Un algorithme générique d'extraction de bi-ensembles sous contraintes

Jérémy Besson (LIRIS, INSA Lyon),
Céline Robardet (Prisma, INSA Lyon),
Jean-François Boulicaut (LIRIS, INSA Lyon)

Contexte

- Deux laboratoires
  - ✓ INSA Lyon-LIRIS UMR 5205 : Data Mining et bases de données inductives
  - ✓ UMR INSERM/INRA 1235 : Obésité et diabète de Type 2

- Objectifs
  - ✓ Aider à l'extraction de connaissances dans les données transcriptionnelles, mécanismes de régulation des gènes chez l'homme en réponse à l'insuline

Données transcriptionnelles

	G_1	G_2	G_3	G_4
CE_1	3.12	0.31	0.21	3.2
CE_2	0.53	1.01	0.01	1.58
CE_3	0.25	1.05	0.80	2.08
CE_4	2.01	1.02	4.6	2.9
CE_5	2.1	1.5	0.8	0.5

	G_1	G_2	G_3	G_4
FT_1	1	1	0	1
FT_2	1	1	0	1
FT_3	1	0	1	1
FT_4	0	1	1	1

	CE_1	CE_2	CE_3	CE_4	CE_5
Inf_1	oui	non	non	oui	Non
Inf_2	12	25	35	?	11
Inf_3	2h	1h10	1h32	2h	1h12

```
>101174:MAPT
CCTTGGTCAGTAACAAAAAAGGTGGGAAAAAA
CCGGCCACGTGGAAGGCCGCTCAGGACTTCTG
AAGGAGGACACCCACCCCCACAACGACACAAA
GGGACCGCGAAAGGGCAGCGCCGAGAGGAACG
GCCCCC
>10283:RBM8A
TTTCCTGAAGAGAAATTTGGTGCTGCAGGTTC
CCTGGGATTTTTGAAAGAAAAGGGTTTCTTCA
TCAGCAGCTGCCGCGGTTAAGAGGAAGCCACG
AGGCCCGCCCCAATTTGCGTGTTTTTACCGTG
AGCGGC
```

Fonctions *GO*

Questions

- Quels sont les groupes de gènes qui varient significativement et simultanément dans des cellules musculaires humaines en réponse à l'insuline ?
- Quels sont les groupes de facteurs de transcription qui peuvent s'accrocher simultanément sur la région promotrice d'ensembles de gènes ?
- Quels sont les sites de fixation de FT qui sont très présents sur les séquences promotrices de gènes ayant la même fonction ?
- …

Analyse du transcriptome

- Difficultés
  - ✓ Grands volumes de données hétérogènes
  - ✓ Données manquantes et erronées
- Approches « traditionnelles »
  - ✓ Extractions heuristiques et incomplètes
    - – Motifs globaux (partitions)
    - – Motifs locaux (bi-partitions)

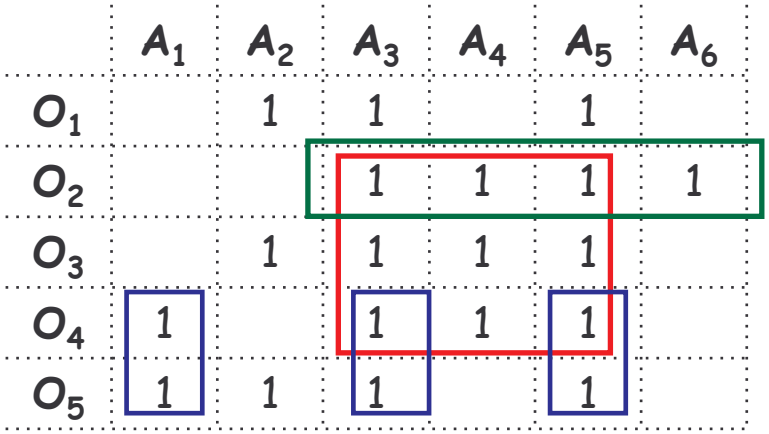
Analyse du transcriptome

- Notre approche

  - ✓ Rechercher tous les motifs locaux ensemblistes qui satisfont une certaine contrainte dans une relation binaire
  - ✓ Justesse et complétude des extractions
  - ✓ Utilisation active des contraintes

