}|\mathbf{h})$
- Independance of meiosis between each son :

$$p(\mathbf{T}|\mathbf{h}) = \prod_{i=1}^{n} p(\mathbf{T}^i|\mathbf{h})$$

Prbability for a son i

$$p(\mathbf{T}^i|\mathbf{h}) = p(T_1^i|\mathbf{h}).p(T_2^i|\mathbf{h}, T_1^i).p(T_3^i|\mathbf{h}, T_1^i, T_2^i)\ldots p(T_L^i|\mathbf{h}, T_1^i, \ldots, T_{L-1}^i)$$

$$T_m^i = \quad \star : \quad p(T_m^i|\mathbf{h}, T_1^i, \ldots, T_{m-1}^i) = 1$$

sire genotype : son i genotype : $\mathbf{T}^i$ $\mathbf{h}$ sire haplotype :

AB	AA	1	$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$	$\Longrightarrow$	A | B
AB	BB	2			B | A

$\Longrightarrow$ no recombination : $p(T_2^i = 2|\mathbf{h} = (1,\text{-}1), T_1^i = 1) = 1\text{-}r_{1,2}$

Prbability for a son i

$$p(\mathbf{T}^i|\mathbf{h}) = p(T_1^i|\mathbf{h}).p(T_2^i|\mathbf{h}, T_1^i).p(T_3^i|\mathbf{h}, T_1^i, T_2^i)\ldots p(T_L^i|\mathbf{h}, T_1^i, \ldots, T_{L-1}^i)$$

$$T_m^i = \quad \star : \quad p(T_m^i|\mathbf{h}, T_1^i, \ldots, T_{m-1}^i) = 1$$

Recombination event

sire genotype :	son i genotype :	$\mathbf{T}^i$	$\mathbf{h}$		sire haplotype :
AB	**A**A	1	$\begin{bmatrix}1\\1\end{bmatrix}$	$\Longrightarrow$	**A** | **B**
AB	B**B**	2			**A** | **B**

$\Longrightarrow$ recombination : $p(T_2^i = 2|\mathbf{h} = (1,\ 1), T_1^i = 1) = \quad r_{1,2}$

Likelihood

After the log transformation we obtain :

$$\log\left[p(\mathbf{T}|\mathbf{h})\right] = const + \sum_{m=1}^{L}\sum_{k<m} h_m h_k \underbrace{\frac{1}{2}\log\left(\frac{1-r_{km}}{r_{km}}\right)\left(N_{km}^{+} - N_{km}^{-}\right)}_{=W_{km}}$$

N_{km}^{+} = number of descendants i such that $T_m^i = T_k^i \neq \star$ and $\forall k < l < m \; T_l^i = \star$

N_{km}^{-} = number of descendants i such that $T_m^i \neq T_k^i \neq \star$ and $\forall k < l < m \; T_l^i = \star$

N^+ and N^- computation

$$W_{km} = \frac{1}{2} \log \left(\frac{1 - r_{km}}{r_{km}} \right) (N^+_{km} - N^-_{km})$$

sire	son 1	son 2	son 3
AB	**BB**	**AA**	AB
AB	AB	AB	AB
AB	**BB**	**AA**	**AA**
AA	AB	AB	AB
AB	**BB**	**BB**	**BB**

sire	T^1	T^2	T^3
AB	2	1	$\star$
AB	$\star$	$\star$	$\star$
AB	2	1	1
AA	$\star$	$\star$	$\star$
AB	2	2	2

$$N^+_{1,3} = 0 \qquad \mathrm{N}^-_{1,3} = 0$$
$$N^+_{3,5} = 0 \qquad \mathrm{N}^-_{3,5} = 0$$

