Réseau bayésien et génétique

Simon de Givry
INRA-UBIA Toulouse
21 juin 2011

PLAN

Réseau Bayésien

- définition
- réseaux de génotypes, d'allèles, de ségrégations
- séparation dirigée et indépendance conditionnelle
- principales requêtes
- méthode exacte pour l'inférence probabiliste
- méthode exacte pour l'optimisation combinatoire
- quelques résultats en reconstruction d'haplotypes à partir d'un pedigree et de génotypes

Supports de cours

Transparents des cours, articles et tutoriaux disponibles à
`http://mulcyber.toulouse.inra.fr/scm/viewvc.php/MAB/?root=mposc`

- Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret et Anna Becker, *Réseaux bayésiens*, 2008
  BNT Matlab (Kevin Murphy) `http://bnt.insa-rouen.fr/`
- Adnan Darwiche, *Modeling and Reasoning with Bayesian Networks*, 2009
  Samlam `http://reasoning.cs.ucla.edu/samiam`
- Christopher M. Bishop, *Pattern Recognition and Machine Learning* (chapitre 8), 2006
  `http://research.microsoft.com/~cmbishop/PRML/index.htm`
- Finn V. Jensen, *Bayesian Networks and Decision Graphs*, 2001
- Association for Uncertainty in Artificial Intelligence (AUAI)
  `http://auai.org/`
- Logiciels payants : (usa) Hugin, Netica, (fr) Bayesia, ProBayes,...

Autres liens sur internet

- Rina Dechter, *courses on Belief Networks (slides and homework)*, 2009 (with focus on Genetic Linkage Analysis, 2005) `http://www.ics.uci.edu/~dechter`
- Dan Geiger, *Advanced Topics in Bioinformatics* (genetics), 2006 `http://www.cs.technion.ac.il/~anna_bi/cs236633/`
- Kevin Murphy, *A Brief Introduction to Graphical Models and Bayesian Networks*, 1998
  `http://people.cs.ubc.ca/~murphyk/Bayes/bayes.html`
- Bruno Garcia, *notes de cours sur Recherche Opérationnelle*, 2000
  `http://www.bruno-garcia.net/www/polyro/polyro.html`
- Daphne Koller, Nir Friedman, *Probabilistic Graphical Models*, 2009

Définition d'un réseau bayésien

Définition

- Un réseau bayésien est défini par
  - la description qualitative des dépendances (ou des indépendances conditionnelles) entre des variables S_i

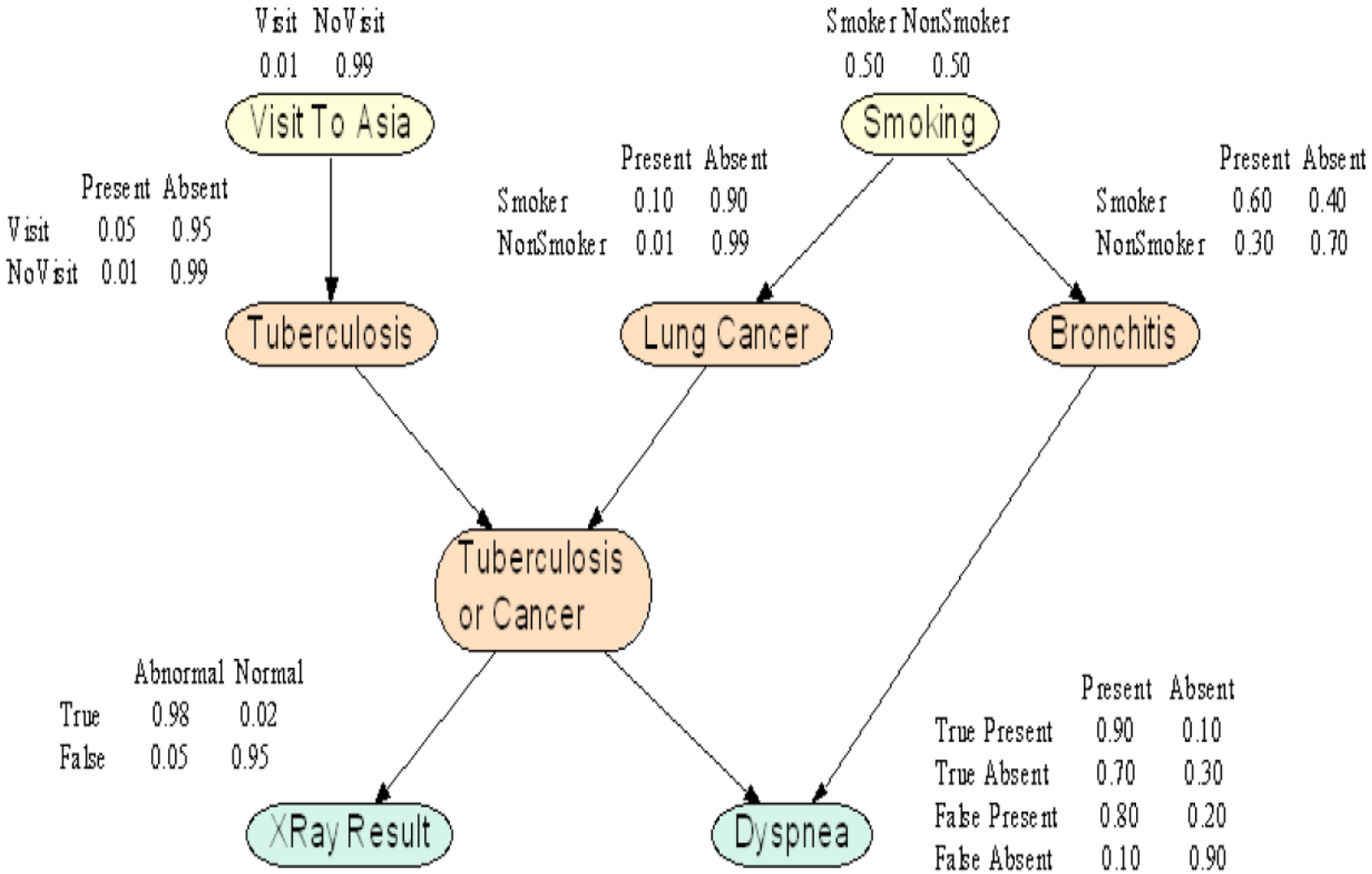
    graphe orienté sans circuit (DAG)
  - la description quantitative de ces dépendances

    probabilités conditionnelles (CPD)

Conséquence

- $P(S) = \Pi_{i=1}^{n} P(S_i | parents(S_i))$
- La loi jointe (globale) se décompose en un produit de lois conditionnelles locales
- RB = représentation compacte de la loi jointe $P(S)$

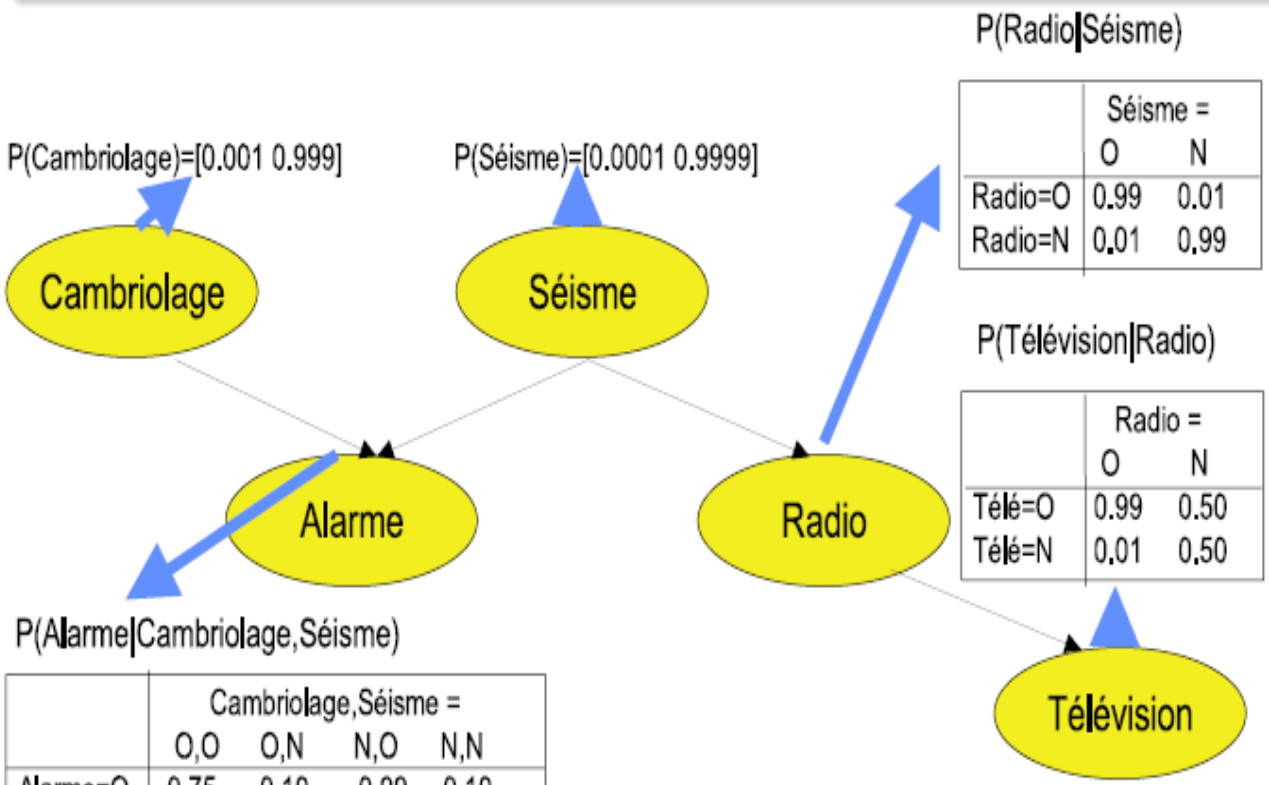
Exemple


Visit NoVisit
0.01 0.99
Visit To Asia
Present Absent
Visit 0.05 0.95
NoVisit 0.01 0.99
Tuberculosis
Smoker NonSmoker
0.50 0.50
Smoking
Present Absent
Smoker 0.10 0.90
NonSmoker 0.01 0.99
Lung Cancer
Present Absent
Smoker 0.60 0.40
NonSmoker 0.30 0.70
Bronchitis
Tuberculosis or Cancer
Abnormal Normal
True 0.98 0.02
False 0.05 0.95
XRay Result
Present Absent
True Present 0.90 0.10
True Absent 0.70 0.30
False Present 0.80 0.20
False Absent 0.10 0.90
Dyspnea

Dimension d'un réseau bayésien

Définition

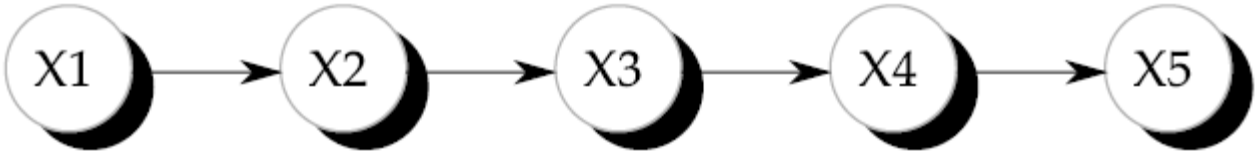
Nombre de paramètres (indépendants) nécessaires pour décrire l'ensemble des CPD associées au RB



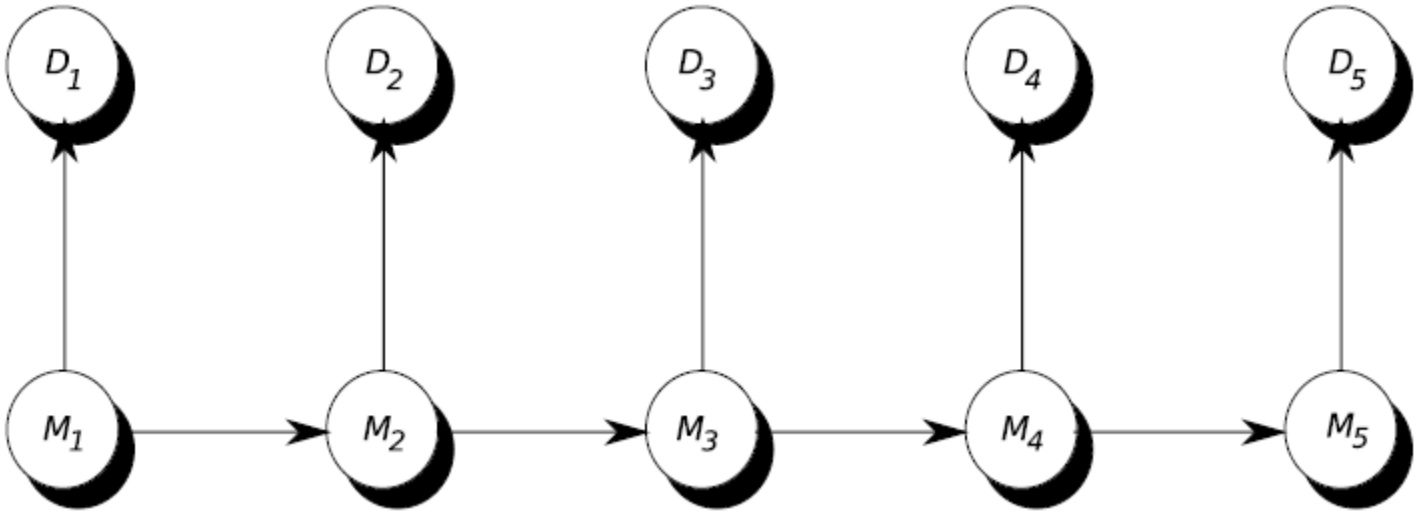
1 + 1 + 4 + 2 + 2

Cas particulier : HMM à horizon fixé

- Chaîne de Markov homogène d'ordre 1

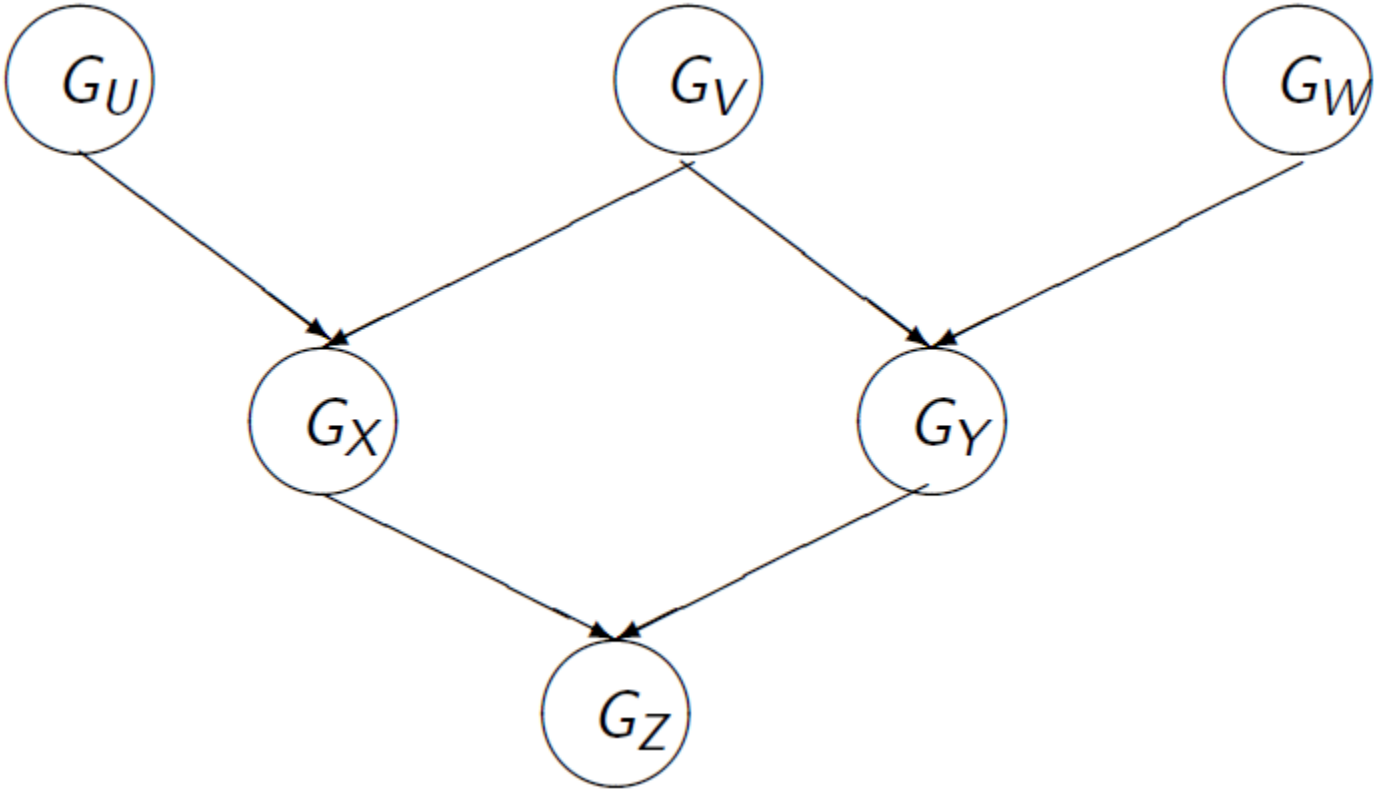


- Chaîne de Markov cachée homogène d'ordre 1



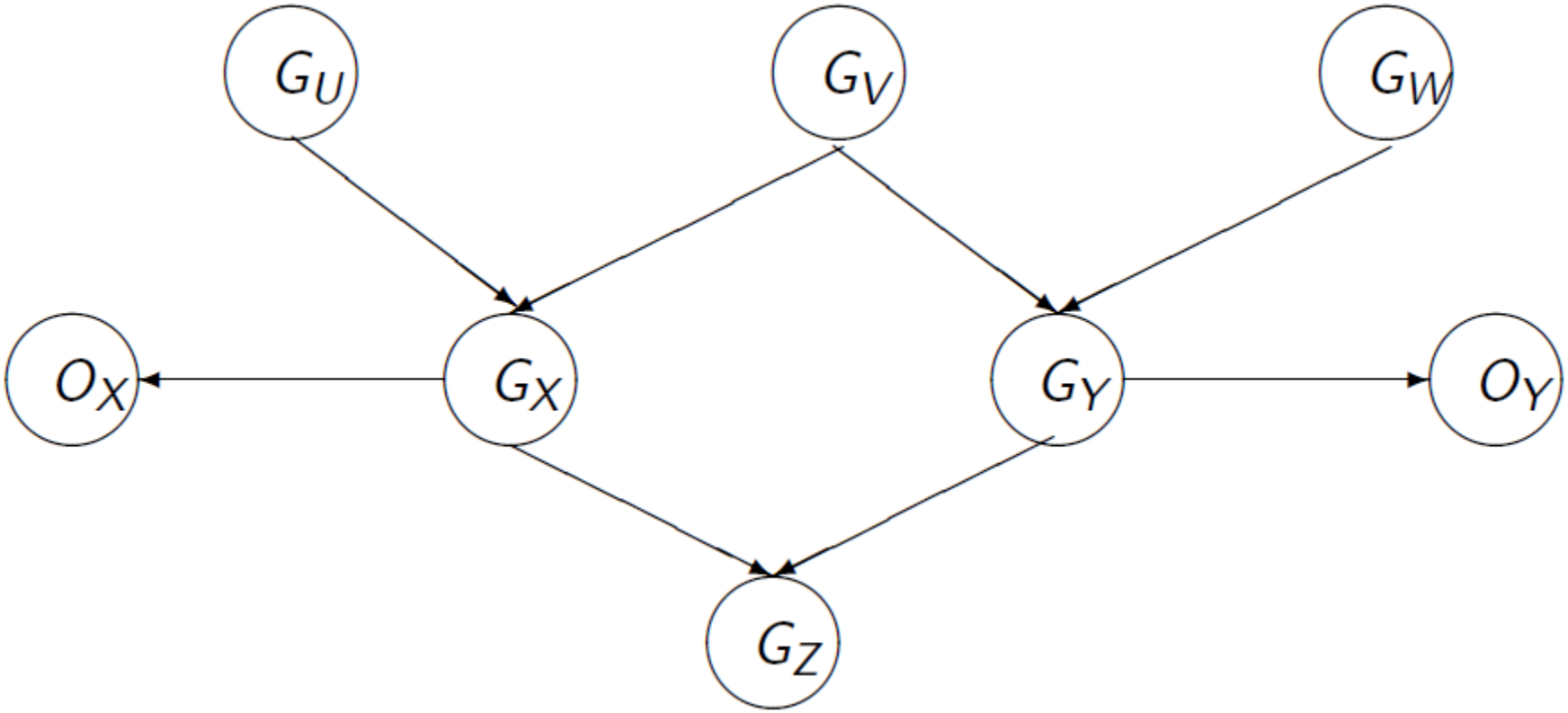
Réseau de génotypes

Soit un pedigree avec 6 individus, notés U, V, W, X, Y, Z, tels que les parents de l'individu X soient $\{U, V\}$, de Y soient $\{V, W\}$ et de Z soient $\{X, Y\}$.