Question : Motifs + contraintes

Concepts formels



	A_1	A_2	A_3	A_4	A_5	A_6
O_1		1	1		1	
O_2			1	1	1	1
O_3		1	1	1	1	
O_4	1		1	1	1	
O_5	1	1	1		1	

« 1 » : un gène (A_i) a une variation d'expression significative dans une condition expérimentale (O_i)

« 1 » : le facteur de transcription (O_i) s'accroche sur la région promotrice d'un gène (A_i)

Concepts formels

$\mathcal{A}$

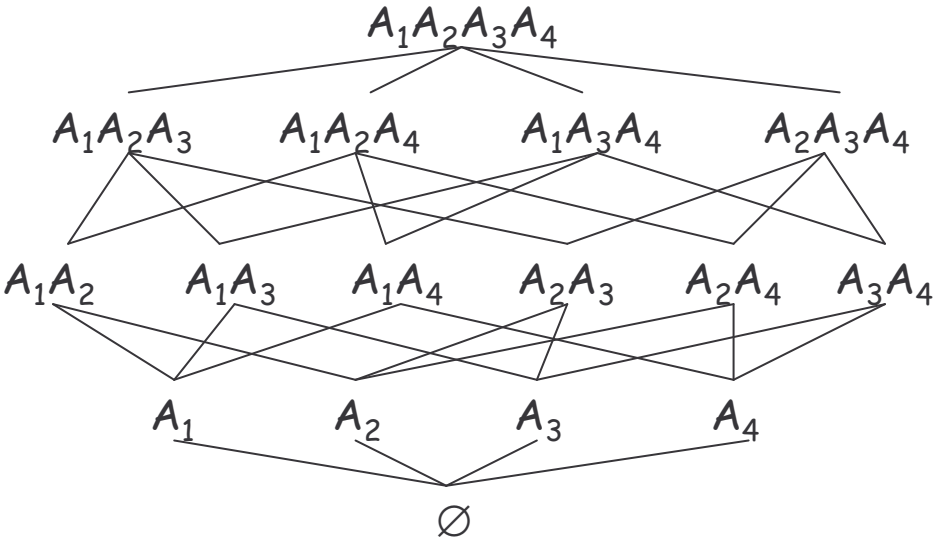
$\mathcal{O}$	A_1	A_2	A_3	A_4
O_1	1	1		
O_2	1	1	1	
O_3		1	1	1

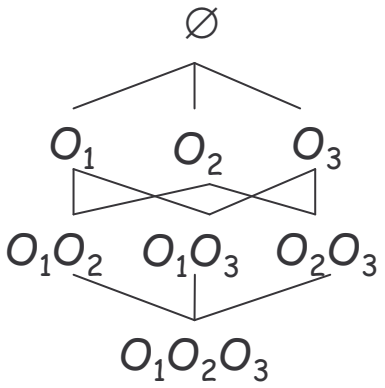
- Connection de Galois
  $f(X) = \{y \in \mathcal{A} \mid \forall x \in X,\ (x,y) \in r\}$
  $g(Y) = \{x \in \mathcal{O} \mid \forall y \in Y,\ (x,y) \in r\}$

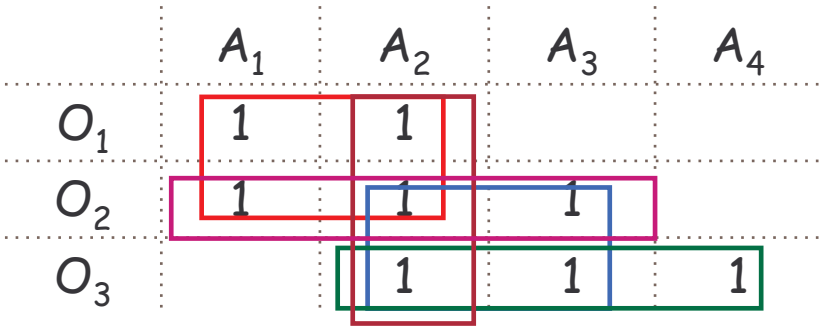
- (X,Y) est un concept formel ssi
  $X = g(Y) \wedge Y = f(X)$