N^+ and N^- computation

$$W_{km} = \frac{1}{2} \log \left(\frac{1 - r_{km}}{r_{km}} \right) (N^+_{km} - N^-_{km})$$

sire	son 1	son 2	son 3
AB	**BB**	**AA**	AB
AB	AB	AB	AB
AB	**BB**	**AA**	**AA**
AA	AB	AB	AB
AB	**BB**	**BB**	**BB**

sire	T^1	T^2	T^3
AB	**2**	1	$\star$
AB	$\star$	$\star$	$\star$
AB	**2**	1	1
AA	$\star$	$\star$	$\star$
AB	**2**	2	2

$$N^+_{1,3} = \mathbf{1} \qquad N^-_{1,3} = 0$$
$$N^+_{3,5} = \mathbf{1} \qquad N^-_{3,5} = 0$$

N^+ and N^- computation

$$W_{km} = \frac{1}{2} \log \left(\frac{1 - r_{km}}{r_{km}} \right) (N^+_{km} - N^-_{km})$$

sire	son 1	son 2	son 3
AB	**BB**	**AA**	AB
AB	AB	AB	AB
AB	**BB**	**AA**	**AA**
AA	AB	AB	AB
AB	**BB**	**BB**	**BB**

sire	T^1	T^2	T^3
AB	2	**1**	$\star$
AB	$\star$	$\star$	$\star$
AB	2	**1**	1
AA	$\star$	$\star$	$\star$
AB	2	**2**	2

$$N^+_{1,3} = \mathbf{2} \qquad N^-_{1,3} = 0$$
$$N^+_{3,5} = 1 \qquad N^-_{3,5} = \mathbf{1}$$

N^+ and N^- computation

$$W_{km} = \frac{1}{2} \log \left(\frac{1 - r_{km}}{r_{km}} \right) (N^+_{km} - N^-_{km})$$

sire	son 1	son 2	son 3
AB	**BB**	**AA**	AB
AB	AB	AB	AB
AB	**BB**	**AA**	**AA**
AA	AB	AB	AB
AB	**BB**	**BB**	**BB**

sire	T^1	T^2	T^3
AB	2	1	$\star$
AB	$\star$	$\star$	$\star$
AB	2	1	**1**
AA	$\star$	$\star$	$\star$
AB	2	2	**2**

$$N^+_{1,3} = 2 \qquad N^-_{1,3} = 0$$
$$N^+_{3,5} = 1 \qquad N^-_{3,5} = \mathbf{2}$$

Weighted Constraint Satisfaction Problem (1/3)

binary Weighted Constraint Satisfaction Problem

A binary WCSP is a triplet (X, D, F)

- X the set of m variables
- D the set of finite domains
- F the set of binary cost functions

A binary cost function between X_i and X_j : $f_{ij} \in F : D_i \times D_j \rightarrow \mathbb{N}$

Minimizing the total cost function :

$$\sum_{f_{ij} \in F} f_{ij}(t[i], t[j])$$

with $t[i]$ value assigned to variable i in the assignment t of all variables

Haplotype Reconstruction in WCSP formulation

- $X = \{h_1, \ldots, h_L\}$ (L number of markers)
- $D_i = \{-1, 1\}$ for $i \in \{1, \ldots, L\}$

We want to maximize $\sum_{l=1}^{L} \sum_{k<l} h_k h_l W_{kl}$

- $F = \{f_{kl} | W_{kl} \neq 0, k < l\}$

$$f_{kl} = -\, h_k h_l W_{kl} + |W_{kl}|$$

minimization

cost function must be positive

If $W_{kl} > 0$: f_{kl} is **soft equality function**

If $W_{kl} < 0$: f_{kl} is **soft disequality function**

Dataset and methods

Simulation :

- 1500 markers on 1 Morgan
- no linkage desequilibrium
- allele frequencies $\beta(x, y)$ as found in beef cattle
- 1 to 100 descendants
- 50 families were simulated for each set of parameters

Exact Methods :