Réseau de génotypes

Comment étendre ce réseau pour prendre en compte des observations incertaines (possiblement erronées) sur les génotypes G_X et G_Y ?



Probabilités a priori des génotypes fondateurs

$p(G_f = \{a, a\})$	0.25
$p(G_f = \{a, b\})$	0.5
$p(G_f = \{b, b\})$	0.25

Probabilités de transmission des génotypes

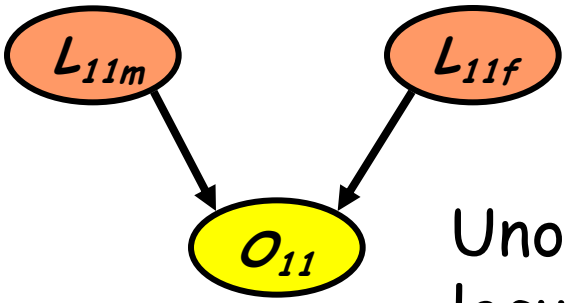
$p(G \mid G_p, G_m)$	$G_m = \{a, a\}$		$G_m = \{a, b\}$		$G_m = \{b, b\}$	
$G_p = \{a, a\}$	$G = \{a, a\}$	1	$G = \{a, a\}$	0.5	$G = \{a, a\}$	0
	$G = \{a, b\}$	0	$G = \{a, b\}$	0.5	$G = \{a, b\}$	1
	$G = \{b, b\}$	0	$G = \{b, b\}$	0	$G = \{b, b\}$	0
$G_p = \{a, b\}$	$G = \{a, a\}$	0.5	$G = \{a, a\}$	0.25	$G = \{a, a\}$	0
	$G = \{a, b\}$	0.5	$G = \{a, b\}$	0.5	$G = \{a, b\}$	0.5
	$G = \{b, b\}$	0	$G = \{b, b\}$	0.25	$G = \{b, b\}$	0.5
$G_p = \{b, b\}$	$G = \{a, a\}$	0	$G = \{a, a\}$	0	$G = \{a, a\}$	0
	$G = \{a, b\}$	1	$G = \{a, b\}$	0.5	$G = \{a, b\}$	0
	$G = \{b, b\}$	0	$G = \{b, b\}$	0.5	$G = \{b, b\}$	1

Probabilités d'erreur de génotypage

$p(O \mid G)$	$G = \{a, a\}$	$G = \{a, b\}$	$G = \{b, b\}$
$O = \{a, a\}$	$1 - \varepsilon$	$\frac{\varepsilon}{2}$	$\frac{\varepsilon}{2}$
$O = \{a, b\}$	$\frac{\varepsilon}{2}$	$1 - \varepsilon$	$\frac{\varepsilon}{2}$
$O = \{b, b\}$	$\frac{\varepsilon}{2}$	$\frac{\varepsilon}{2}$	$1 - \varepsilon$

Allele network

Maternal allele at locus 1 of person 1



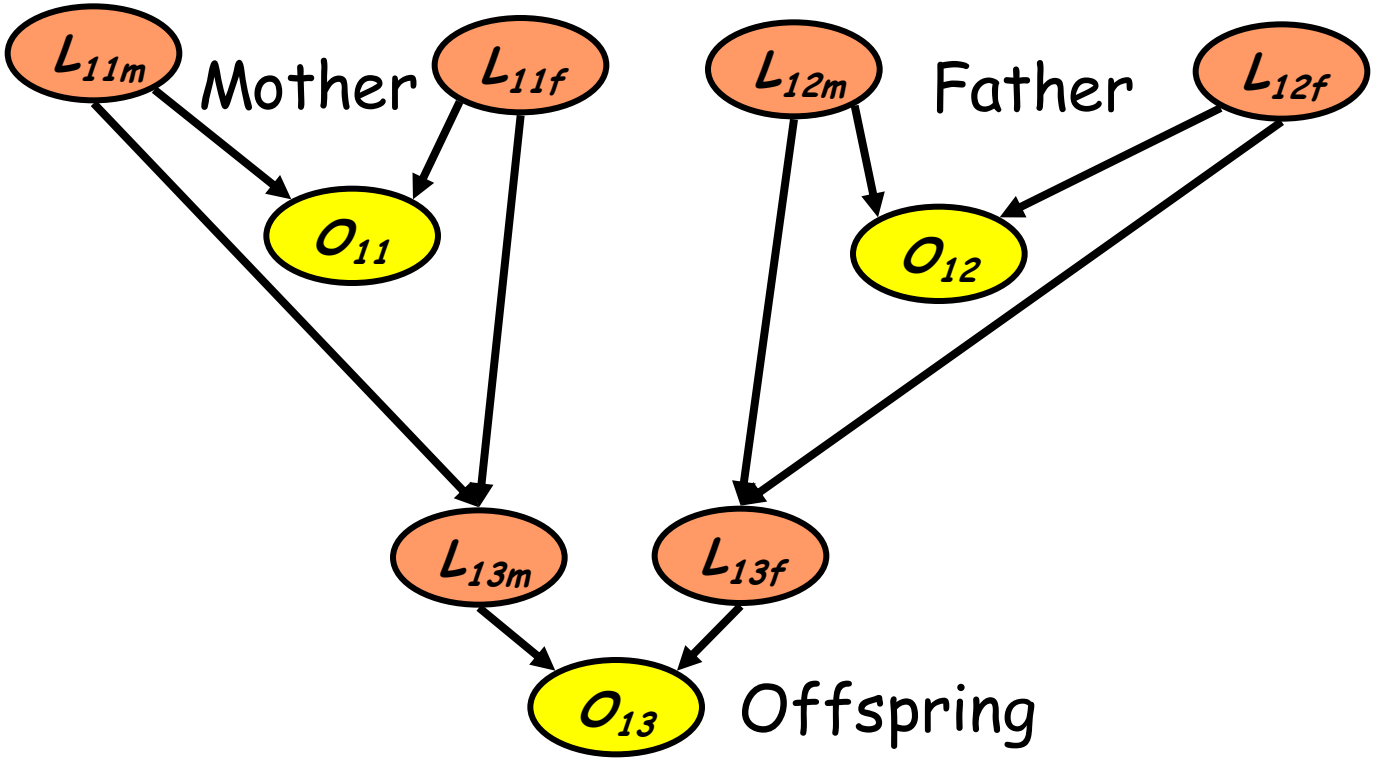
Paternal allele at locus 1 of person 1

Unordered allele pair at locus 1 of person 1 = data

$p(L_{11m} = a)$ is the frequency of allele a.

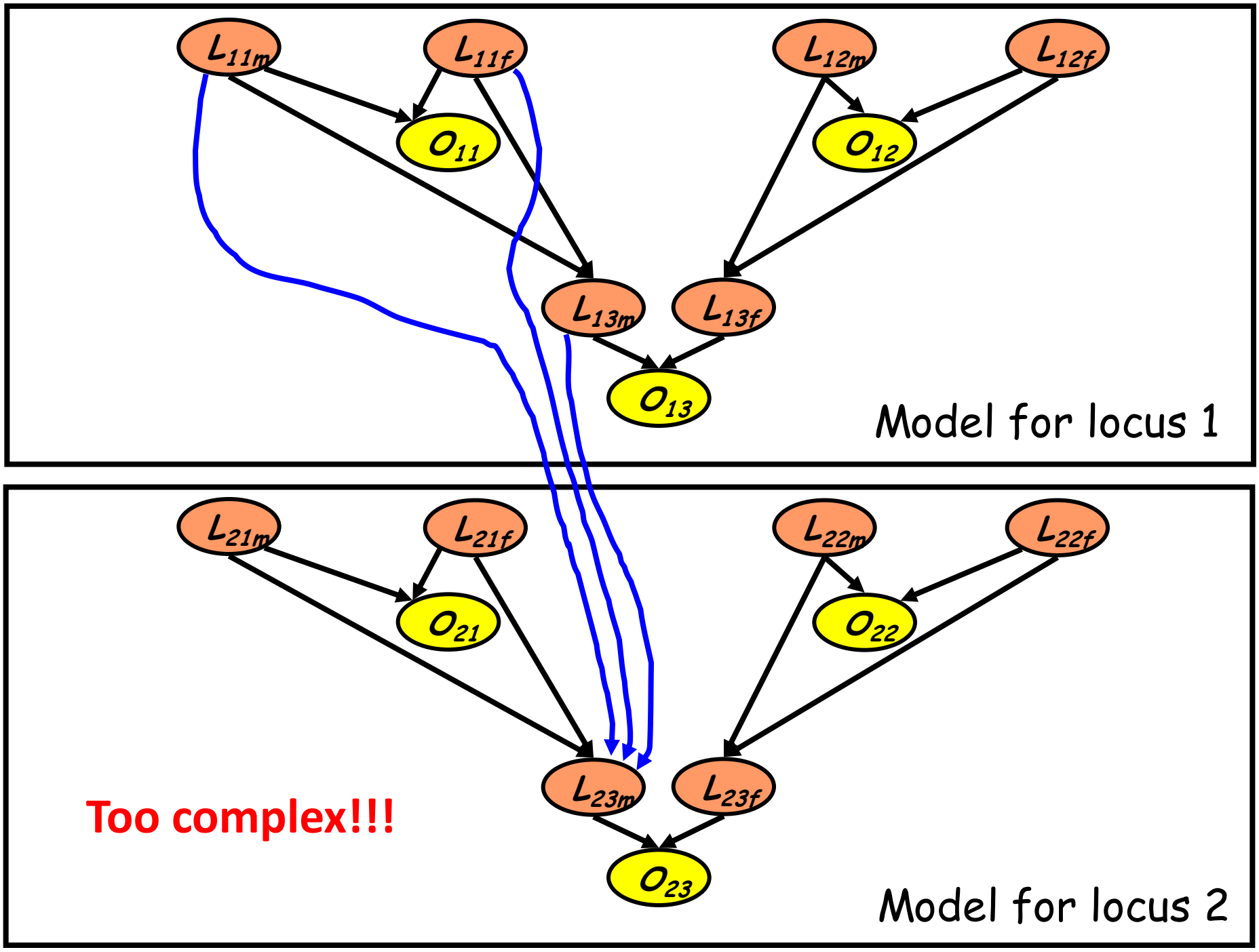
$p(O_{11} \mid l_{11m}, l_{11f}) = 0$ or 1 depending on consistency

Allele network



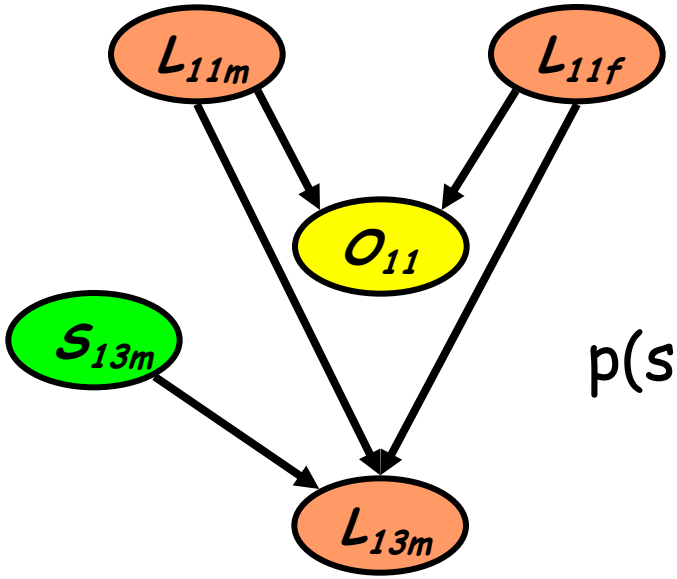
$p(l_{13m} \mid l_{11m}, l_{11f}) = 1/2$ if $l_{13m} = l_{11m}$ or $l_{13m} = l_{11f}$

$p(l_{13m} \mid l_{11m}, l_{11f}) = 0$ otherwise


L_{11m}
L_{11f}
L_{12m}
L_{12f}
O_{11}
O_{12}
L_{13m}
L_{13f}
O_{13}
Model for locus 1
L_{21m}
L_{21f}
L_{22m}
L_{22f}
O_{21}
O_{22}
L_{23m}
L_{23f}
O_{23}
Too complex!!!
Model for locus 2

Adding a selector variable

Selector of maternal allele at locus 1 of person 3

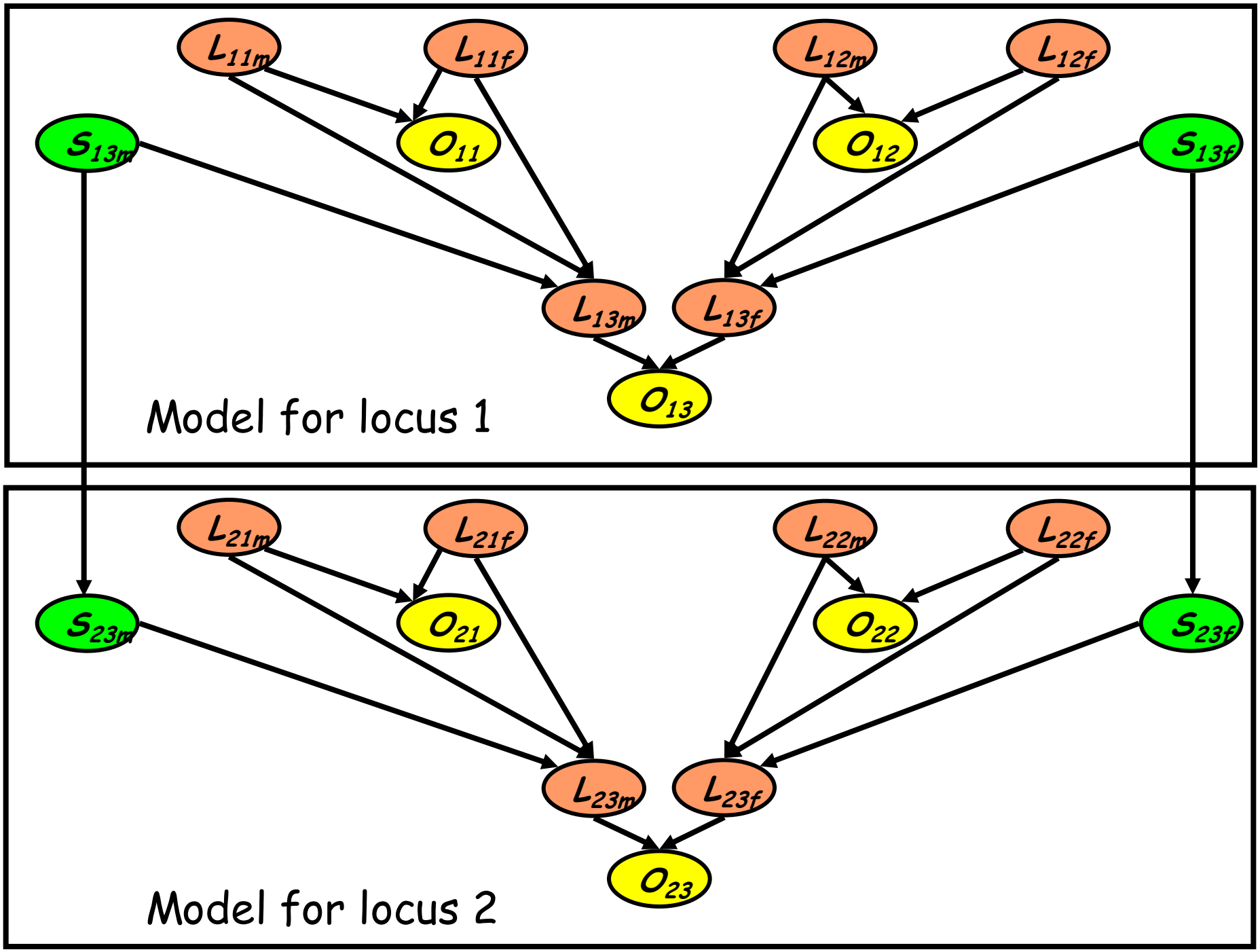


$p(s_{13m}) = \frac{1}{2}$

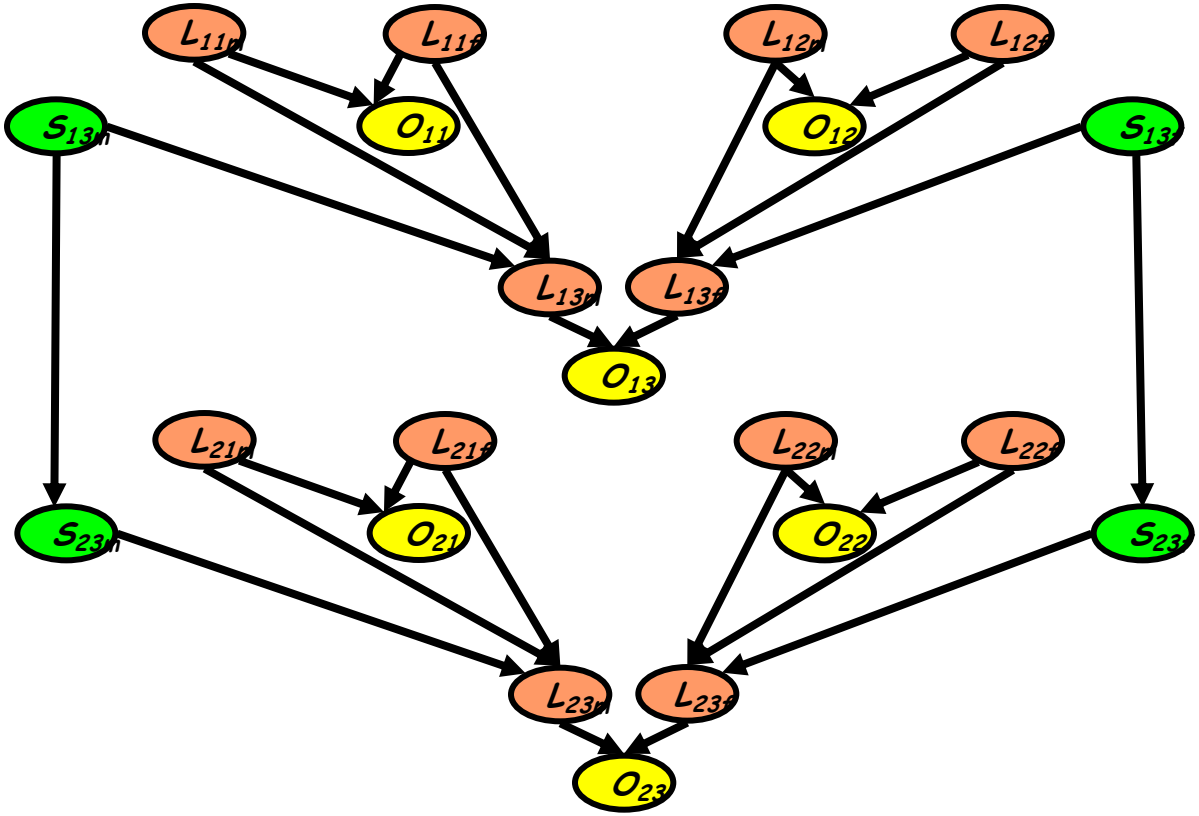
Maternal allele at locus 1 of person 3 (offspring)

Selector variables S_{ijm} are 0 or 1 depending on whose allele is transmitted to offspring i at maternal locus j.