- f et g sont des fonctions décroissantes : $A \subseteq B \Rightarrow f(B) \subseteq f(A)$

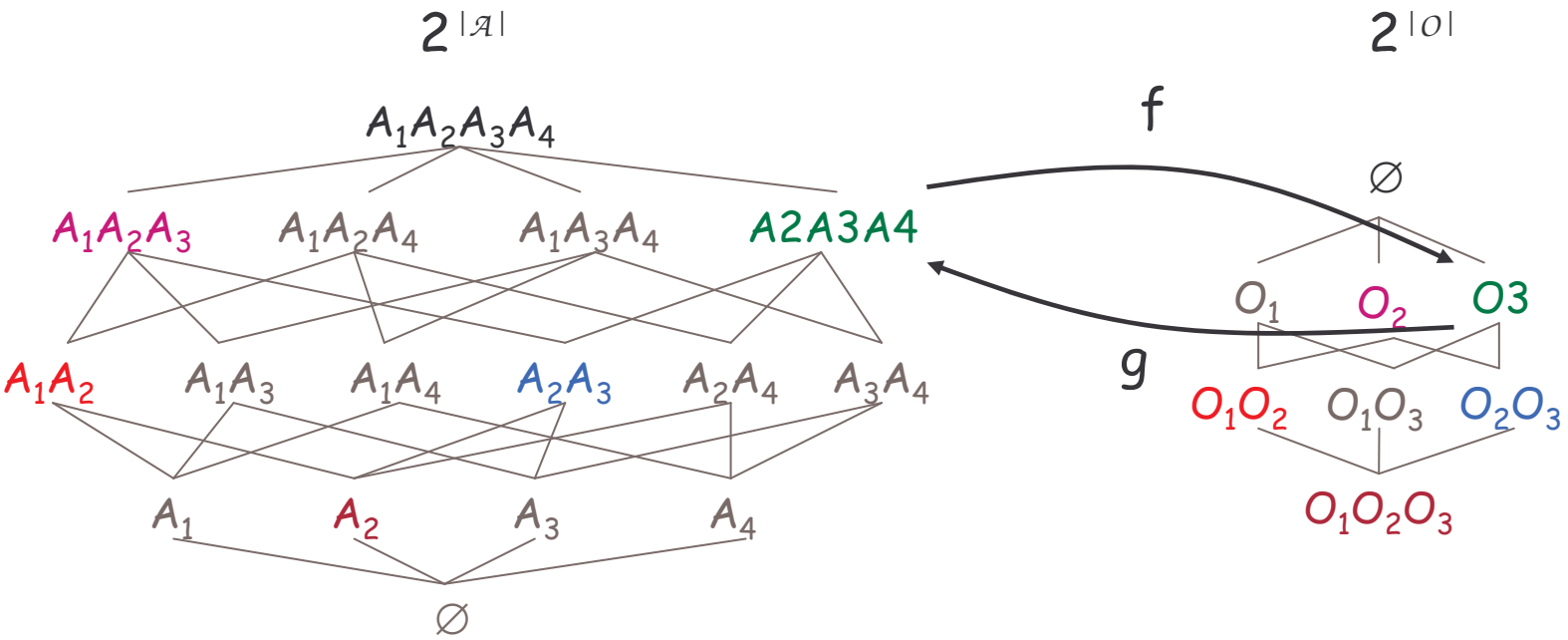
Treillis des attributs et des objets

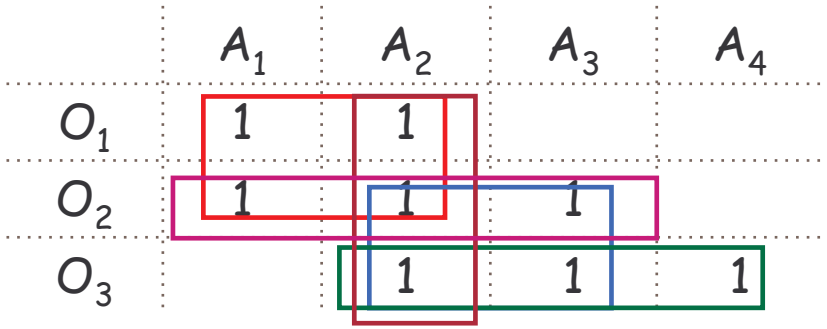






Exemple


$2^{|\mathcal{A}|}$
$2^{|O|}$
$A_1A_2A_3A_4$
$A_1A_2A_3$
$A_1A_2A_4$
$A_1A_3A_4$
A2A3A4
A_1A_2
A_1A_3
A_1A_4
A_2A_3
A_2A_4
A_3A_4
A_1
A_2
A_3
A_4
∅
f
g
∅
O_1
O_2
O3