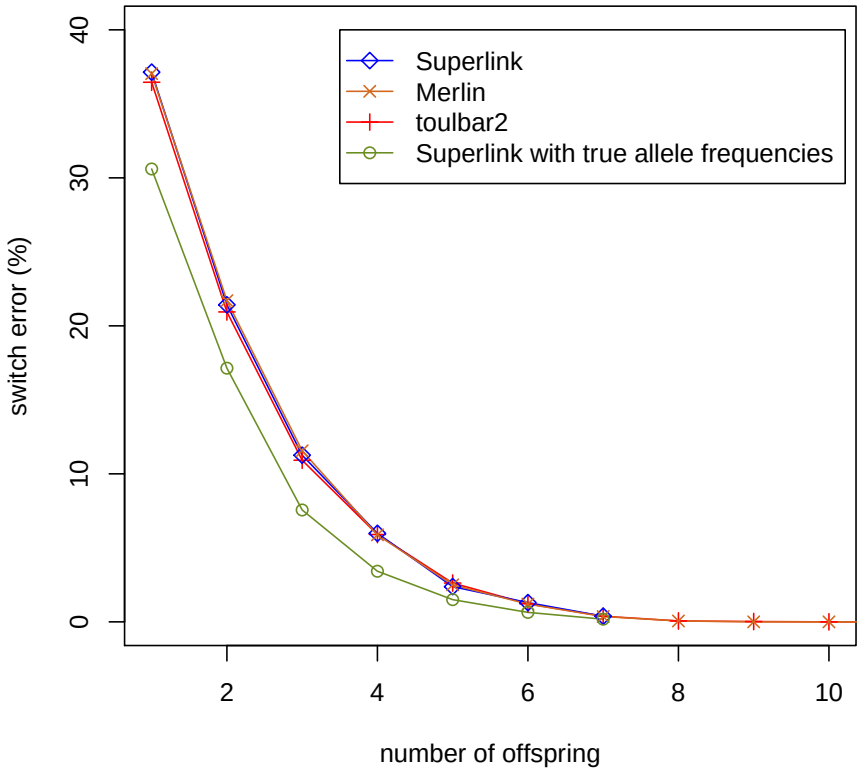
- toulbar2 [1] (using our WCSP formulation)
- Merlin [Abecasis *et al.*(2002)]
- Superlink [Fichelson *et al.*(2005)]

Approximated Methods :

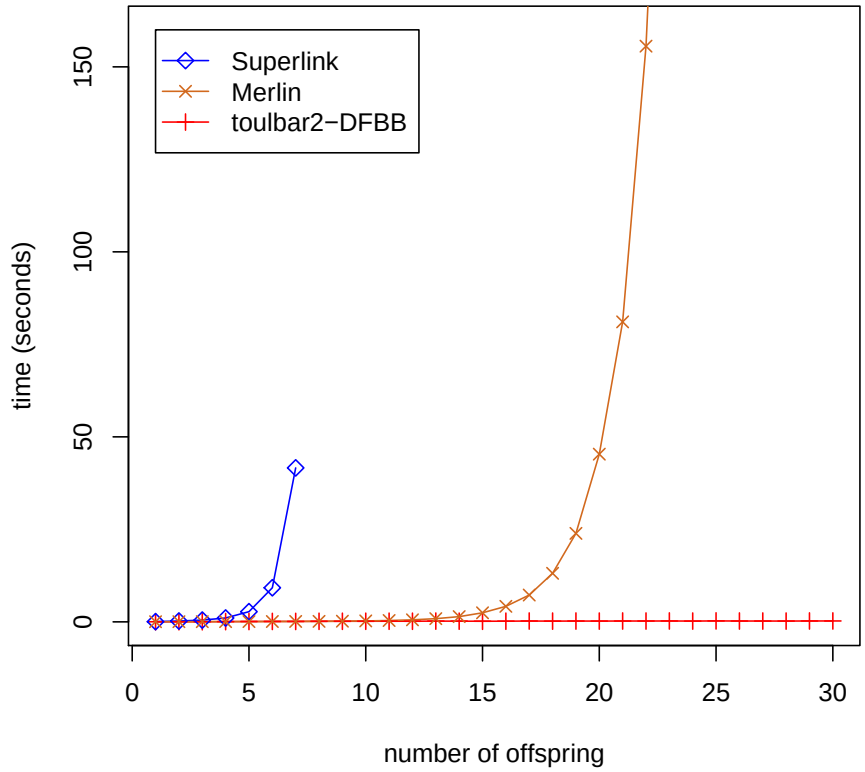
- W&M algorithm [Windig & Meuwissen(2004)]
- LinkPhase [Druet & Georges(2010)]