$p(l_{13m} \mid l_{11m}, l_{11f,}, S_{13m}=0) = 1$ if $l_{13m} = l_{11m}$
$p(l_{13m} \mid l_{11m}, l_{11f,}, S_{13m}=1) = 1$ if $l_{13m} = l_{11f}$
$p(l_{13m} \mid l_{11m}, l_{11f,}, s_{13m}) = 0$ otherwise


L_{11m}
L_{11f}
L_{12m}
L_{12f}
S_{13m}
O_{11}
O_{12}
S_{13f}
L_{13m}
L_{13f}
O_{13}
Model for locus 1
L_{21m}
L_{21f}
L_{22m}
L_{22f}
S_{23m}
O_{21}
O_{22}
S_{23f}
L_{23m}
L_{23f}
O_{23}
Model for locus 2

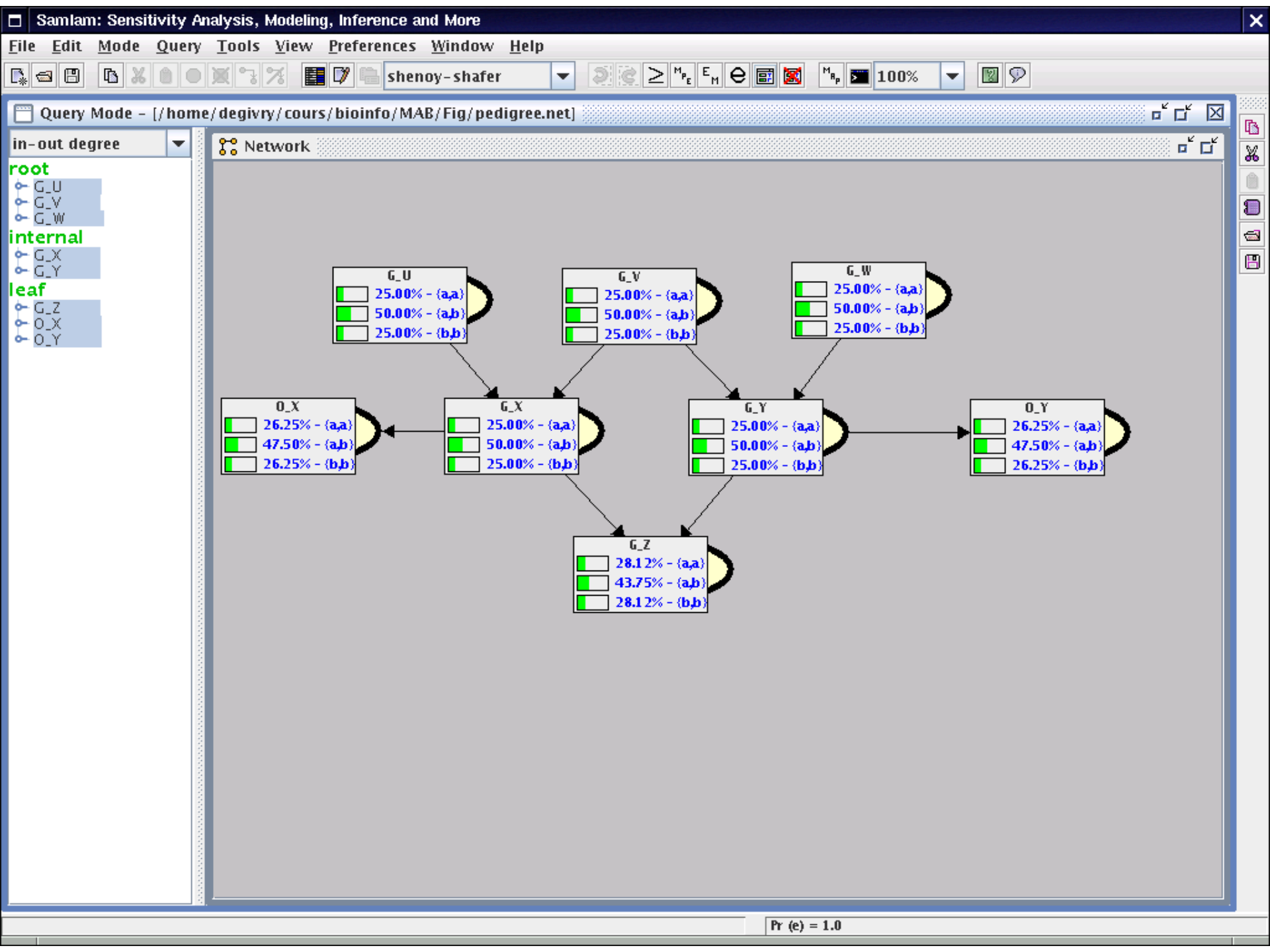
Segregation network

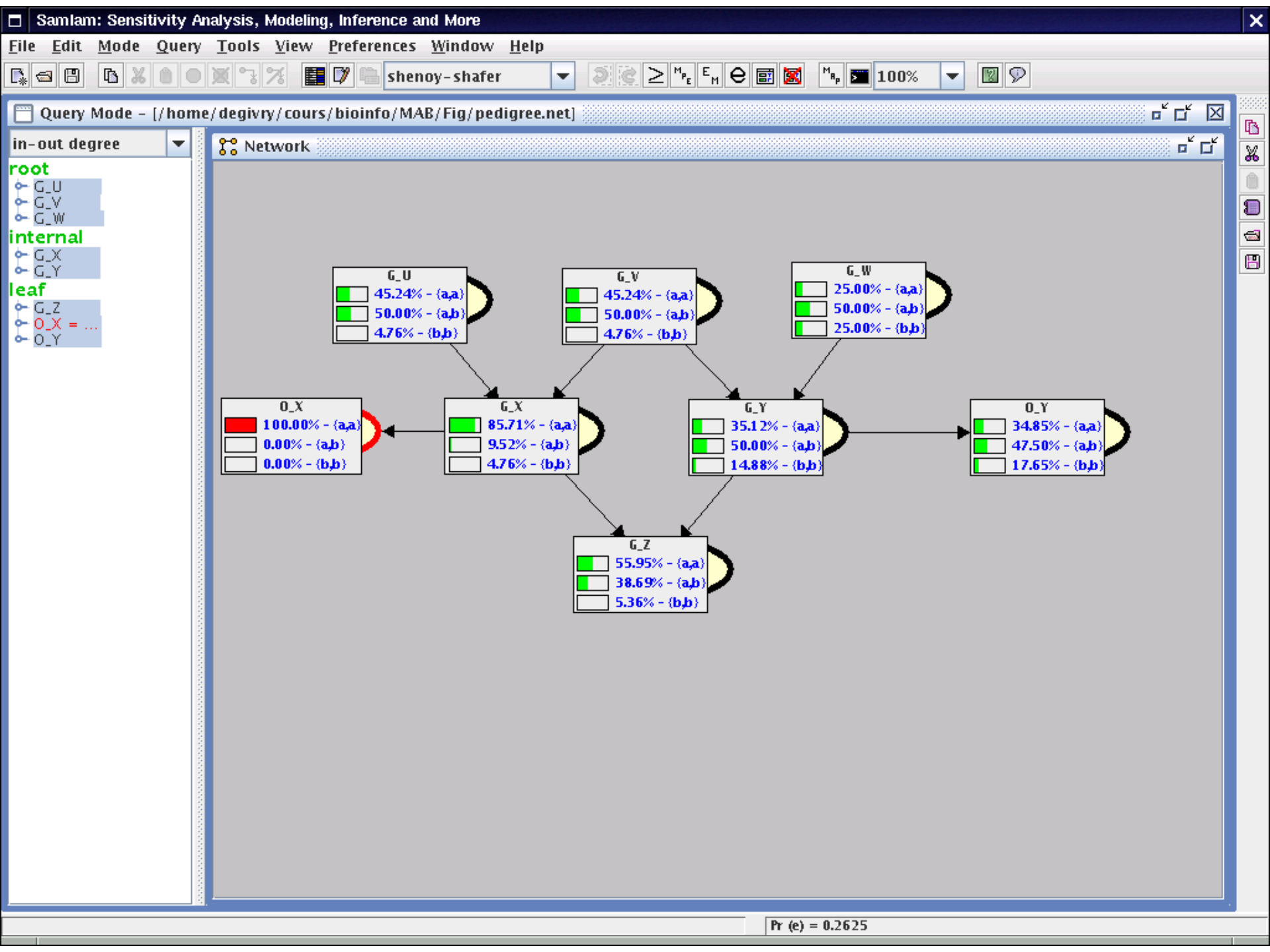


Probabilistic Model for Recombination:

$$p(s_{23t} \mid s_{13t}, \theta) = \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix} \quad \text{where } t \in \{m, f\}$$

θ is the recombination fraction between loci 2 & 1.


Samlam: Sensitivity Analysis, Modeling, Inference and More
File Edit Mode Query Tools View Preferences Window Help
shenoy-shafer
100%
Query Mode - [/home/degivry/cours/bioinfo/MAB/Fig/pedigree.net]
in-out degree
root
G_U
G_V
G_W
internal
G_X
G_Y
leaf
G_Z
O_X
O_Y
Network
G_U
25.00% - (a,a)
50.00% - (a,b)
25.00% - (b,b)
G_V
25.00% - (a,a)
50.00% - (a,b)
25.00% - (b,b)
G_W
25.00% - (a,a)
50.00% - (a,b)
25.00% - (b,b)
O_X
26.25% - (a,a)
47.50% - (a,b)
26.25% - (b,b)
G_X
25.00% - (a,a)
50.00% - (a,b)
25.00% - (b,b)
G_Y
25.00% - (a,a)
50.00% - (a,b)
25.00% - (b,b)
O_Y
26.25% - (a,a)
47.50% - (a,b)
26.25% - (b,b)
G_Z
28.12% - (a,a)
43.75% - (a,b)
28.12% - (b,b)
Pr (e) = 1.0


Samlam: Sensitivity Analysis, Modeling, Inference and More
File Edit Mode Query Tools View Preferences Window Help
shenoy-shafer
100%
Query Mode - [/home/degivry/cours/bioinfo/MAB/Fig/pedigree.net]
in-out degree
root
G_U
G_V
G_W
internal
G_X
G_Y
leaf
G_Z
O_X = ...
O_Y
Network
G_U
45.24% - (a,a)
50.00% - (a,b)
4.76% - (b,b)
G_V
45.24% - (a,a)
50.00% - (a,b)
4.76% - (b,b)
G_W
25.00% - (a,a)
50.00% - (a,b)
25.00% - (b,b)
O_X
100.00% - (a,a)
0.00% - (a,b)
0.00% - (b,b)
G_X
85.71% - (a,a)
9.52% - (a,b)
4.76% - (b,b)
G_Y
35.12% - (a,a)
50.00% - (a,b)
14.88% - (b,b)
O_Y
34.85% - (a,a)
47.50% - (a,b)
17.65% - (b,b)
G_Z
55.95% - (a,a)
38.69% - (a,b)
5.36% - (b,b)
Pr (e) = 0.2625

Samlam: Sensitivity Analysis, Modeling, Inference and More

File Edit Mode Query Tools View Preferences Window Help

shenoy-shafer 100%

Query Mode - [/home/degivry/cours/bioinfo/MAB/Fig/pedigree.net]

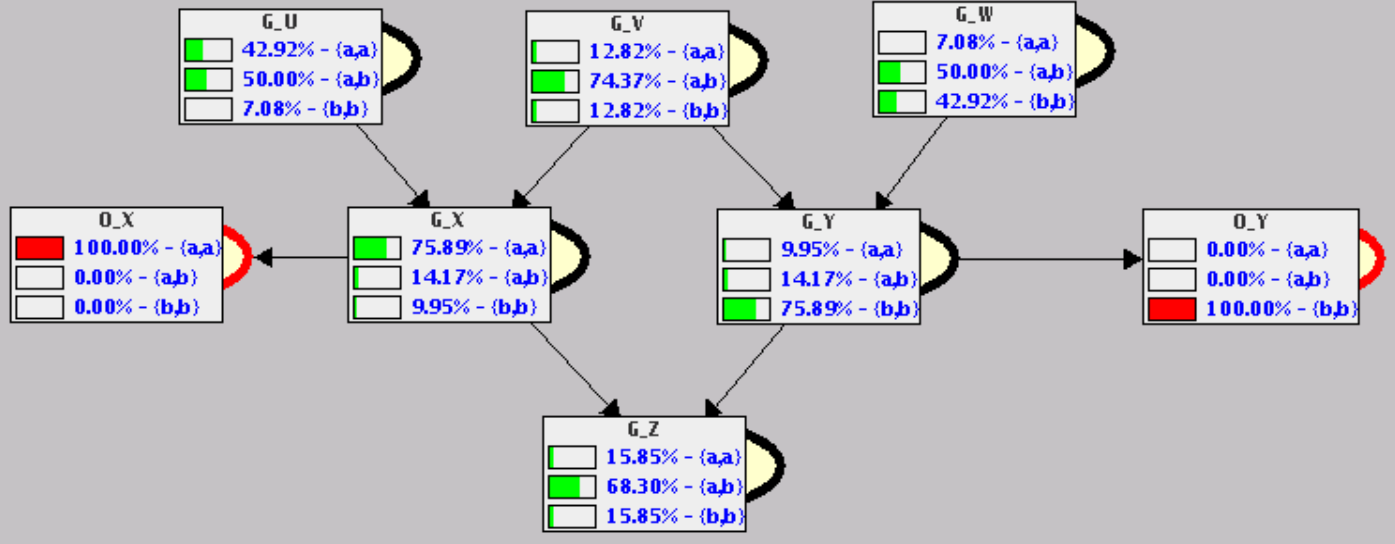
in-out degree

root
- G_U
- G_V
- G_W

internal
- G_X
- G_Y

leaf
- G_Z
- O_X = ...
- O_Y = ...

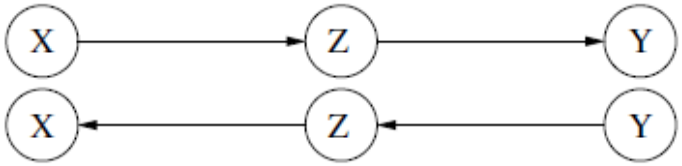
Network



Pr (e) = 0.046328125

Circulation de l'information

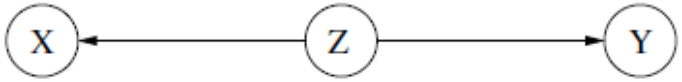
Connexion en série



ne circule de X vers Y
que si Z n'est pas connu

$$G_X = \{a, a\} \implies p(G_Y = \{a, a\}) > p(G_Y = \{b, b\})$$

Connexion divergente



ne circule de X vers Y
que si Z n'est pas connu

$$G_X = \{a, a\} \implies p(G_Y = \{a, a\}) > p(G_Y = \{b, b\})$$

Connexion convergente

ne circule de X vers Y
que si Z est connu

$$G_X = \{a, a\}, G_Z = \{a, b\} \implies b \in G_Y$$

Séparation dirigée (d-separation)

X et Y sont d-séparés par Z (noté $\langle X|Z|Y\rangle$) si pour **toutes** les chaînes entre X et Y :

- La chaîne converge en un sommet W, tel que $W \neq Z$ et W n'est pas une cause directe ou *indirecte* de Z ;
- *OU* la chaîne passe par Z, et est soit divergente, soit en série au sommet Z.

Il y alors blocage de l'information par la connaissance de Z.

Extension au cas où **X**, **Y** et **Z** sont des ensembles disjoints.
L'absence de chaîne entre X et Y implique la d-séparation entre X et Y quelque soit Z.
Complexité de tester $\langle X|Z|Y\rangle$ linéaire en la taille du réseau.

Lien entre d-separation et indépendence conditionnelle

Indépendance conditionnelle

Soit des ensembles disjoints $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$. $\mathbf{X}$ est indépendant de $\mathbf{Y}$ conditionnellement à $\mathbf{Z}$ (noté $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$) ssi :

$$\begin{aligned}
\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z} &\iff \begin{cases} p(\mathbf{X}|\mathbf{Y},\mathbf{Z}) = p(\mathbf{X}|\mathbf{Z}) \\ \text{et} \quad p(\mathbf{Y}|\mathbf{X},\mathbf{Z}) = p(\mathbf{Y}|\mathbf{Z}) \end{cases} \\
&\iff p(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = p(\mathbf{X}|\mathbf{Z}) \cdot p(\mathbf{Y}|\mathbf{Z}) \\
&\iff p(\mathbf{X},\mathbf{Y},\mathbf{Z}) = p(\mathbf{X}|\mathbf{Z}) \cdot p(\mathbf{Y}|\mathbf{Z}) \cdot p(\mathbf{Z})
\end{aligned}$$

Théorème

Soit un réseau Bayésien $\vec{G}$ définissant la loi de probabilité jointe $p(\mathbf{V})$. Il vérifie la *propriété orientée de Markov globale (OG)* :

$$\forall \mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V} \text{ disjoints }, \langle \mathbf{X}|\mathbf{Z}|\mathbf{Y} \rangle \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$$

Limites d'expressivité des réseaux bayésiens

- Il existe des lois dont le *modèle d'indépendence* n'est pas représentable par un réseau bayésien

  Exemple: M = { $X \perp\!\!\!\perp W \mid \{Y,Z\}$, $Y \perp\!\!\!\perp Z \mid \{X,W\}$ }

- *Réseau de Markov* : graphe non-orienté avec séparation directe
- Quelque soit le formalisme (réseau bayésien ou réseau de Markov), il existe des lois non représentables

Requêtes

Vraisemblance des observations (PR)

Calculer la probabilité $p(\mathbf{E} = \mathbf{e})$ ayant observé les variables $\mathbf{E} \subset \mathbf{V}$.

Probabilités Marginales (MAR)

Calculer la probabilité *a posteriori* $p(V_i|\mathbf{E} = \mathbf{e})$ de toutes les variables $V_i \in \mathbf{V} \setminus \mathbf{E}$.