O_1O_2
O_1O_3
O_2O_3
$O_1O_2O_3$


A_1
A_2
A_3
A_4
O_1
O_2
O_3
1
1
1
1
1
1
1
1
1

Contraintes (anti-)monotones

- Soient C une contrainte sur E et $\subseteq$ une relation d'ordre sur E alors

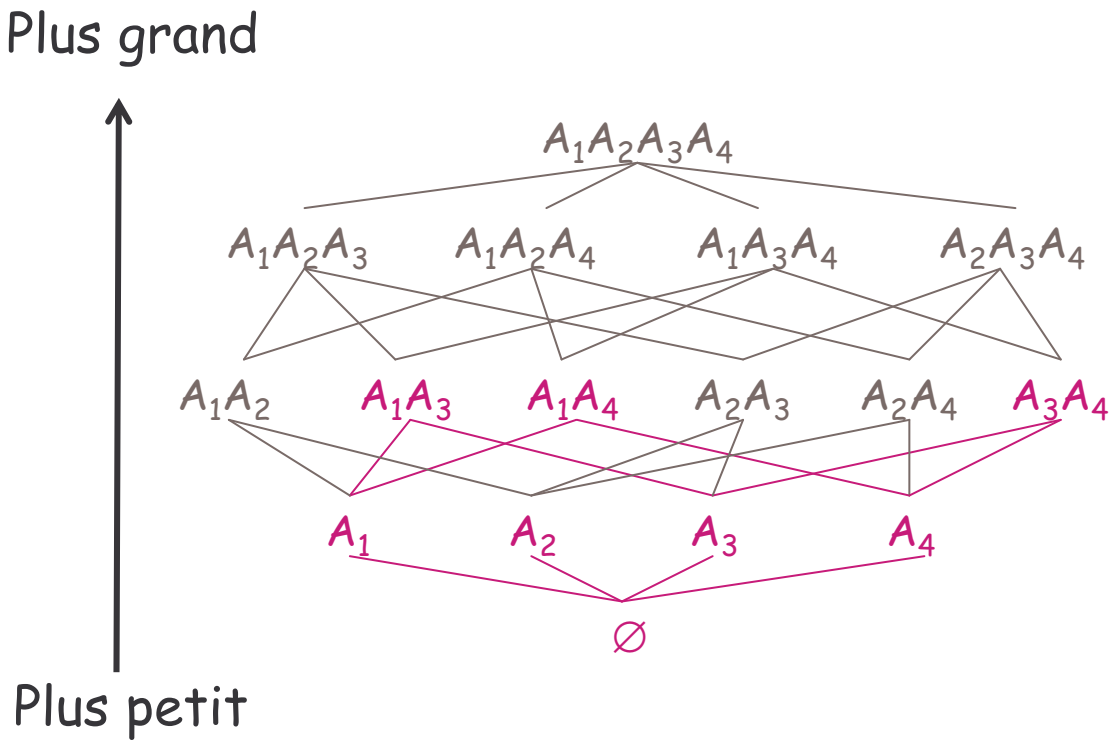
  - ✓ C est monotone sur $(\subseteq, E)$ ssi

$$\forall A,B \subseteq E \text{ tq } A \subseteq B,\ C(A) \Rightarrow C(B)$$
$$\neg C(B) \Rightarrow \neg\, C(A)$$