1. `carlit.toulouse.inra.fr/cgi-bin/awki.cgi/ToolBarIntro`

Experimental Results : Exact methods



(a) switch errors

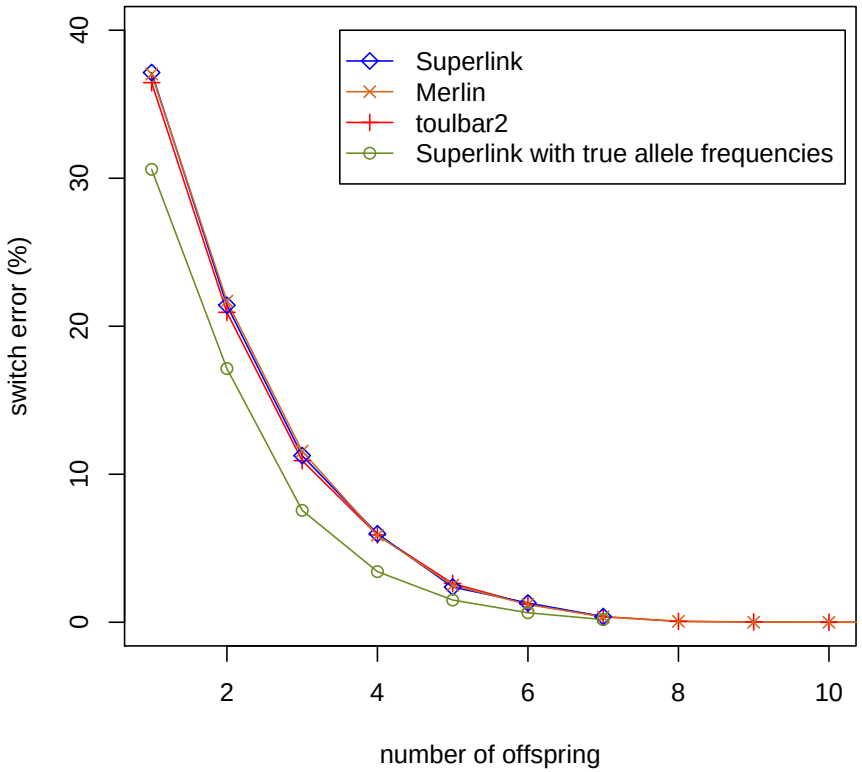


(b) time

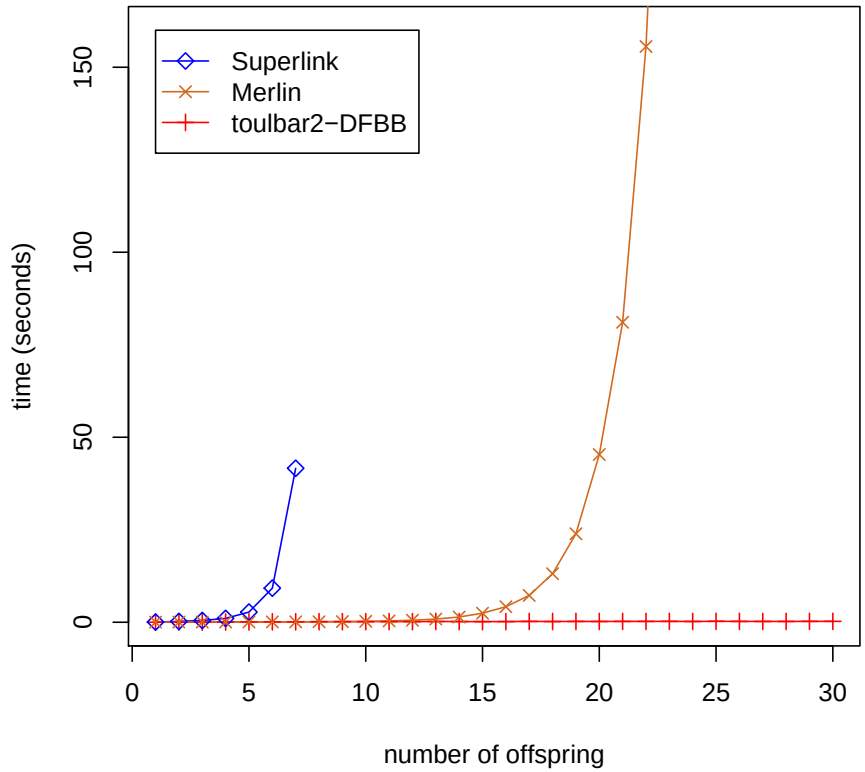
switch error

$$\frac{\begin{matrix}A & A & A & A\end{matrix}}{\begin{matrix}B & B & B & B\end{matrix}} \qquad \frac{\begin{matrix}A & A & B & B\end{matrix}}{\begin{matrix}B & B & A & A\end{matrix}} \quad \Longrightarrow \quad \text{one switch error}$$