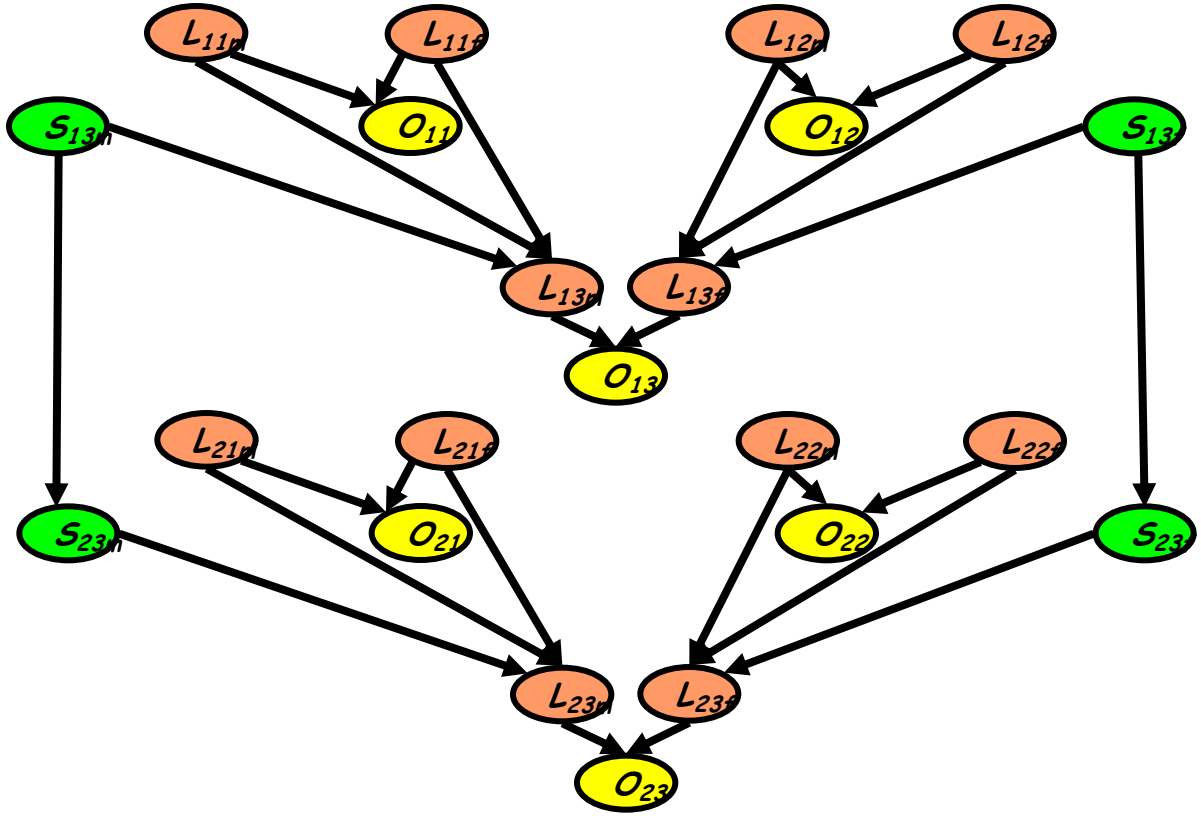
Maximum a Posteriori hypothesis (MAP)

Rechercher une affectation *partielle* $\mathbf{u}$ d'un sous-ensemble de variables $\mathbf{U} \subseteq \mathbf{V} \setminus \mathbf{E}$ de probabilité $p(\mathbf{U} = \mathbf{u}|\mathbf{E} = \mathbf{e})$ maximum.

Most Probable Explanation (MPE)

Rechercher une affectation *complète* $\mathbf{u}$ des variables $\mathbf{U} = \mathbf{V} \setminus \mathbf{E}$ de probabilité $p(\mathbf{U} = \mathbf{u}|\mathbf{E} = \mathbf{e})$ maximum.

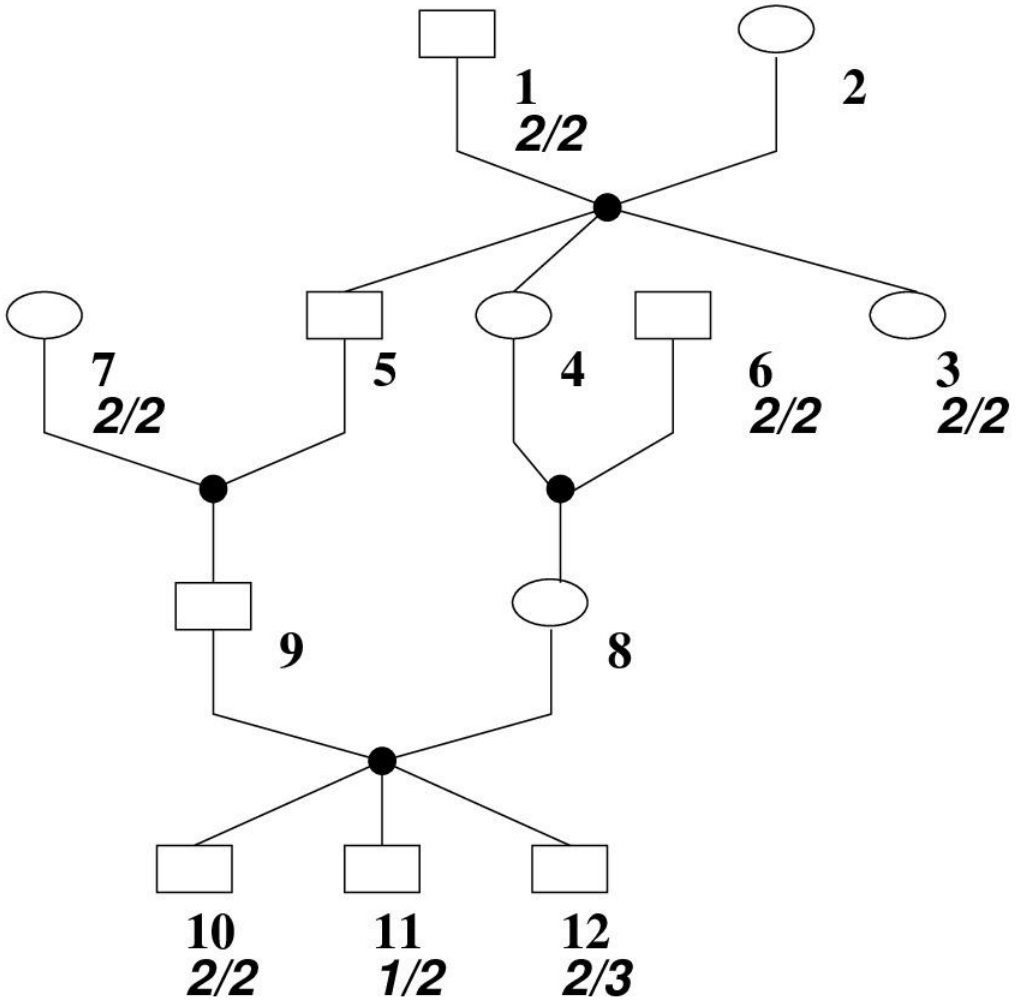
Segregation network



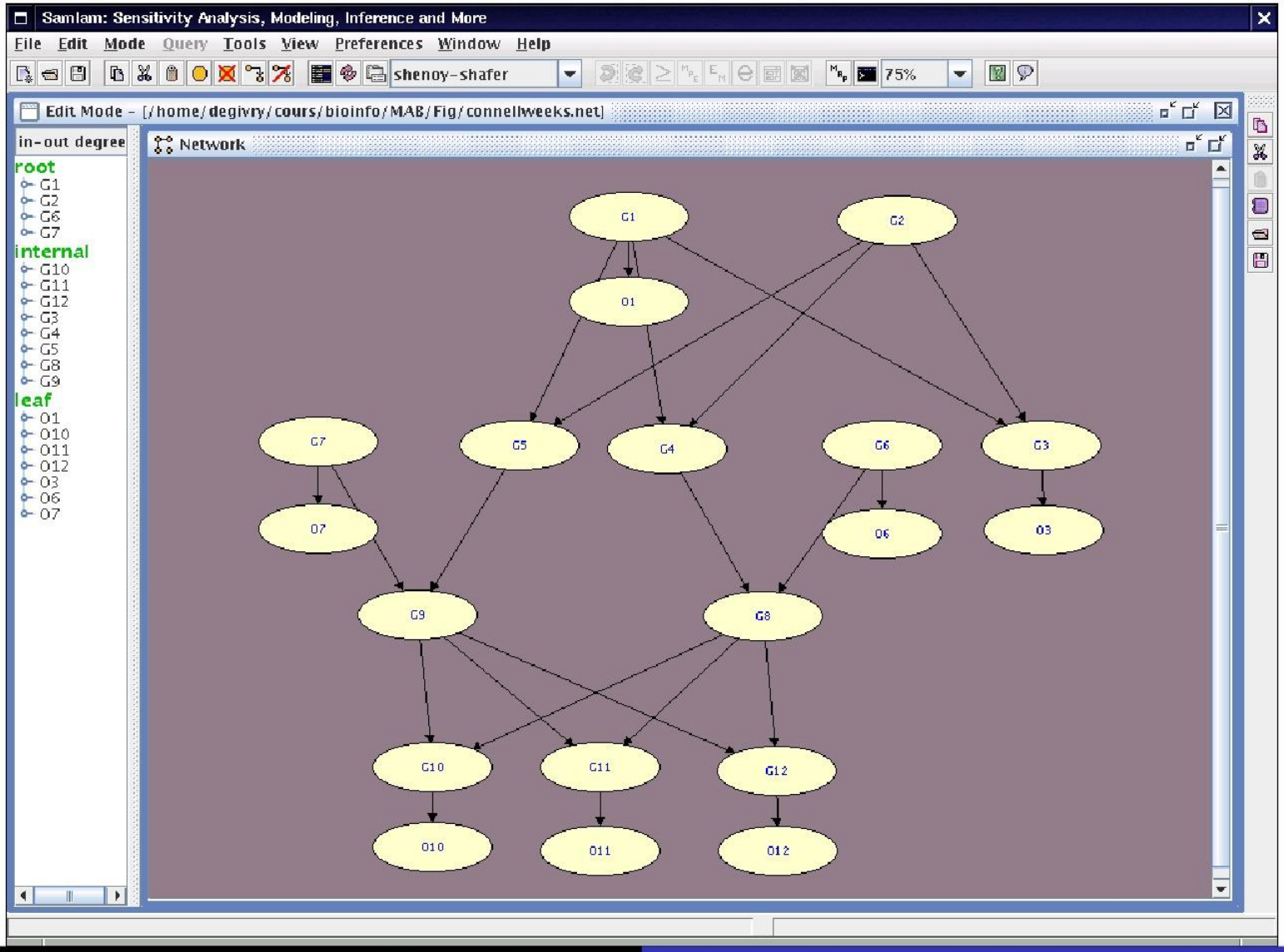
Genetic linkage analysis (PR): $\text{argmax}_{\theta} \sum_{\mathbf{L,S}} p(\mathbf{L, S, O} \mid \theta)$

Haplotype reconstruction (MPE): $\text{argmax}_{\mathbf{L,S}} p(\mathbf{L, S \mid O}, \theta)$

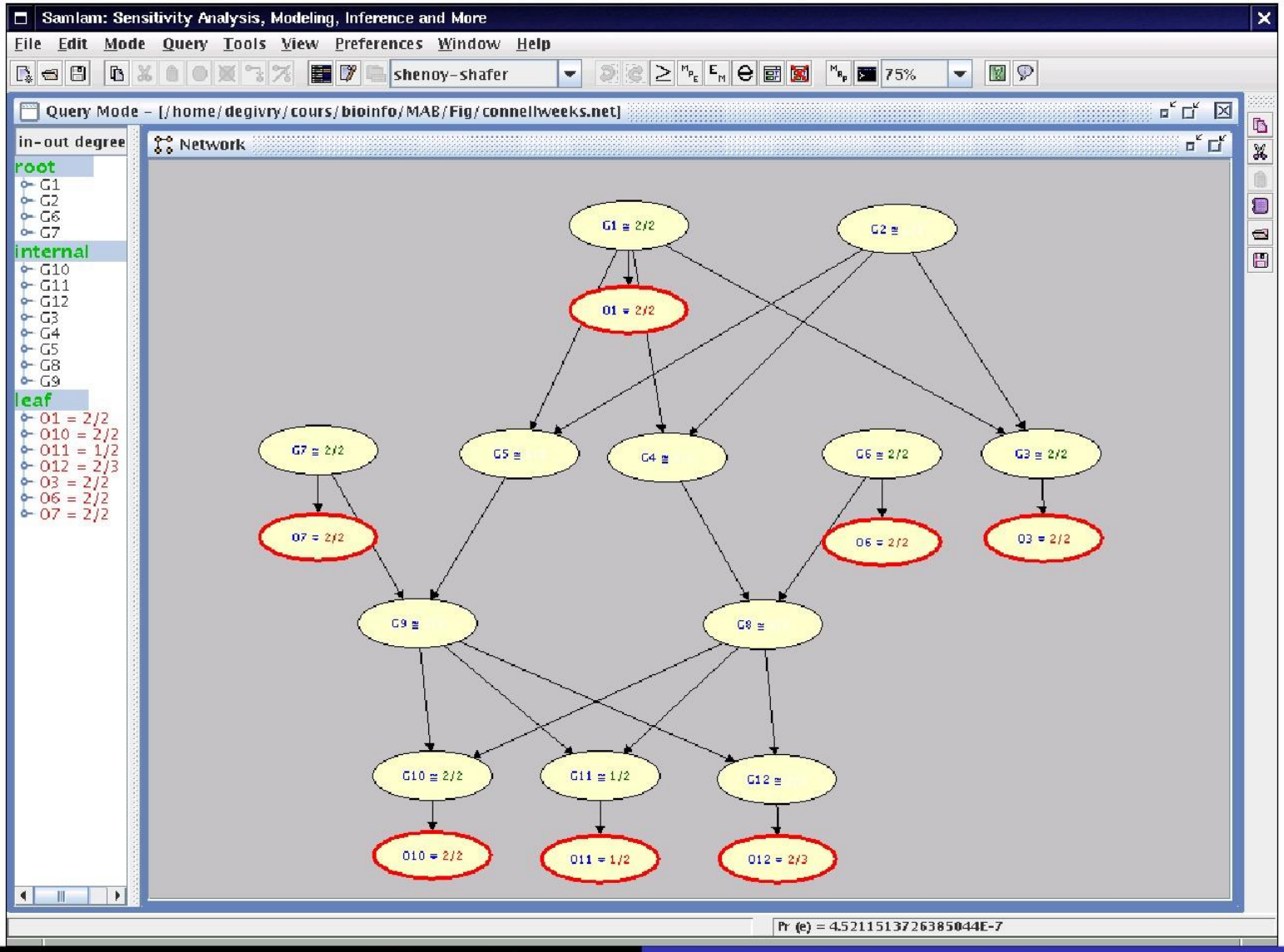
Exemple de pedigree avec génotypes observés