  - ✓ C est anti-monotone sur $(\subseteq, E)$ ssi

$$\forall A,B \subseteq E \text{ tq } A \subseteq B,\ C(B) \Rightarrow C(A)$$
$$\neg C(A) \Rightarrow \neg\, C(B)$$

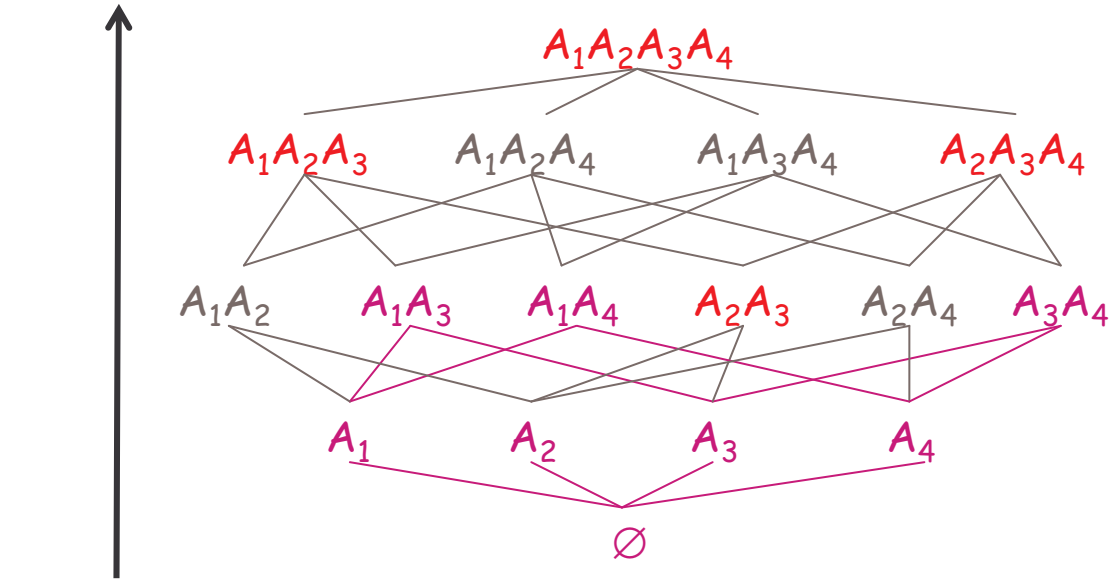
Contraintes monotones - Exemples



$$|E| < 3 \wedge A_2 \notin E$$

Contraintes monotones - Exemples

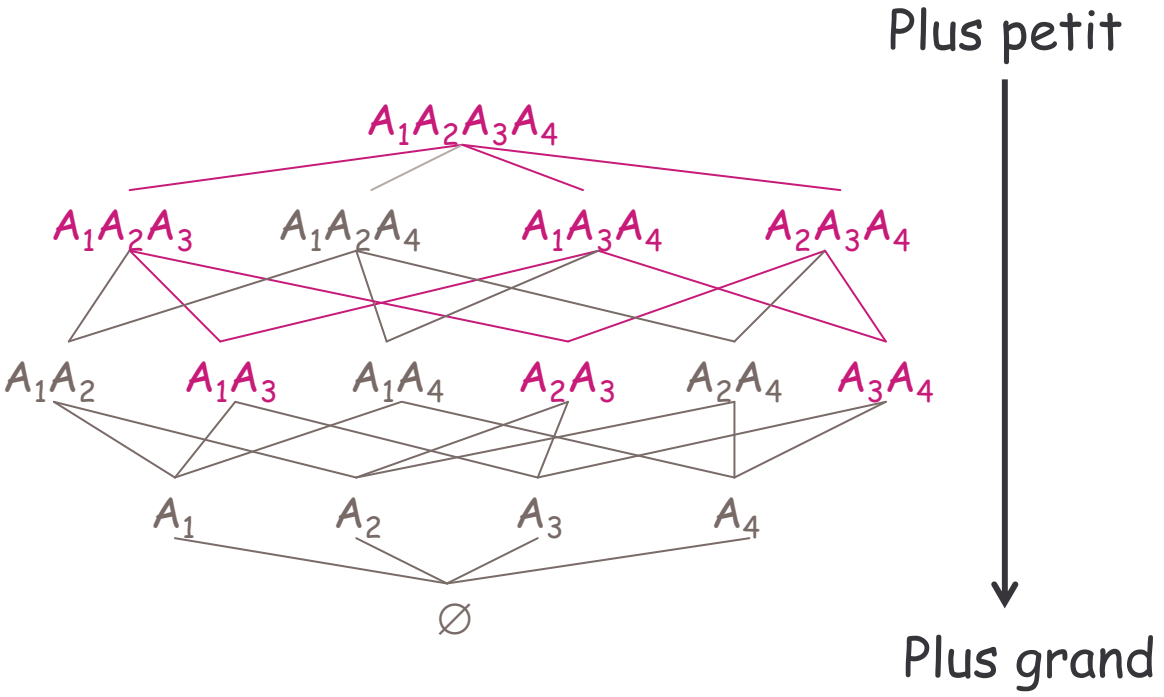
Plus grand



Plus petit

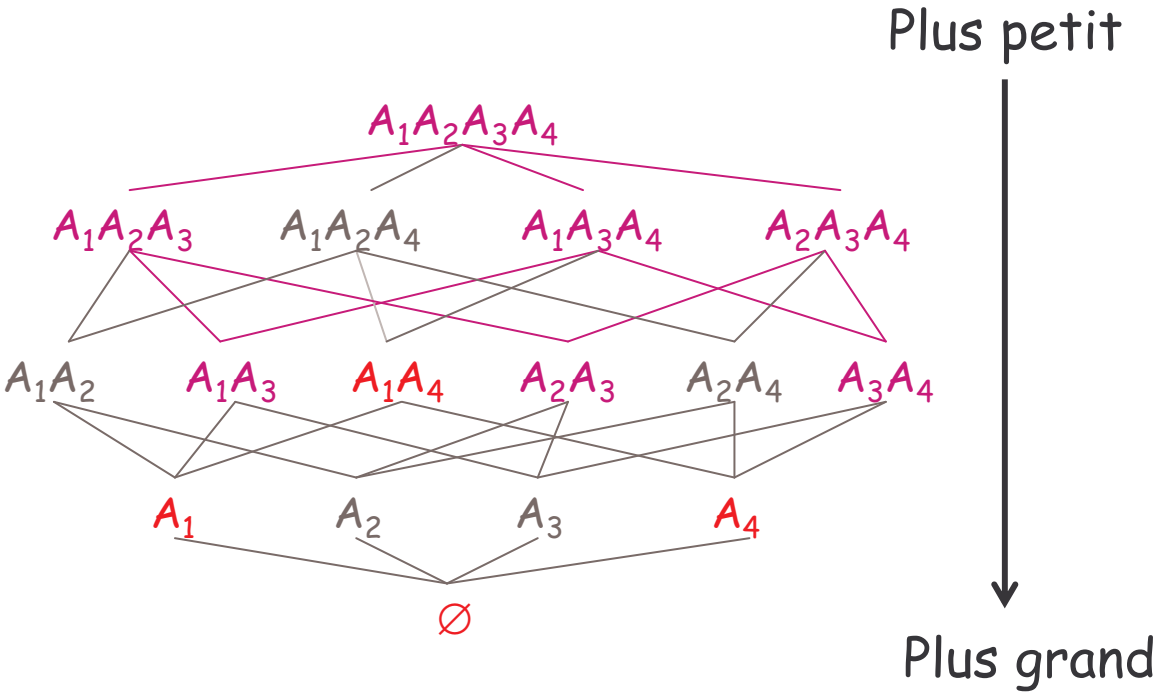
$$|E| < 3 \wedge A_2 \notin E$$

Contraintes monotones - Exemples



$|E| > 1 \wedge A_3 \in E$

Contraintes monotones - Exemples



$$|E| > 1 \wedge A_3 \in E$$

Propriétés

- Soit C une contrainte monotone sur $(E,\subseteq)$ alors
  - ✓ $\neg C$ est anti-monotone sur $(E,\subseteq)$
  - ✓ C est anti-monotone sur $(E,\supseteq)$
- Une conjonction et disjonction de contraintes monotones est monotone