Experimental Results : Exact methods



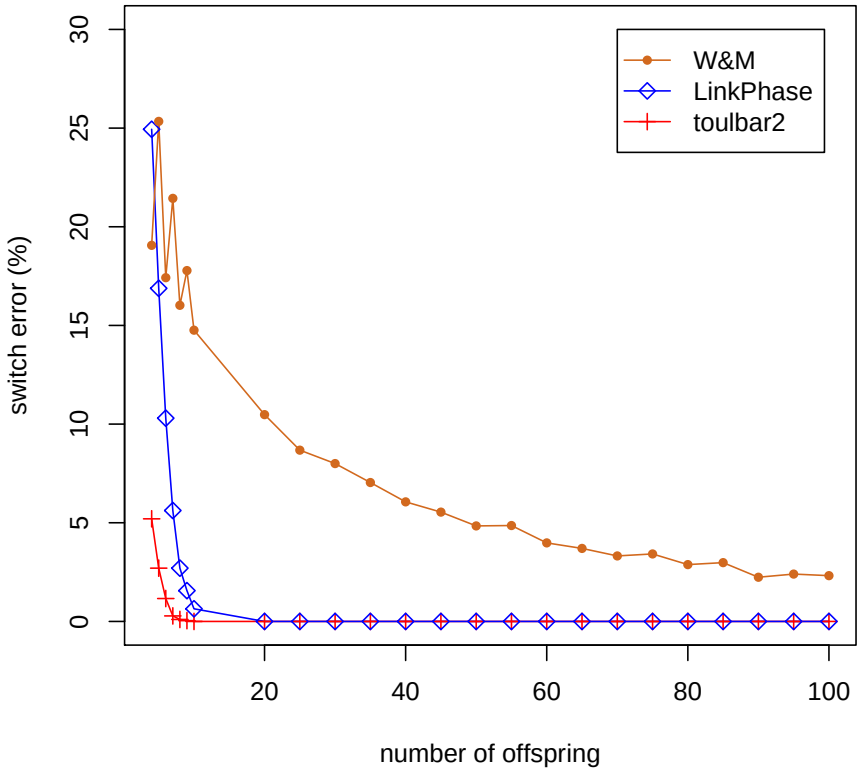
(c) switch errors



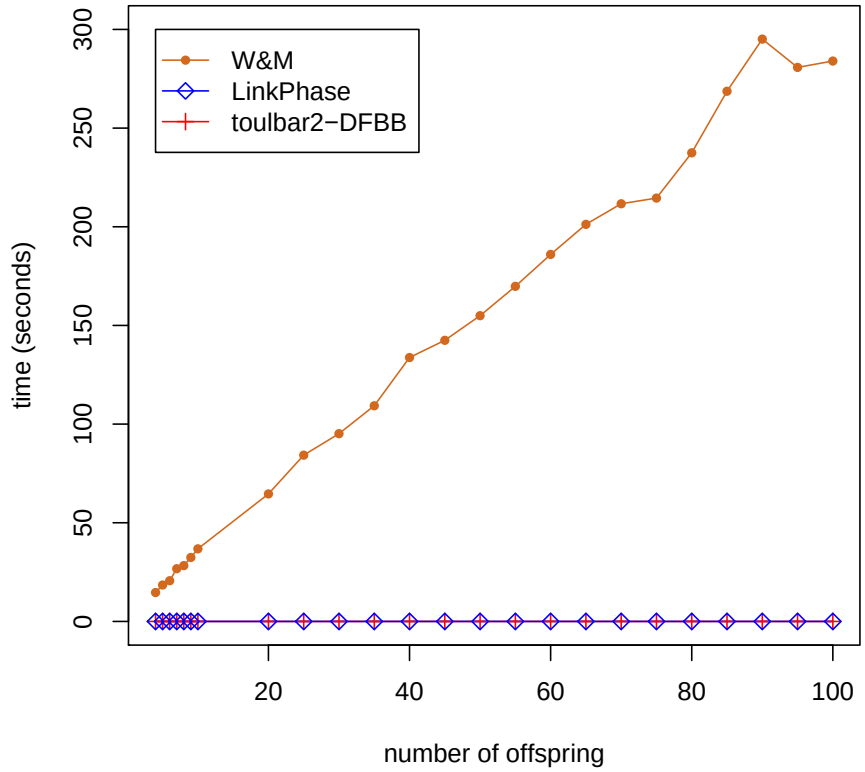
(d) time

- toulbar2 is efficient even with many offspring

Experimental Results : Approximated methods



(a) switch errors