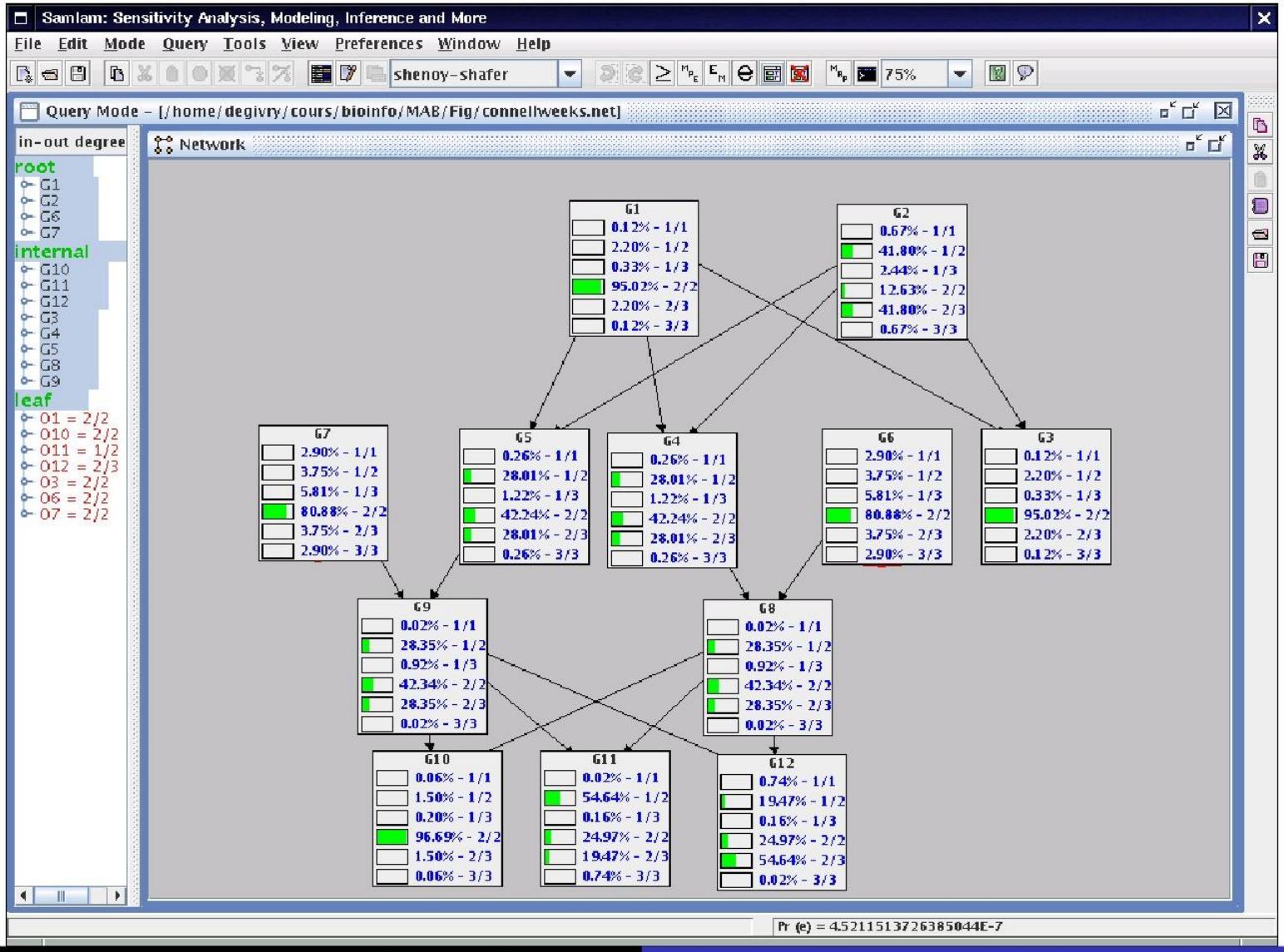
Réseau de génotypes



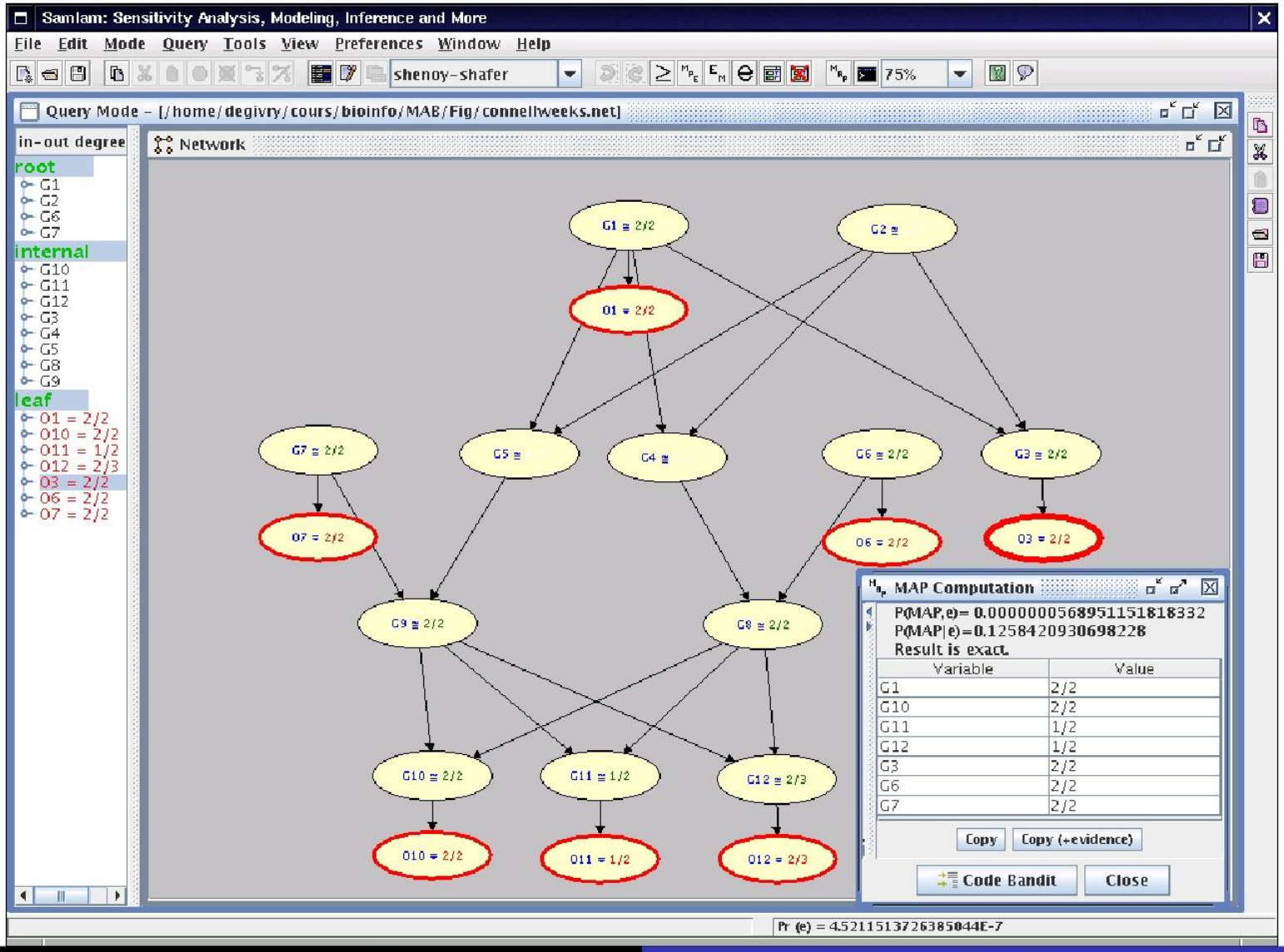
PR


Samlam: Sensitivity Analysis, Modeling, Inference and More
File Edit Mode Query Tools View Preferences Window Help
shenoy-shafer
75%
Query Mode – [/home/degivry/cours/bioinfo/MAB/Fig/connellweeks.net]
in-out degree
root
G1
G2
G6
G7
internal
G10
G11
G12
G3
G4
G5
G8
G9
leaf
O1 = 2/2
O10 = 2/2
O11 = 1/2
O12 = 2/3
O3 = 2/2
O6 = 2/2
O7 = 2/2
Network
G1 = 2/2
G2 =
O1 = 2/2
G7 = 2/2
G5 =
G4 =
G6 = 2/2
G3 = 2/2
O7 = 2/2
O6 = 2/2
O3 = 2/2
G9 =
G8 =
G10 = 2/2
G11 = 1/2
G12 =
O10 = 2/2
O11 = 1/2
O12 = 2/3
Pr (e) = 4.521151372638504E-7

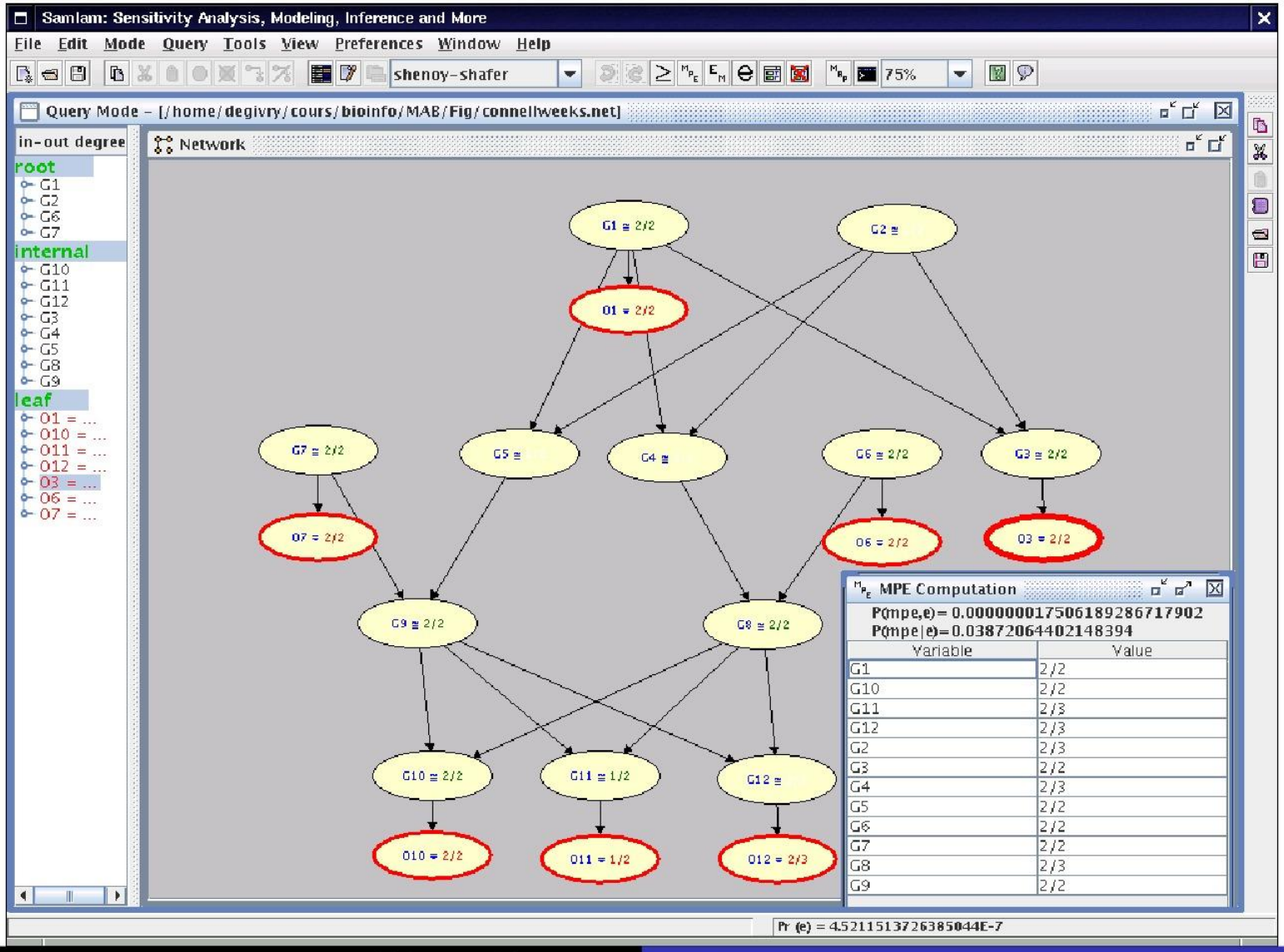
MAR


Samlam: Sensitivity Analysis, Modeling, Inference and More
File Edit Mode Query Tools View Preferences Window Help
shenoy-shafer
75%
Query Mode – [/home/degivry/cours/bioinfo/MAB/Fig/connellweeks.net]
in-out degree
root
G1
G2
G6
G7
internal
G10
G11
G12
G3
G4
G5
G8
G9
leaf
O1 = 2/2
O10 = 2/2
O11 = 1/2
O12 = 2/3
O3 = 2/2
O6 = 2/2
O7 = 2/2
Network
G1
0.12% - 1/1
2.20% - 1/2
0.33% - 1/3
95.02% - 2/2
2.20% - 2/3
0.12% - 3/3
G2
0.67% - 1/1
41.80% - 1/2
2.44% - 1/3
12.63% - 2/2
41.80% - 2/3
0.67% - 3/3
G7
2.90% - 1/1
3.75% - 1/2
5.81% - 1/3
80.88% - 2/2
3.75% - 2/3
2.90% - 3/3
G5
0.26% - 1/1
28.01% - 1/2
1.22% - 1/3
42.24% - 2/2
28.01% - 2/3
0.26% - 3/3
G4
0.26% - 1/1
28.01% - 1/2
1.22% - 1/3
42.24% - 2/2
28.01% - 2/3
0.26% - 3/3
G6
2.90% - 1/1
3.75% - 1/2
5.81% - 1/3
80.88% - 2/2
3.75% - 2/3
2.90% - 3/3
G3
0.12% - 1/1
2.20% - 1/2
0.33% - 1/3
95.02% - 2/2
2.20% - 2/3
0.12% - 3/3
G9
0.02% - 1/1
28.35% - 1/2
0.92% - 1/3
42.34% - 2/2
28.35% - 2/3
0.02% - 3/3
G8
0.02% - 1/1
28.35% - 1/2
0.92% - 1/3
42.34% - 2/2
28.35% - 2/3
0.02% - 3/3
G10
0.06% - 1/1
1.50% - 1/2
0.20% - 1/3
96.69% - 2/2
1.50% - 2/3
0.06% - 3/3
G11
0.02% - 1/1
54.64% - 1/2
0.16% - 1/3
24.97% - 2/2
19.47% - 2/3
0.74% - 3/3
G12
0.74% - 1/1
19.47% - 1/2
0.16% - 1/3
24.97% - 2/2
54.64% - 2/3
0.02% - 3/3
Pr (e) = 4.521151372638504E-7

MAP


Samlam: Sensitivity Analysis, Modeling, Inference and More
File Edit Mode Query Tools View Preferences Window Help
shenoy-shafer
75%
Query Mode – [/home/degivry/cours/bioinfo/MAB/Fig/connellweeks.net]
in-out degree
root
G1
G2
G6
G7
internal
G10
G11
G12
G3
G4
G5
G8
G9
leaf
O1 = 2/2
O10 = 2/2
O11 = 1/2
O12 = 2/3
O3 = 2/2
O6 = 2/2
O7 = 2/2
Network
G1 = 2/2
G2 =
O1 = 2/2
G7 = 2/2
G5 =
G4 =
G6 = 2/2
G3 = 2/2
O7 = 2/2
O6 = 2/2
O3 = 2/2
G9 = 2/2
G8 = 2/2
G10 = 2/2
G11 = 1/2
G12 = 2/3
O10 = 2/2
O11 = 1/2
O12 = 2/3
MAP Computation
P(MAP,e)= 0.00000005689511518183 32
P(MAP|e)=0.12584209306982 28
Result is exact.
Variable
Value
G1
2/2
G10
2/2
G11
1/2
G12
1/2
G3
2/2
G6
2/2
G7
2/2
Copy
Copy (+evidence)
Code Bandit
Close
Pr (e) = 4.521151372638504 4E-7

MPE


Samlam: Sensitivity Analysis, Modeling, Inference and More
File Edit Mode Query Tools View Preferences Window Help
shenoy-shafer
75%
Query Mode – [/home/degivry/cours/bioinfo/MAB/Fig/connellweeks.net]
in-out degree
root
G1
G2
G6
G7
internal
G10
G11
G12
G3
G4
G5
G8
G9
leaf
O1 = ...
O10 = ...
O11 = ...
O12 = ...
O3 = ...
O6 = ...
O7 = ...
Network
G1 = 2/2
G2 =
O1 = 2/2
G7 = 2/2
G5 =
G4 =
G6 = 2/2
G3 = 2/2
O7 = 2/2
O6 = 2/2
O3 = 2/2
G9 = 2/2
G8 = 2/2
G10 = 2/2
G11 = 1/2
G12 =
O10 = 2/2
O11 = 1/2
O12 = 2/3
MPE Computation
P(mpe,e)= 0.00000001750618928671790 2
P(mpe|e)= 0.038720644021483 94
Variable
Value
G1 2/2
G10 2/2
G11 2/3
G12 2/3
G2 2/3
G3 2/2
G4 2/3
G5 2/2
G6 2/2
G7 2/2
G8 2/3
G9 2/2
Pr (e) = 4.521151372638504 4E-7

Requêtes et algorithmes – liens avec HMM

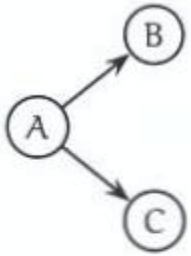
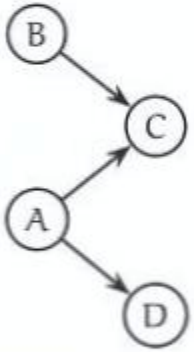
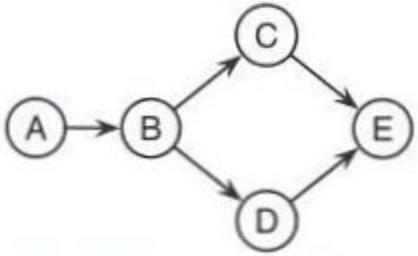
- Recherche de l'état le plus probable (MPE)
  - algorithme de Viterbi (opérateurs max-prod)
- Vraisemblance des observations (PR)
  - algorithme Forward (opérateurs sum-prod)
- Probabilités marginales (MAR)
  - algorithme Forward-Backward (opérateurs sum-prod)

➔ Cadre général :
programmation dynamique non sérielle

Inférence probabiliste

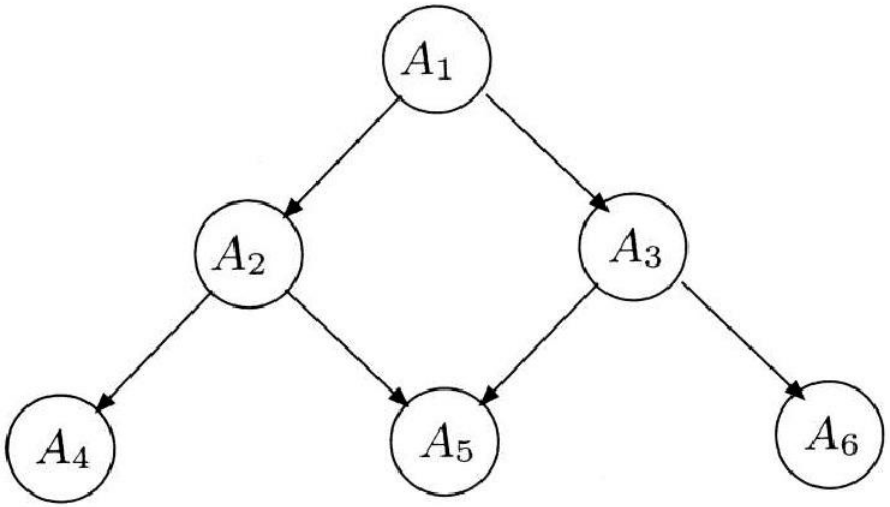
- Méthodes exactes
  - programmation dynamique non sérielle
  - recherche arborescente
  - méthodes hybrides
- Méthodes approchées
  - Méthodes par passage de messages (*loopy belief propagation*,…)
  - Méthodes par simulation (*MCMC*,…)

Inférence probabiliste

	Chaîne	p(C | A)?
	Arbre	p(C | B)?
	Polyarbre	p(D | B)?
	Réseau avec boucles	p(E | A)?

Algorithme d'inférence exacte fondé sur l'élimination de variable

Soit les variables $\mathbf{V} = A_1, A_2, A_3, A_4, A_5, A_6$
et les *potentiels* $\phi_1(A_1) = p(A_1)$, $\phi_2(A_2, A_1) = p(A_2|A_1)$,
$\phi_3(A_3, A_1) = p(A_3|A_1)$, $\phi_4(A_4, A_2) = p(A_4|A_2)$,
$\phi_5(A_5, A_2, A_3) = p(A_5|A_2, A_3)$, $\phi_6(A_6, A_3) = p(A_6|A_3)$.



Calcul de $p(A_4)$?

Les deux grandes règles

Règle de la somme - élimination de variable - marginale

$$p(X) = \sum_{Y} p(X, Y)$$

Règle du produit

$$p(X, Y) = p(Y|X).p(X) = p(X|Y).p(Y)$$

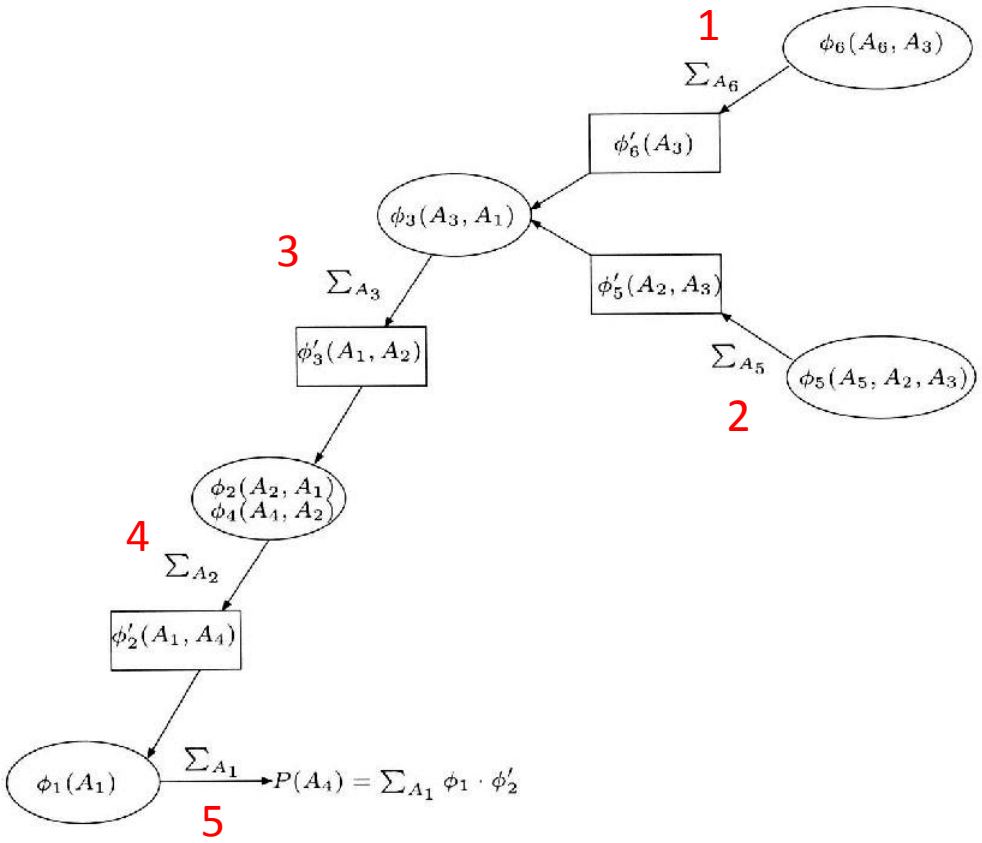
Algorithme d'inférence par élimination de variable

Utilisation de la règle de distribution entre somme et produit.