Extraction de concepts formels sous contraintes

- Enumérer les candidats suivant une relation d'ordre $\subseteq$
  - ✓ Vérifier et propager les contraintes monotones et anti-monotones suivant $\subseteq$ au cours de l'extraction
  - ✓ Utiliser les autres contraintes en post-traitement
- Algorithmes qui exploitent la connection de Galois
  - ✓ Ganter, Ac-miner, Charm, Closet, …
  - ✓ Enumération sur une des deux dimensions et génération par (f,g) des ensembles sur l'autre dimension.

Problèmes - Contraintes

Monotones sur		Anti-monotones sur		Autres	
$(\mathcal{O}, \subseteq)$	$(\mathcal{A}, \subseteq)$	$(\mathcal{O}, \subseteq)$	$(\mathcal{A}, \subseteq)$		
$\lvert X\rvert > \varepsilon$ $a \in X$	$\lvert Y\rvert > \varepsilon$ $a \in Y$	$\lvert X\rvert < \varepsilon$ $a \notin X$	$\lvert Y\rvert < \varepsilon$ $a \notin Y$	$\lvert X\rvert / \lvert Y\rvert > 2$ $\lvert Y\rvert / \lvert X\rvert < 4$	$\lvert X\rvert * \lvert Y\rvert > 10$

- C = {(X,Y) concepts tels que $|X| > 5 \wedge |Y| > 4$}
  - ✓ C1 = {(X,Y) concepts tels que $|X| > 5$ }
  - ✓ C2 = {(X,Y) $\in$ C1 tels que $|Y| > 4$ }

Extraction de bi-ensembles sous-contraintes

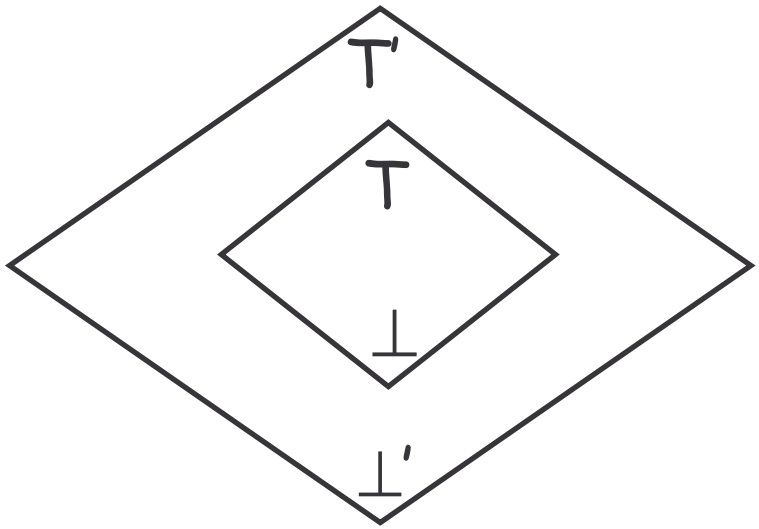
Approche générique

- Motifs
  - ✓ bi-ensembles
- Contraintes
  - ✓ Pour définir le type de motifs à extraire
  - ✓ Pour sélectionner les motifs pertinents pour l'utilisateur final

Bi-ensembles

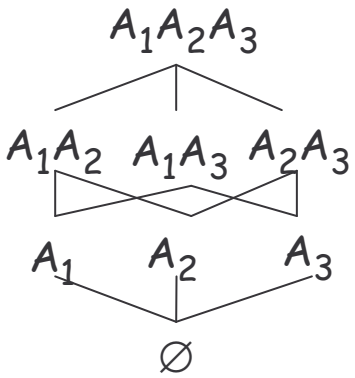
- Un espace de recherche sur les bi-ensembles
  - ✓ 2 treillis $([\bot_O, \top_O], [\bot_A, \top_A])$