(b) time

- toulbar2 is always better in switch errors
- toulbar2 is as efficient as LinkPhase

Real dataset

Human chromosome X (HapMap)

- 36 000 markers on 1.64 Morgan
- with linkage desequilibrium

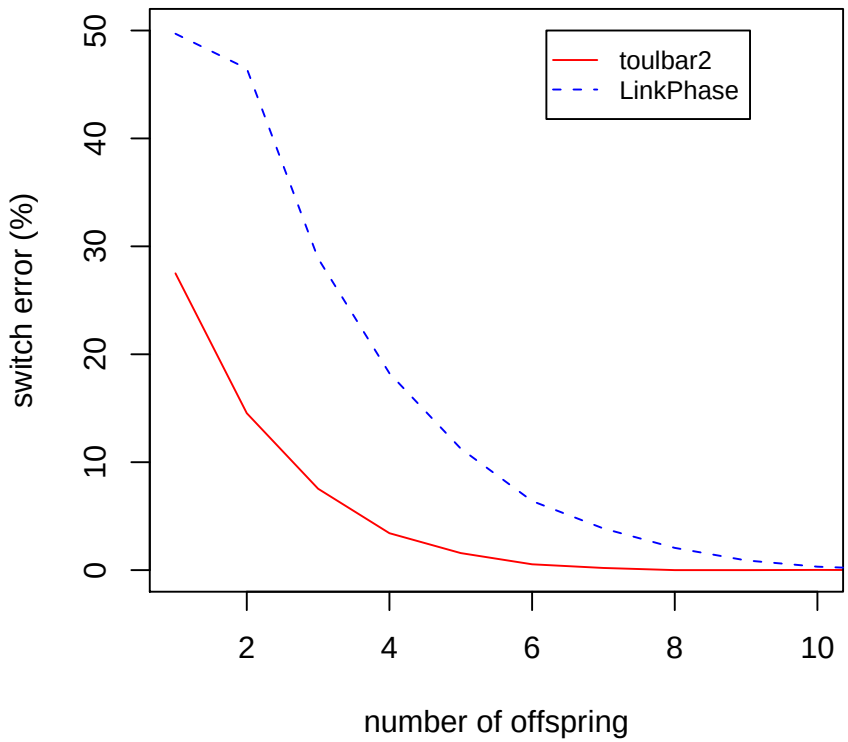
Simulation :