$$
\begin{aligned}
p(A_4) &= \sum_{A_1,A_2,A_3,A_5,A_6} p(\mathbf{V}) \\
&= \sum_{A_1,A_2,A_3,A_5,A_6} \phi_1\phi_2\phi_3\phi_4\phi_5\phi_6 \\
&= \sum_{A_1}\phi_1(A_1)\sum_{A_2}\phi_2(A_2,A_1)\phi_4(A_4,A_2)\sum_{A_3}\phi_3(A_3,A_1) \\
&\quad \sum_{A_5}\phi_5(A_5,A_2,A_3)\sum_{A_6}\phi_6(A_6,A_3)
\end{aligned}
$$

Factorisation des calculs selon un arbre

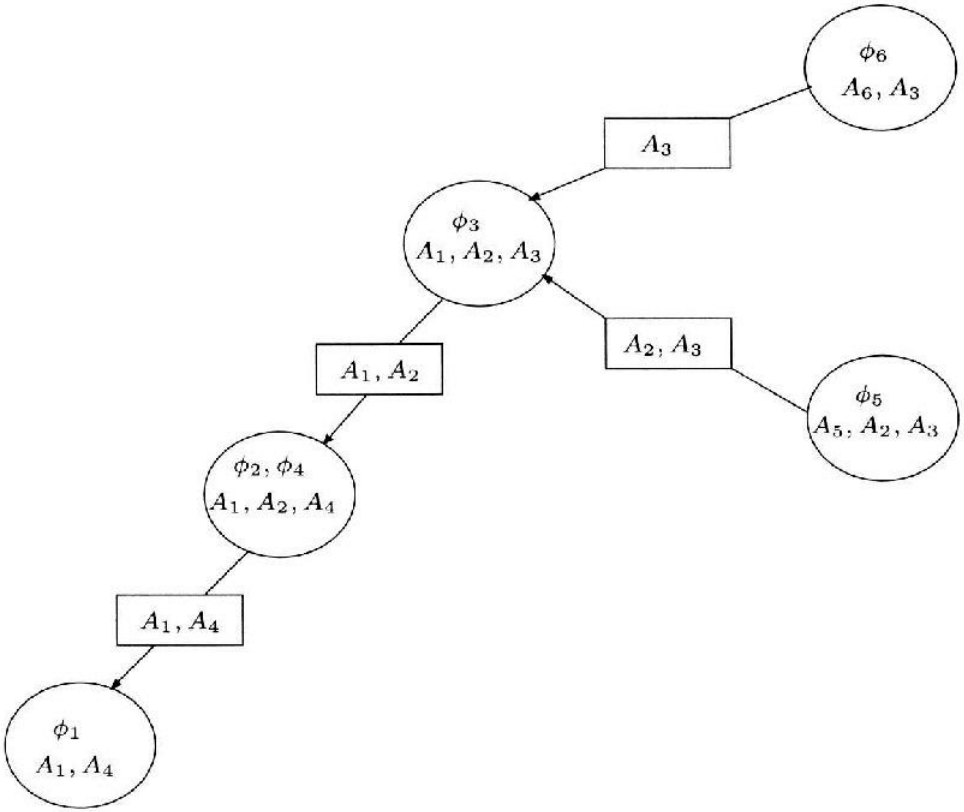
Introduction de **nouveaux** potentiels obtenus par marginalisation (par ex., $\phi'_6(A_3) = \sum_{A_6} \phi_6(A_6, A_3)$)



Multiplication des potentiels dans les ellipses avec les potentiels entrants.

Impact de l'ordre d'élimination de variable

Ordre d'élimination : $(A_6, A_5, A_3, A_2, A_1)$



Complexité exponentielle en le nombre de variables. Ici, temps $O(d^3)$ et espace $O(d^2)$.

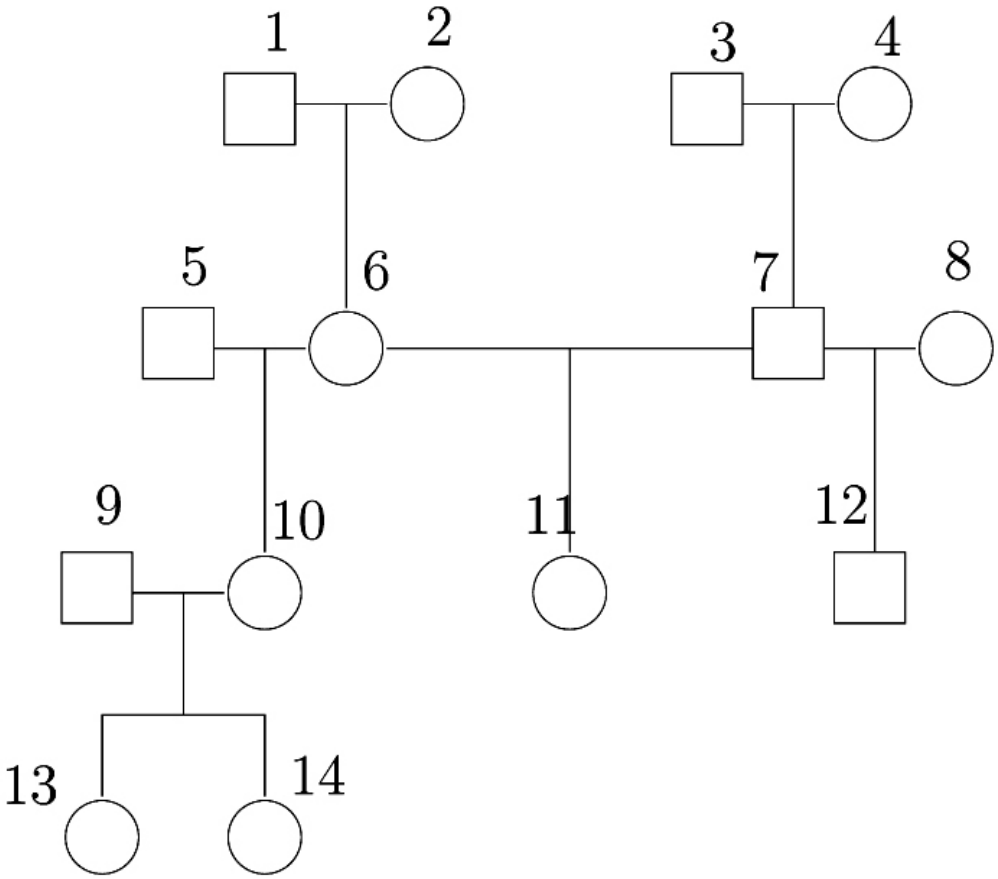
Algorithme d'inférence exacte fondé sur un regroupement des variables

Principe général

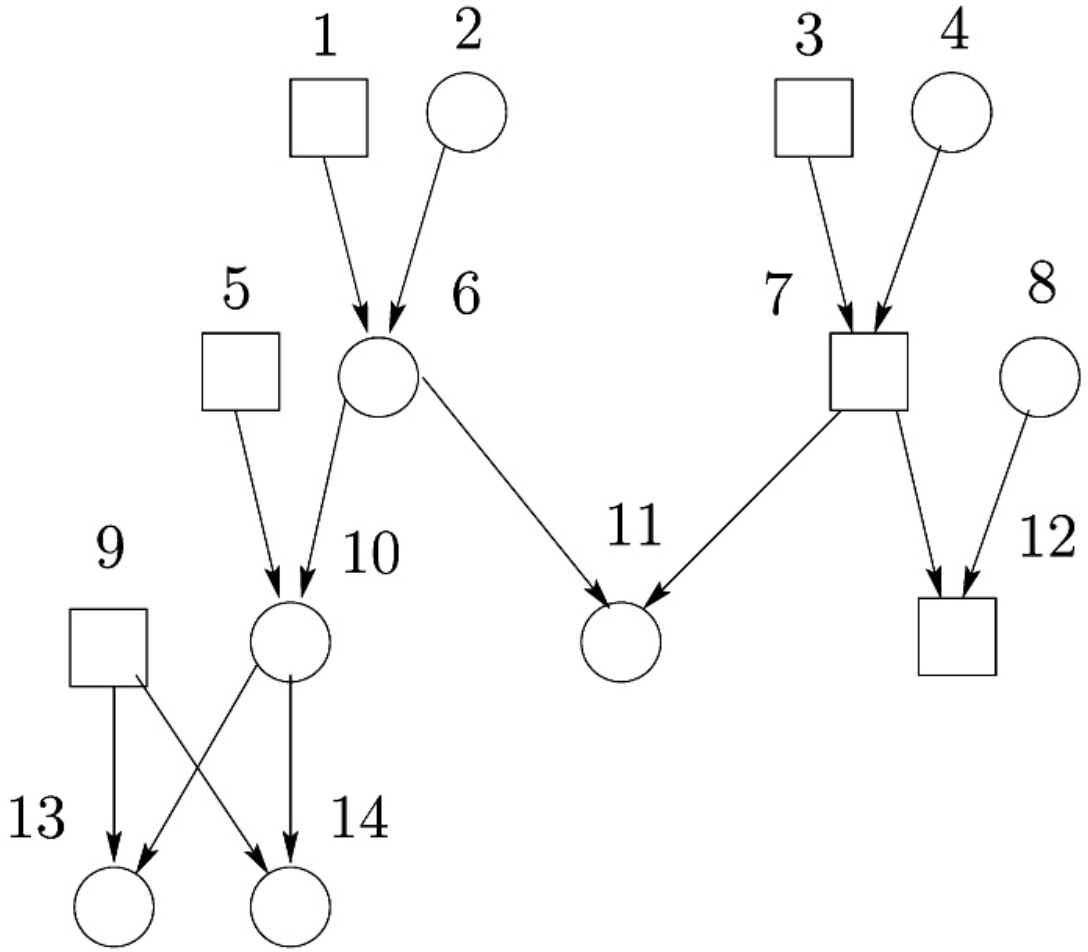
- Phase 1 : Compilation
  Idée : regrouper les variables en noeuds de calcul (*clusters*) de manière à définir un arbre de calcul.
  - Transformation du réseau Bayésien en graphe non-orienté
  - Trouver un jeu d'élimination des variables
  - Construire l'arbre de clusters
  - Remplir les clusters avec les probabilités conditionnelles
  - Ajouter les observations

  Remarque : manipulations de la phase 1 uniquement sur la structure du réseau
- Phase 2 : Propagation
  Idée : appliquer l'algorithme de propagation des polyarbres (version simplifiée au cas d'un arbre) sur l'arbre de clusters

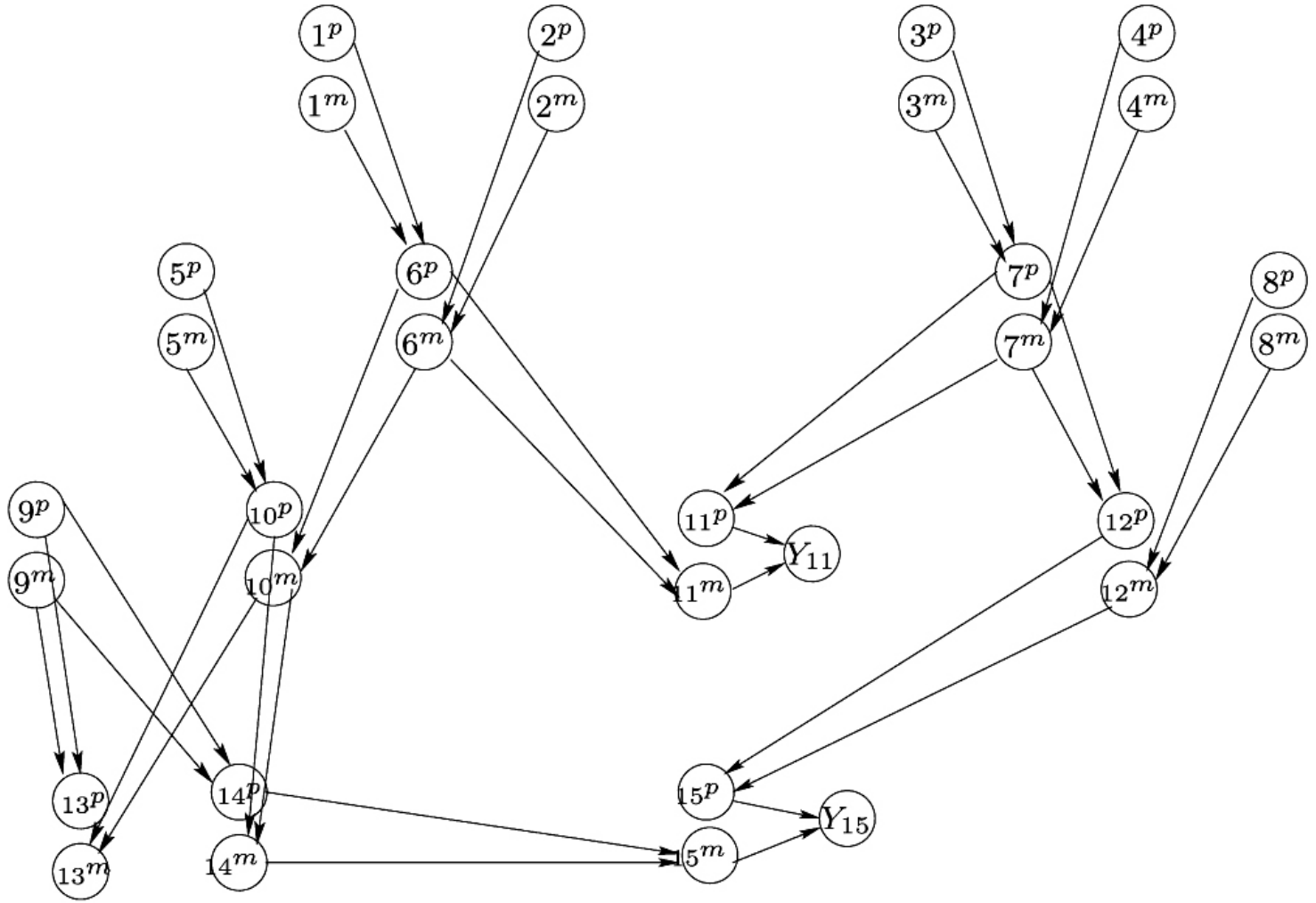
Simple pedigree example



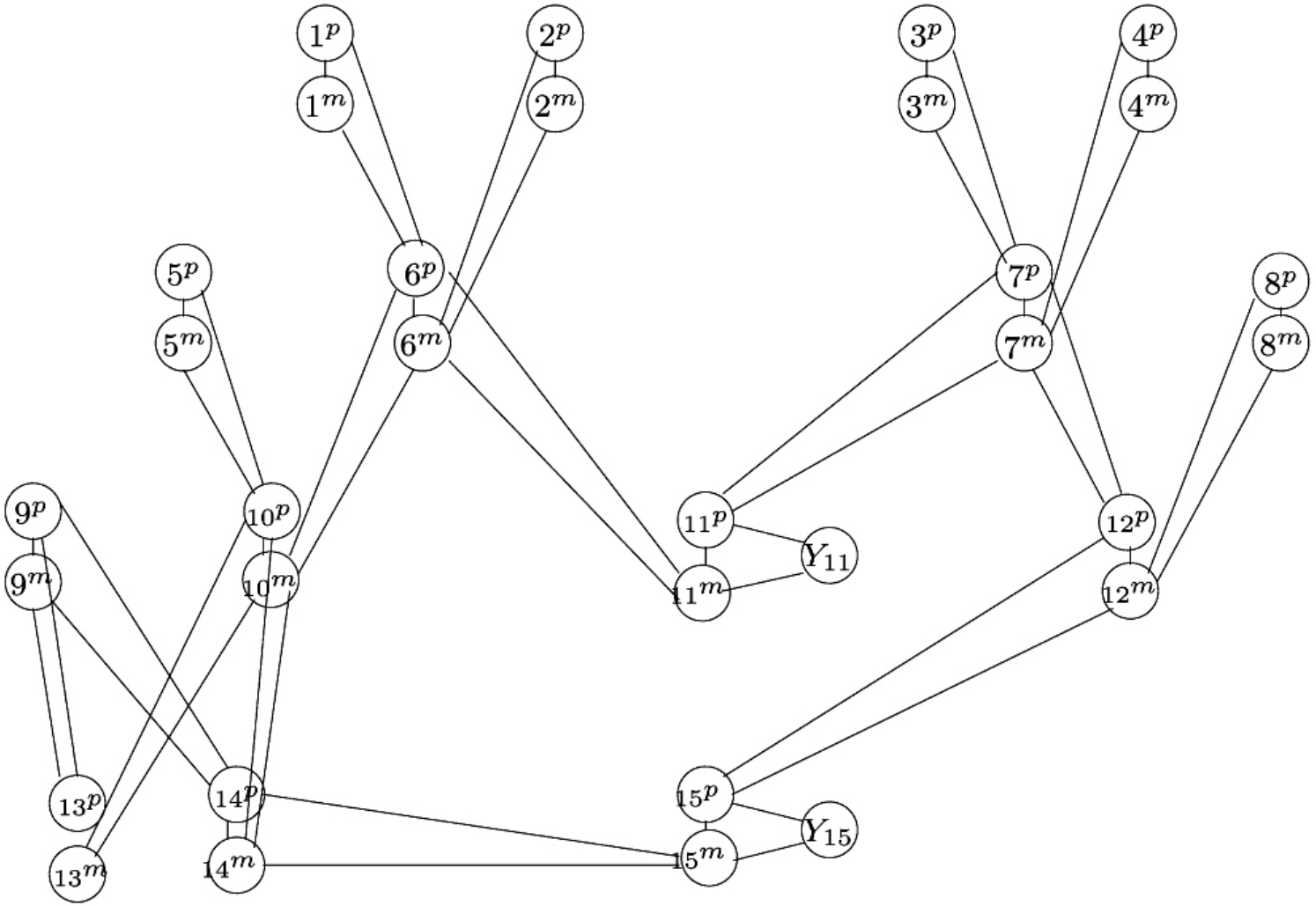
Genotype network



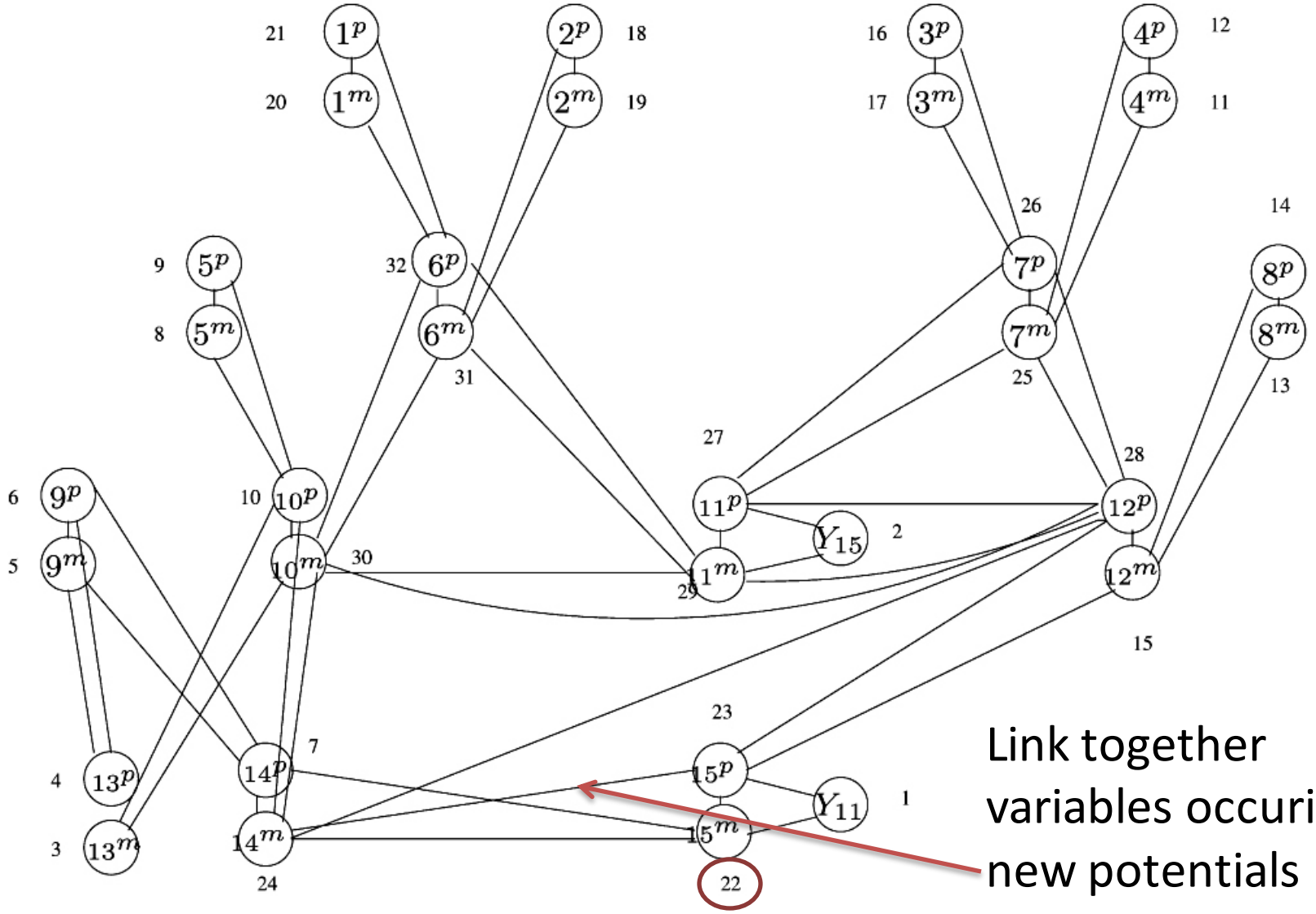
Allele network with phenotypic information on individual 11 and new individual 15



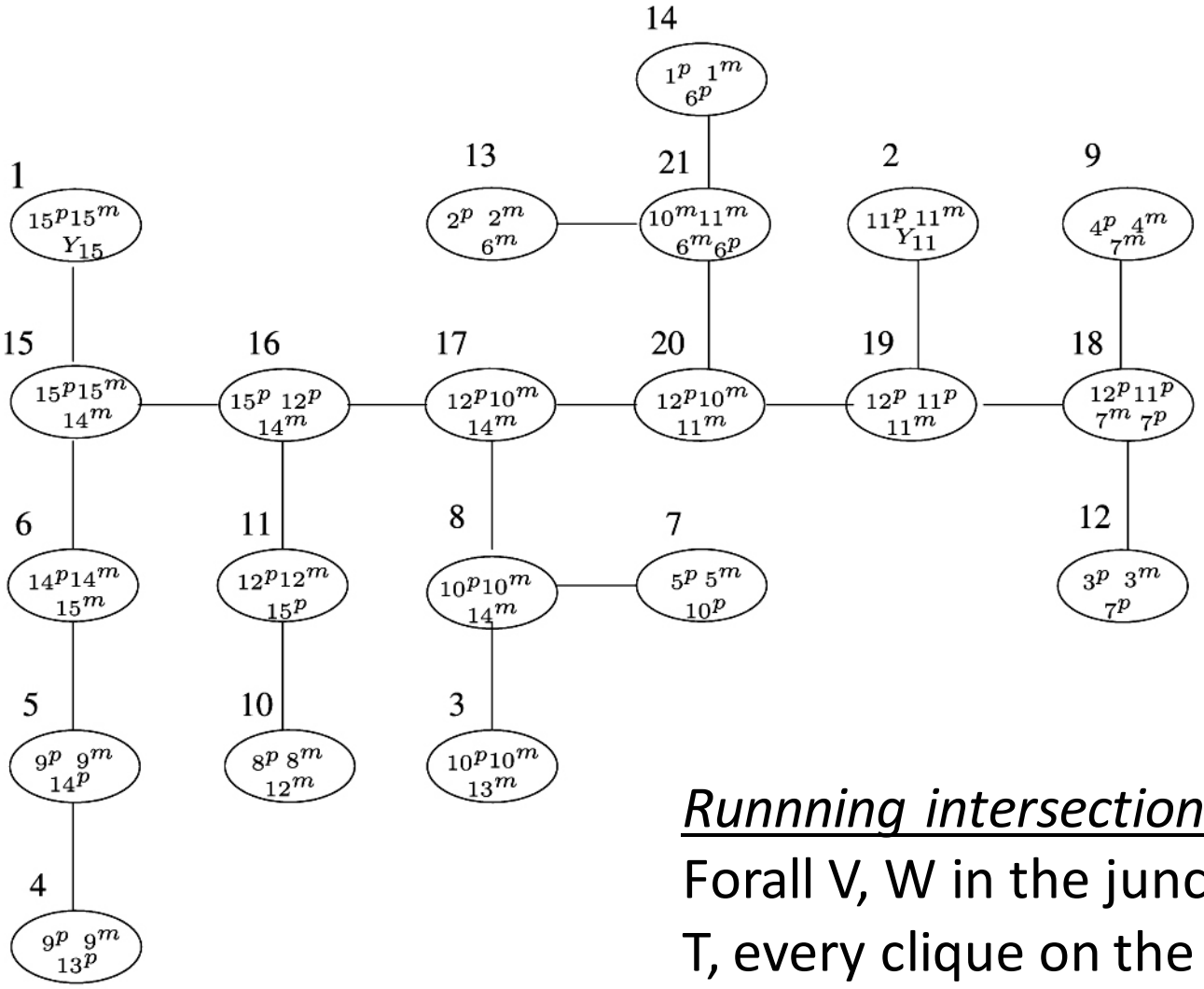
Primal graph (*moral graph*)



Triangulated graph
(*by following an elimination order*)



Junction tree (*find all maximal cliques, build a tree from clique intersections*)



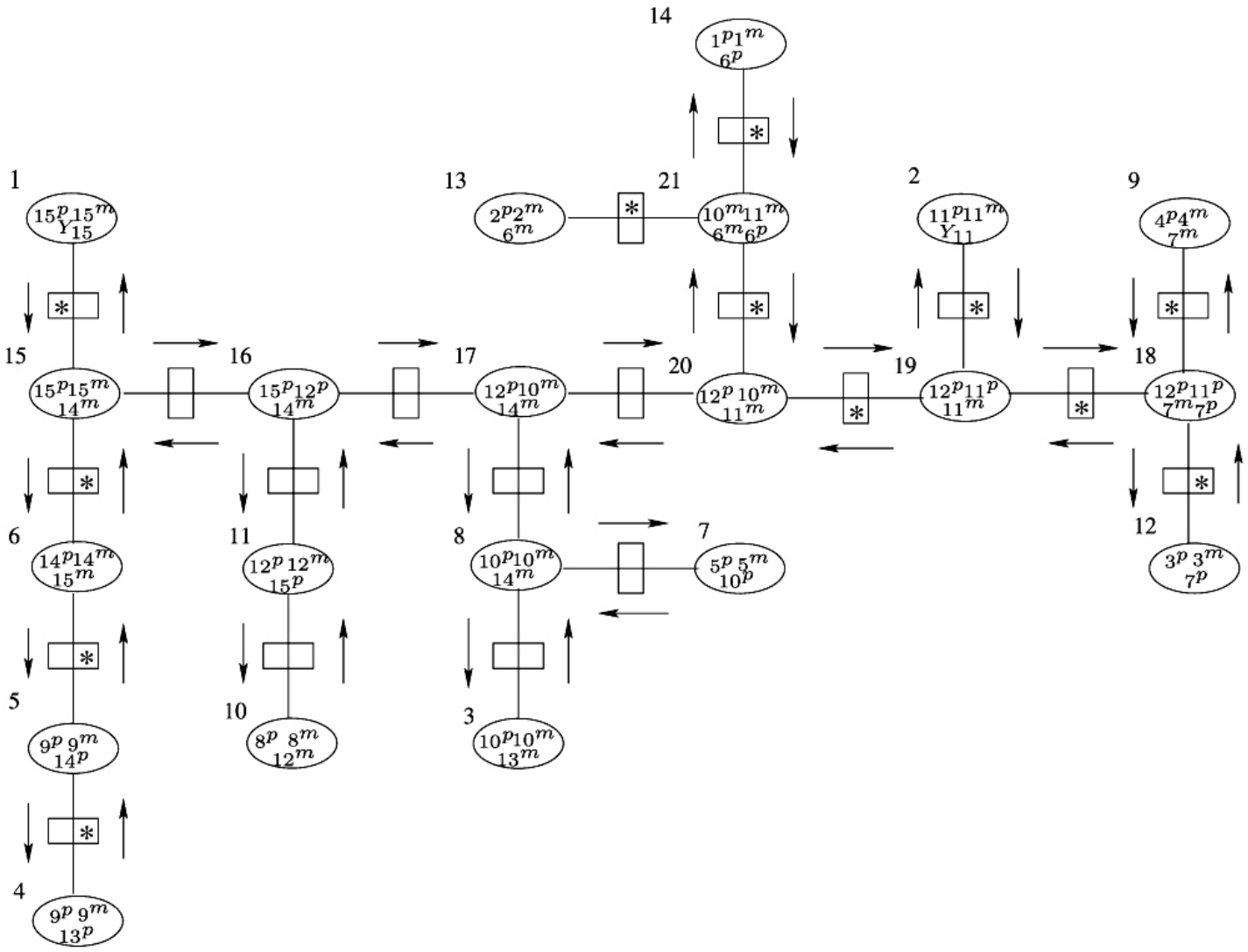
Runnning intersection property

Forall V, W in the junction tree T, every clique on the path from V to W in T must contain V ∩ W

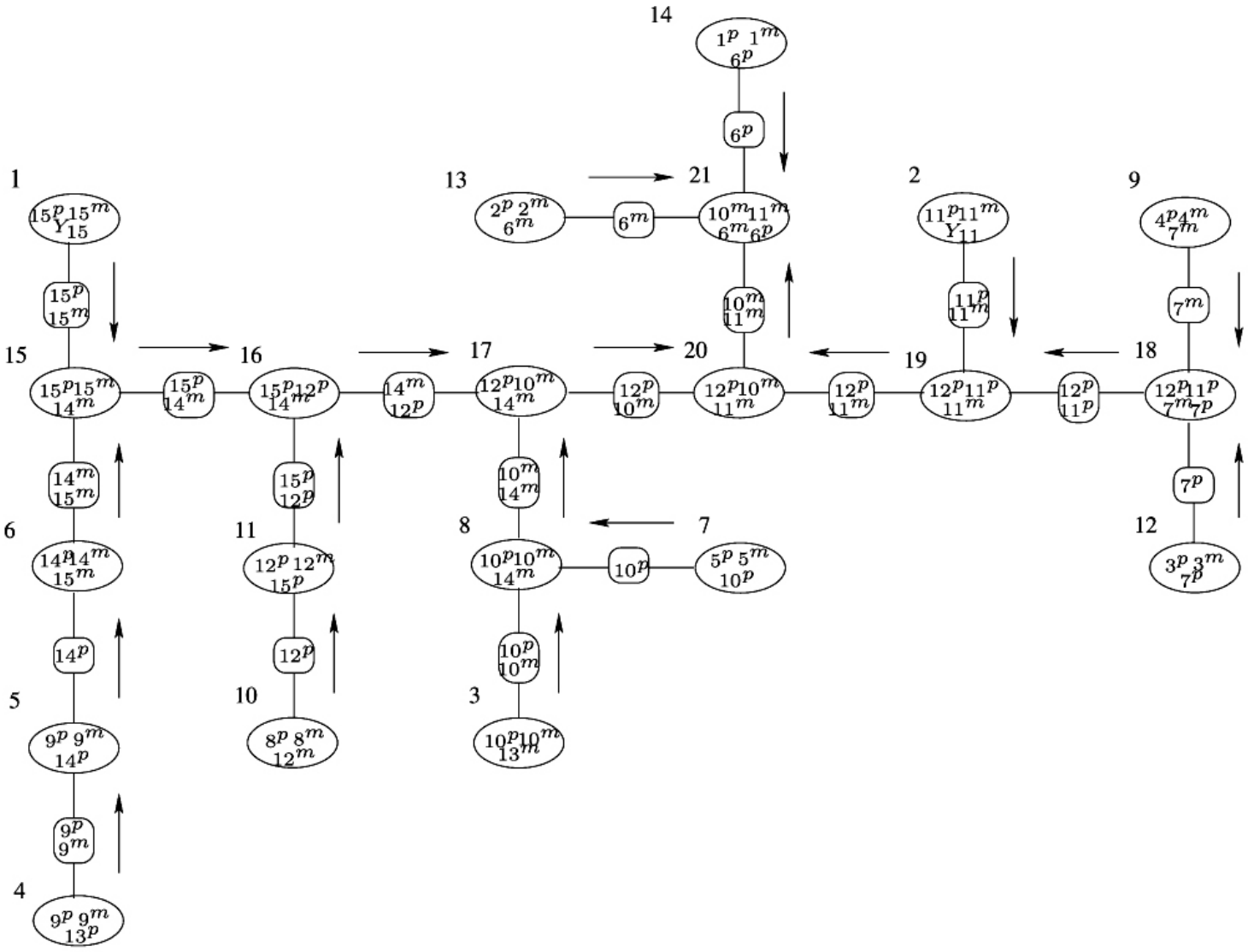
Assignments of potentials

Number	Elements	Assignments	Potential
1	$15^p, 15^m, Y_{15}$	Y_{15}	$f(Y_{15}\|15^p, 15^m)$
2	$11^p, 11^m, Y_{11}$	Y_{11}	$f(Y_{11}\|11^p, 11^m)$
3	$10^p, 10^m, 13^m$	13^m	$f(13^m\|10^p, 10^m)$
4	$9^p, 9^m, 13^p$	13^p	$f(13^p\|9^p, 9^m)$
5	$9^p, 9^m, 14^p$	$14^p, 9^m, 9^p$	$f(14^p\|9^p, 9^m) f(9^p) f(9^m)$
6	$14^p, 14^m, 15^m$	15^m	$f(15^m\|14^p, 14^m)$
7	$5^p, 5^m, 10^p$	$10^p, 5^m, 5^p$	$f(10^m\|5^p, 5^m) f(5^p) f(5^m)$
8	$10^p, 10^m, 14^m$	14^m	$f(14^m\|10^p, 10^m)$
9	$4^p, 4^m, 7^m$	$7^m, 4^m, 4^p$	$f(7^m\|4^p, 4^m) f(4^p) f(4^m)$
10	$8^p, 8^m, 12^m$	$12^m, 8^m, 8^p$	$f(12^m\|8^p, 8^m) f(8^p) f(8^m)$
11	$12^p, 12^m, 15^p$	15^p	$f(15^p\|12^p, 12^m)$
12	$3^p, 3^m, 7^p$	$7^p, 3^m, 3^p$	$f(7^p\|3^p, 3^m) f(3^p) f(3^m)$
13	$2^p, 2^m, 6^m$	$6^m, 2^m, 2^p$	$f(6^m\|2^p, 2^m) f(2^p) f(2^m)$
14	$1^p, 1^m, 6^p$	$6^p, 1^m, 1^p$	$f(6^p\|1^p, 1^m) f(1^p) f(1^m)$
15	$15^p, 15^m, 14^m$		1
16	$15^p, 12^p, 14^m$		1
17	$12^p, 10^m, 14^m$		1
18	$12^p, 11^p, 7^m, 7^p$	$11^p, 12^p$	$f(11^p\|7^p, 7^m) f(12^p\|7^p, 7^m)$
19	$12^p, 11^p, 11^m$		1
20	$12^p, 10^m, 11^m$		1
21	$10^m, 11^m, 6^m, 6^p$	$11^m, 10^m$	$f(11^m\|6^p, 6^m) f(10^m\|6^p, 6^m)$

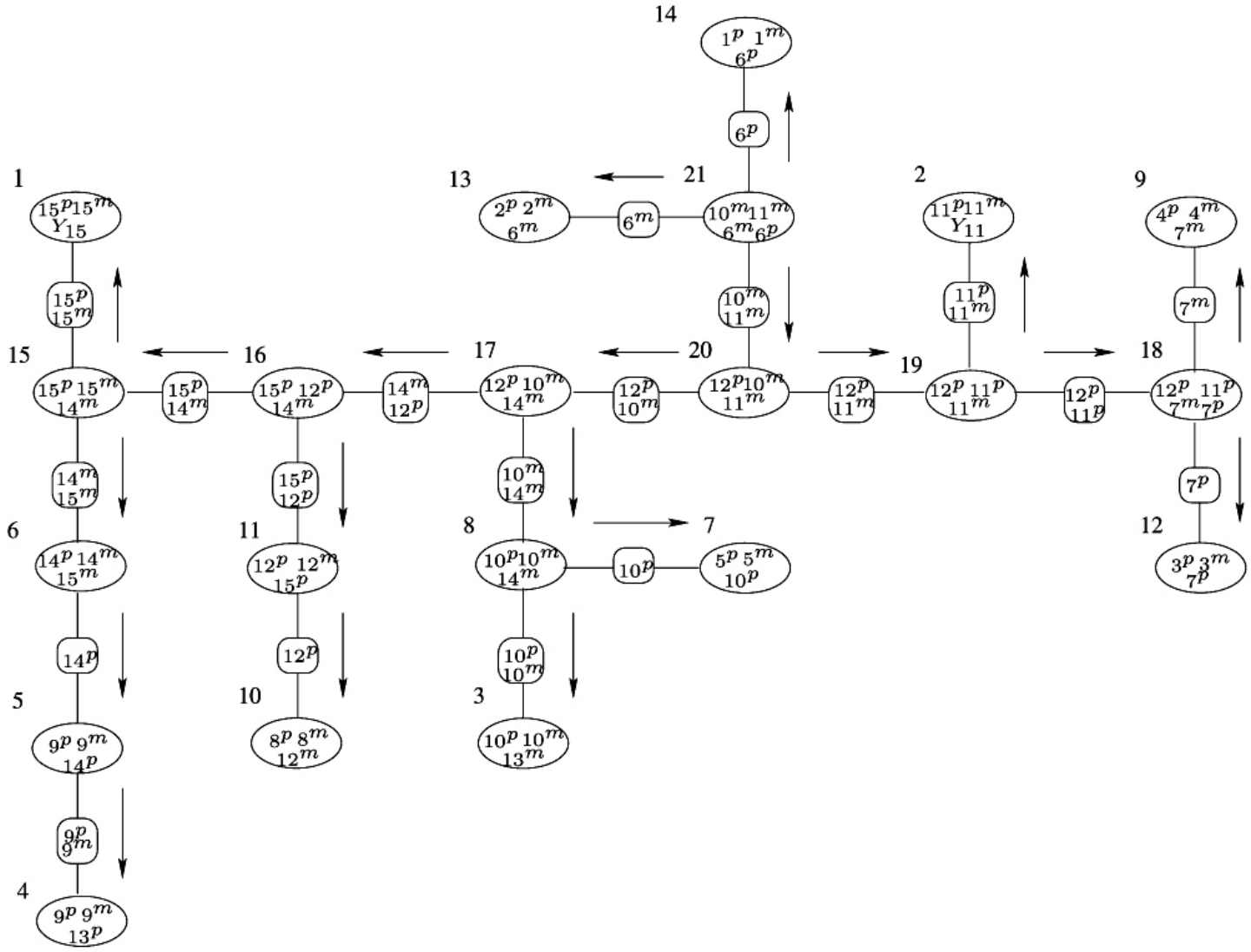
Propagation in the junction tree



Collect evidence



Distribute evidence



Algorithme de propagation dans l'arbre de jointure

Théorème

$$p(V_i, \mathbf{e}) = \prod \Phi_i \prod \psi^{sep(V_i)} \prod_{V_j \in \Xi(V_i)} \prod \psi_{sep(V_j)}$$

$$p(S_k, \mathbf{e}) = \prod \psi_k \prod \psi^k$$

Remarque : choix du plus petit ensemble V_i ou S_k contenant X pour calculer $p(X, \mathbf{e})$. Choix du plus petit S_k pour calculer $p(\mathbf{e})$.