- 1 to 15 descendants
- 50 families were simulated for each set of parameters

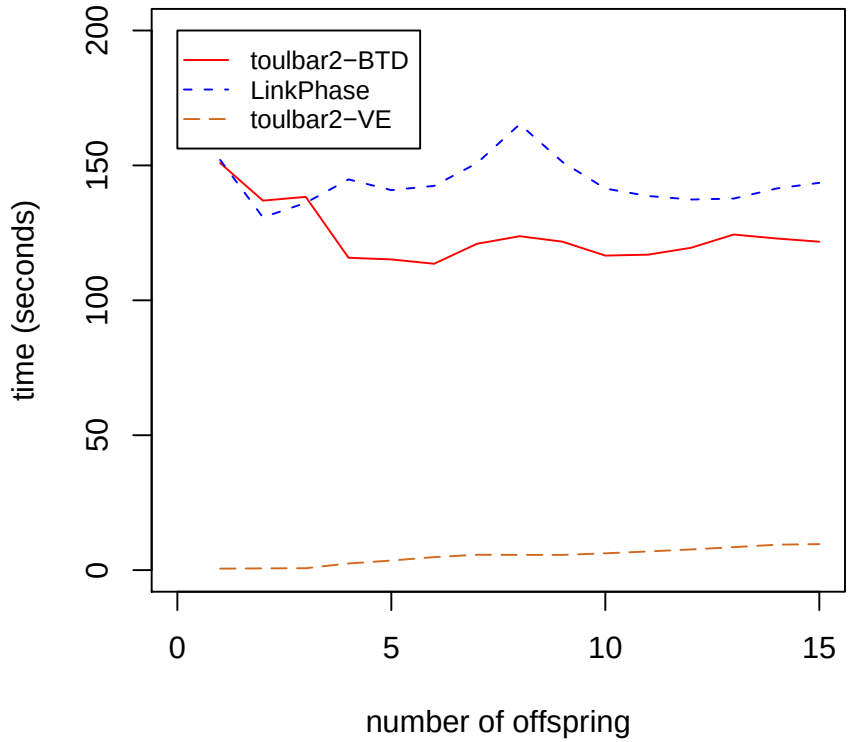
Methods :

- toulbar2
- Linkphase

Experimental Results : Human chromosome X



(a) switch errors



(b) time

BTD : depth first branch and bound exploiting the problem structure
VE (Variable Elimination) : dynamic programming

Conclusion and Perspectives

Conclusion

- A WCSP formulation was proposed for sire haplotype reconstruction in half-sib families
- Our method obtain good results, in terms of accuracy and time, on simulated and real datasets.

Perspectives

- Integration the dam informations (see the paper)
- Integration of Linkage Desequilibrium
- Study other kinds of pedigrees


Abecasis, G., Cherny, S., Cookson, W., & Cardon, L. (2002).
Merlin – rapid analysis of dense genetic maps using sparse gene flow trees.
Nature Genetics, *30*, 97–101.


Druet, T. & Georges, M. (2010).
A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping.
Genetics, *184*.


Fichelson, M., Dovgolevsky, N., & Geiger, D. (2005).
Maximum Likelihood Haplotyping for General Pedigrees.
Human Heredity, *59*(1), 41–60.


Windig, J. & Meuwissen, T. (2004).
Rapid haplotype reconstruction in pedigrees with dense marker maps.
J. of Animal Breeding and Genetics, *121*, 26–39.


Gibbs, R., Belmont, J., Hardenbol, P., Willis, T., Yu, F., Yang, H., Ch'ang, L., Huang, W., Liu, B., Shen, Y., *et al.* (2003).

The international HapMap project.
Nature, *426*(6968), 789–796.


Knott, S., Elsen, J., & Haley, C. (1996).
Methods for multiple-marker mapping of quantitative trait loci in half-sib populations.
TAG Theoretical and Applied Genetics, *93*(1), 71–80.


Calus, M. *et al.* (2008).
Accuracy of genomic selection using different methods to define haplotypes.
Genetics, *178*(1), 553.


Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J., Kvasz, A., Mni, M., Simon, P., Frère, J., Coppieters, W., *et al.* (2004).
Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition.
Proceedings of the National Academy of Sciences, *101*(8), 2398.


Doi, K., Li,J. , Jiang, T.(2003).

Minimum Recombinant Haplotype Configuration on Tree Pedigrees

Proceedings of WABI'03