Complexité

Temps $O(n^2 d^{w+1})$, espace $O(nsd^s)$ avec
$n = |\mathbf{A}|, d = max_j dom(A_j), w = max_i |V_i| - 1, s = max_k |S_k|$

Remarque : w est appelé largeur d'arbre (tree-width).
Dans le cas d'un polyarbre, $w = r, s = 1$, avec $r = max_j |\Pi(A_j)|$.

Graphical model formalisms

Graphical Model (GM)

A GM is a triplet $(\mathcal{X}, \mathcal{D}, \mathcal{P})$.

- $\mathcal{X} = \{X_1, \ldots, X_n\}$ a set of variables,
- $\mathcal{D} = \{D_{X_1}, \ldots, D_{X_n}\}$ a set of finite domains
- $\mathcal{F} = \{f_1, \ldots, f_e\}$, a set of nonnegative functions, each defined over a subset of variables $\mathbf{S}_i \subseteq \mathcal{X}$ (i.e. the scope)

Probabilistic

joint distribution :

$$\mathbb{P}(\mathcal{X}) \propto \prod_{i=1}^{e} f_i(\mathbf{S}_i)$$

Non probabilistic : WCSP

$$\texttt{score} = \sum_{i=1}^{e} f_i(\mathbf{S}_i)$$

Combinatorial Optimization

Probabilistic

The *Most Probable Explanation* (MPE) problem is to find the most likely assignment to all variables in $\mathcal{X}$ maximizing $\mathbb{P}(\mathcal{X})$.

Non probabilistic : WCSP

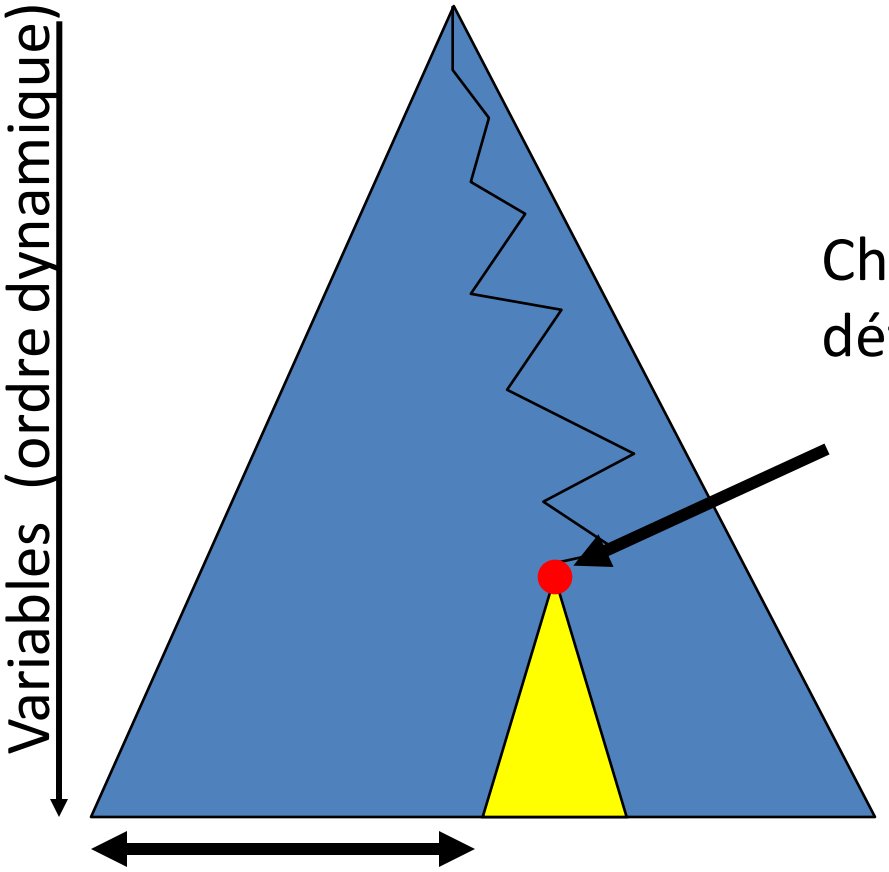
To find a feasible assignment of $\mathcal{X}$ minimizing `score`.

Equivalence

$$\underset{\mathcal{X}}{\operatorname{argmax}}\, \mathbb{P}(\mathcal{X}) = \underset{\mathcal{X}}{\operatorname{argmin}} - \sum_{i=1}^{e} log(f_i(\mathbf{S}_i))$$

- Inference (Variable Elimination : VE, Join Tree)
- Depth First Branch and Bound
- Hybrids
  - DFBB + local reasonning : bounded VE + soft local consistencies

Algorithme de séparation et évaluation



Chaque nœud de l'arbre est un WCSP défini par l'affectation courante

(LB) Minorant = $\mathbf{f_\varnothing}$

= sous-estimation du coût de la meilleure solution dans le sous-arbre courant

si $f_\varnothing \geq k$ alors couper

$\alpha +_k \beta = \min \{k, \alpha+\beta\}$

Local operators

Projection : $f[\mathbf{S}']$

$\mathbf{S}' \subseteq \mathbf{S}$ and $\forall t' \in D_{\mathbf{S}'} f[\mathbf{S}'](t') = \min_{t \in D_{\mathbf{S}} \text{ s.t. } t[\mathbf{S}']=t'} f(t)$

X	Y	f
0	0	5
0	1	7
1	0	6
1	1	4

X	$f[X]$
0	5
1	4

Join : $f = f_1 + f_2$

$f(t) = f_1(t[\mathbf{S}_1]) + f_2(t[\mathbf{S}_2]), \forall t \in D_{\mathbf{S}_1 \cup \mathbf{S}_2}$

i-bounded Variable Elimination (VE(i)) :
X such that $|\cap_{f_j : X \in S_j} S_j| < i$ $(\sum_{f_j : X \in S_j} f_j)$ $[\cap_{f_j : X \in S_j} S_j \; X]$

Local operators

Projection : $f[\mathbf{S}']$

$\mathbf{S}' \subseteq \mathbf{S}$ and $\forall t' \in D_{\mathbf{S}'} f[\mathbf{S}'](t') = \min_{t \in D_{\mathbf{S}} \text{ s.t. } t[\mathbf{S}']=t'} f(t)$

X	Y	f
0	0	5
0	1	7
1	0	6
1	1	4

X	$f[X]$
0	5
1	4

Join : $f = f_1 + f_2$

$f(t) = f_1(t[\mathbf{S}_1]) + f_2(t[\mathbf{S}_2]), \forall t \in D_{\mathbf{S}_1 \cup \mathbf{S}_2}$

Subtraction : $f = f_1 - f_2$

$\mathbf{S}_2 \subseteq \mathbf{S}_1$ and $f(t) = f_1(t) - f_2(t[\mathbf{S}_2]), \forall t \in D_{\mathbf{S}_1}$

soft Directed Arc Consistency (DAC) [Cooper, *Fuzzy Set Sys.*, 2003] :

$f_1(X) \leftarrow f_1(X) + (f(X,Y) + f_2(Y))[X]$ X < Y

$f(X,Y) \leftarrow (f(X,Y) + f_2(Y)) - (f(X,Y) + f_2(Y))[X]$ $f(Y) \leftarrow 0$

(de Givry et al, Constraints 2008)

MENDEL

- toulbar2 v0.5 avec last conflict
- Temps CPU en secondes pour trouver et prouver l'optimum

	ind	*vars*	*genotyped*	alleles	*nf*	*ngen*	treewidth ub	DFBB-VE(2) errors	time	nodes
eye	36	36	28	6	11	4	2	1	0.02	0
cancer	49	48	37	8	18	5	2	1	0.21	0
parkinson	37	34	13	4	7	7	5	0	0	6
$berrichon_{1nc}$	129516	9947	2448	4	8821	17	262	2	4.73	8805
$berrichon_{1}$	129516	10017	2483	4	8786	17	330	23	5.81	8384
$berrichon_{2nc}$	27255	19337	10215	4	4719	19	-	41	5.89	6170
$berrichon_{2}$	27255	19562	10215	4	2381	19	-	106	17.23	15445
$langlade_{1}$	1355	1209	711	9	298	13	84	38	12.28	391
$langlade_{2}$	1355	1223	715	7	298	13	82	89	60.56	17857
$langlade_{3}$	1355	1258	787	5	298	13	85	39	14.19	6731
$langlade_{4}$	1355	1186	672	8	298	13	83	43	59.7	3520
$moissac_{1}$	283	260	183	2	81	5	6	0	0	5
$moissac_{2}$	283	244	167	7	81	5	6	0	0.51	6
$moissac_{3}$	283	225	151	3	81	5	6	0	0	4
$moissac_{4}$	283	256	179	2	81	5	6	0	0	5
$moissac_{5}$	283	237	161	8	81	5	6	0	1.02	5
$moissac_{6}$	283	201	131	11	81	5	5	0	5.64	6

Reconstruction d'haplotypes dans des pedigrees *en arbre* (Favier et al, IJCAI 2011)

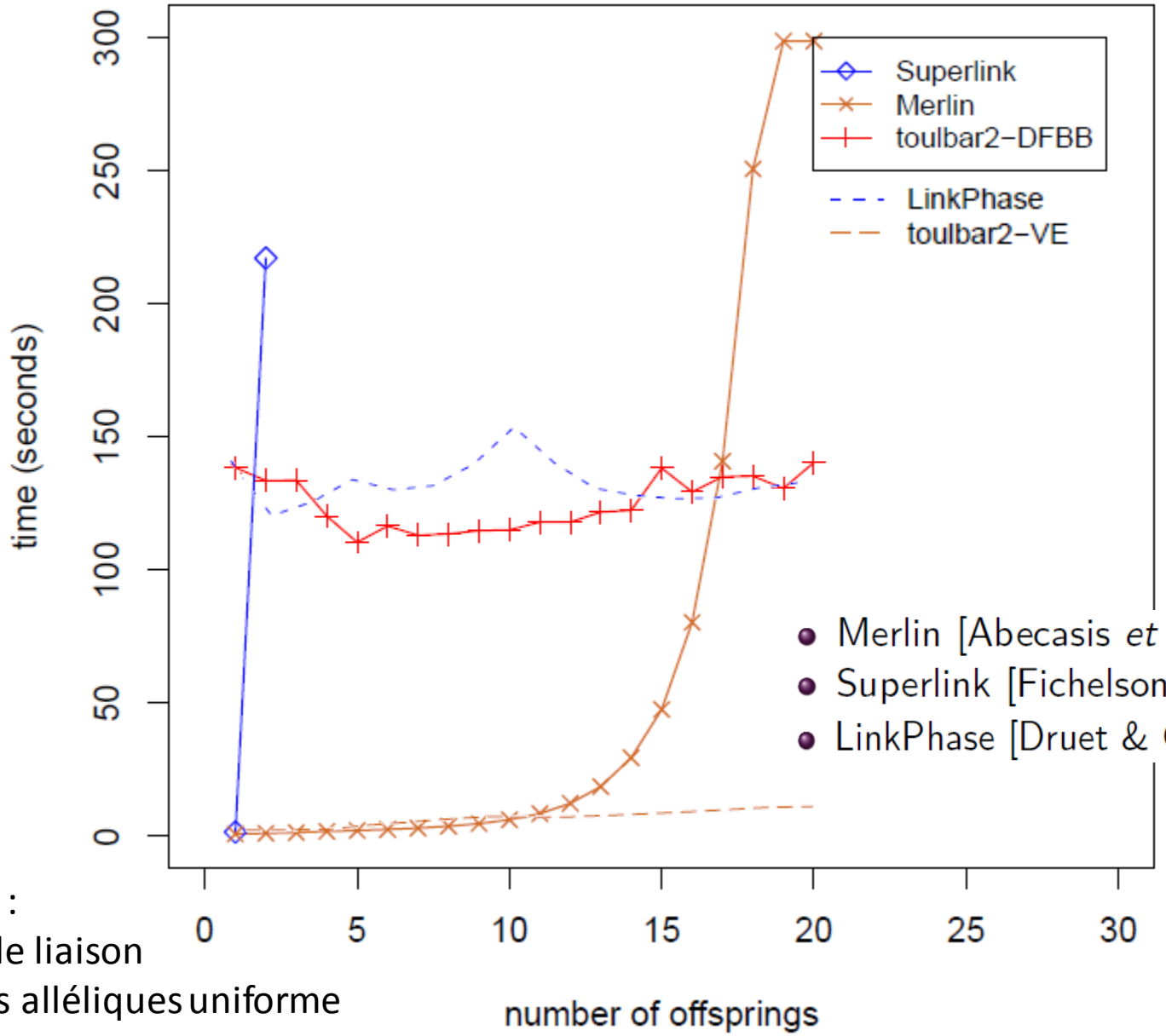
Indiv. x Loci	Problème	DFBB-VE(i)		DFBB-VE(i) +dec+ps		AOBB-C+SMB(j)+VE(i)			AOBB-C+SMB(j)+VE(i) +dec+ps
		time (s)	i	time (s)	i	j	i	time (s)	time (s)
	Linkage								
25x20	ped7	4.04	2	1.18	4	20	4	1915.56	131.14
	ped9	-		3.36	6	20	6	-	104.62
	ped18	149.71	4	3.19	5	20	5	54.67	18.82
	ped19	-			-	20	6	-	-
	ped20	3.46	4	0.39	6	16	6	407.6	66.53
	ped23	0.09	4	0.05	3	12	3	6.05	1.52
	ped25	1207.97	4	0.65	6	20	6	25.91	32.51
	ped30	543.20	2	5.44	5	20	5	28.39	10.10
	ped34	1.13	2	0.36	5	20	5	-	24.56
57x6	ped37	0.21	5	0.11	4	10	4	87.18	9.58
	ped39	16.68	4	0.24	5	18	5	6.87	2.60
	ped41	-		302.05	4	20	4	-	1271.31
	ped42	1.94	4	0.34	6	16	6	240.94	155.39
20x20	ped44	-		505.46	5	20	5	2631.71	333.89
	ped50	0.90	4	0.18	4	12	4	316.92	521.92

Hypothèse : équilibre de liaison

Marqueurs multiallèliques (≤ 9 allèles par marqueur)

Pedigree de demi-frères (Favier et al, WCB'10)

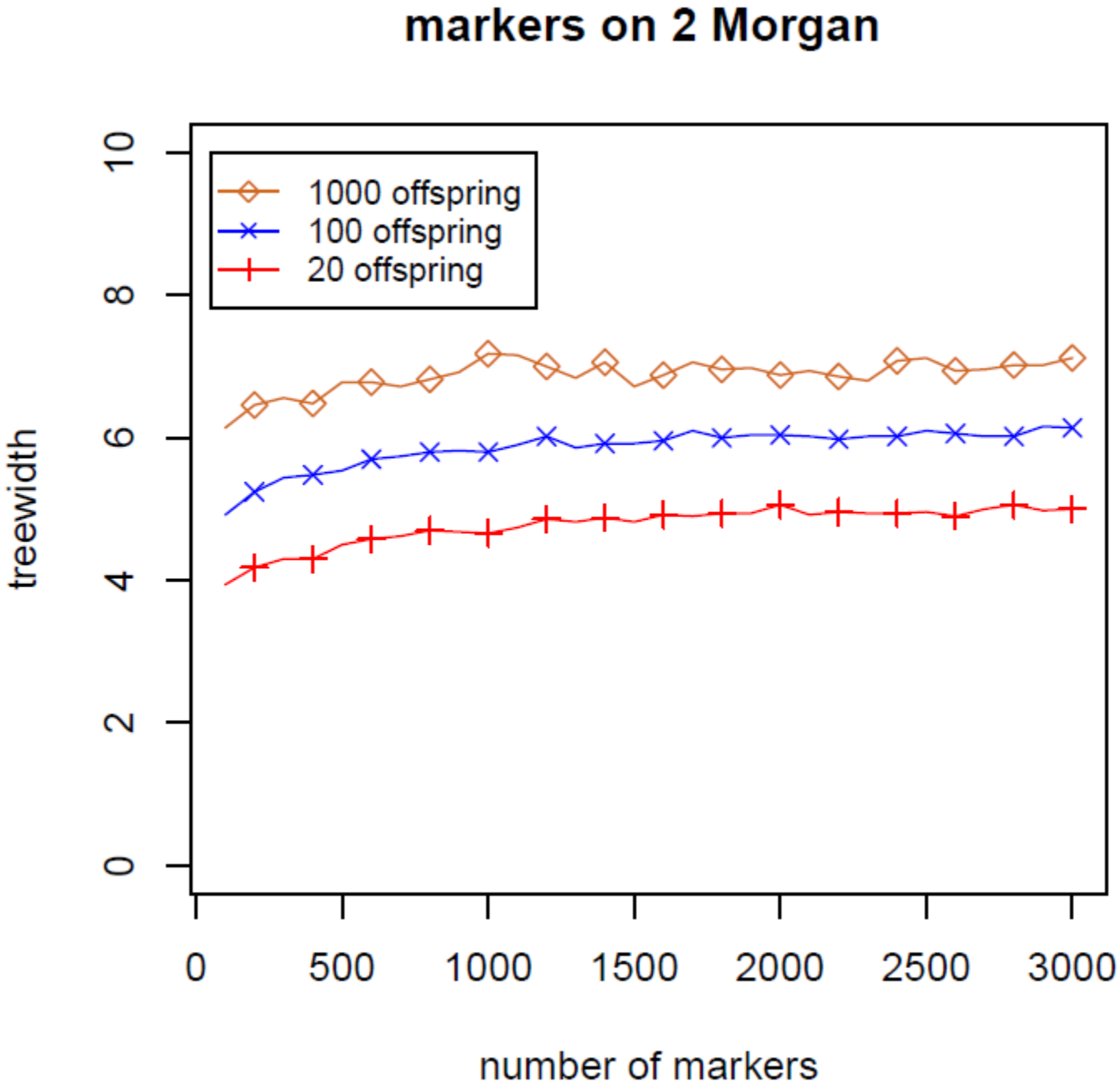
Chromosome X humain, 36000 marqueurs SNP sur 1,64 Morgan(hapmap)



- Merlin [Abecasis *et al.*(2002)]
- Superlink [Fichelson *et al.*(2005)]
- LinkPhase [Druet & Georges(2010)]

Hypothèses :

- équilibre de liaison
- fréquences alléliques uniforme


markers on 2 Morgan
1000 offspring
100 offspring
20 offspring
treewidth
0
2
4
6
8
10
0
500
1000
1500
2000
2500
3000
number of markers