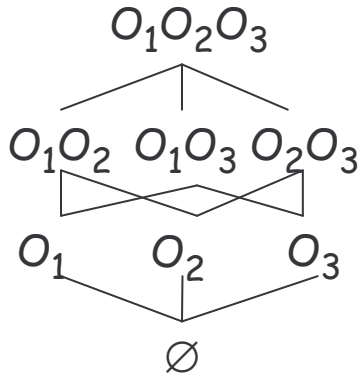
- $([\bot_O, \top_O], [\bot_A, \top_A]) \subseteq_{be} ([\bot_O', \top_O'], [\bot_A', \top_A'])$ ssi
  - ✓ $\bot_O' \subseteq \bot_O$
  - ✓ $\bot_A' \subseteq \bot_A$
  - ✓ $\top_O \subseteq \top_O'$
  - ✓ $\top_A \subseteq \top_A'$



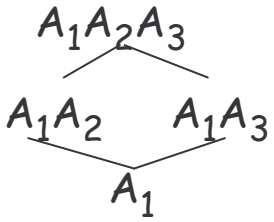
Bi-ensembles

- Exemple : $([\varnothing, A_1A_2A_3],[\varnothing, O_1O_2O_3])$





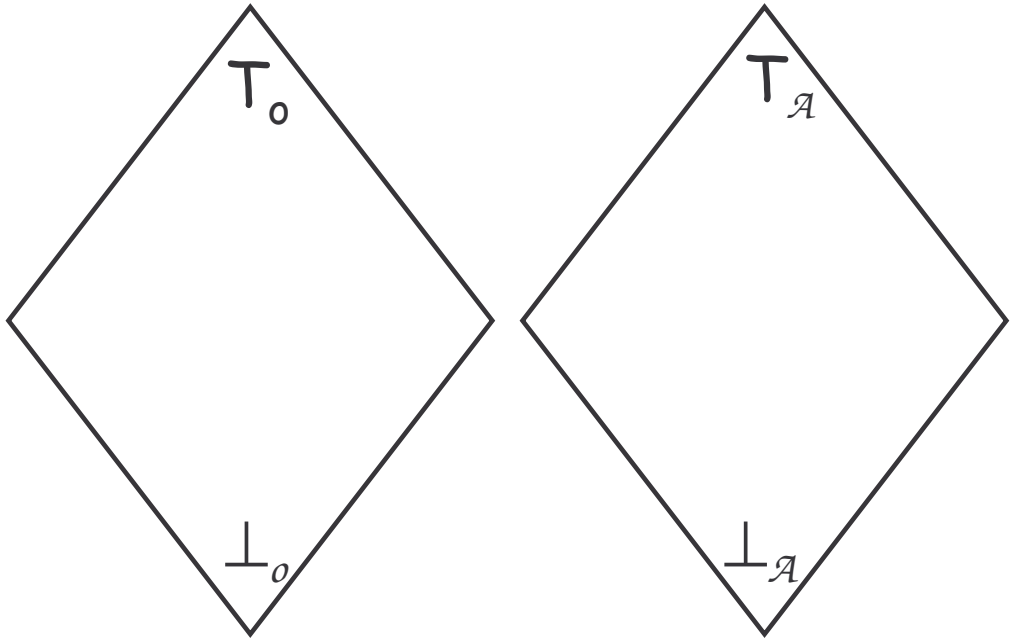
- Exemple : $([A_1, A_1A_2A_3],[O_2, O_2])$



O_2

- $([A_1, A_1A_2A_3],[O_2, O_2]) \subseteq_{be} ([\varnothing, A_1A_2A_3],[\varnothing, O_1O_2O_3])$

Contraintes



$|X| > \varepsilon \wedge |Y| > \varepsilon$

$|T_o| > \varepsilon \wedge |T_{\mathcal{A}}| > \varepsilon$

$b \notin Y$

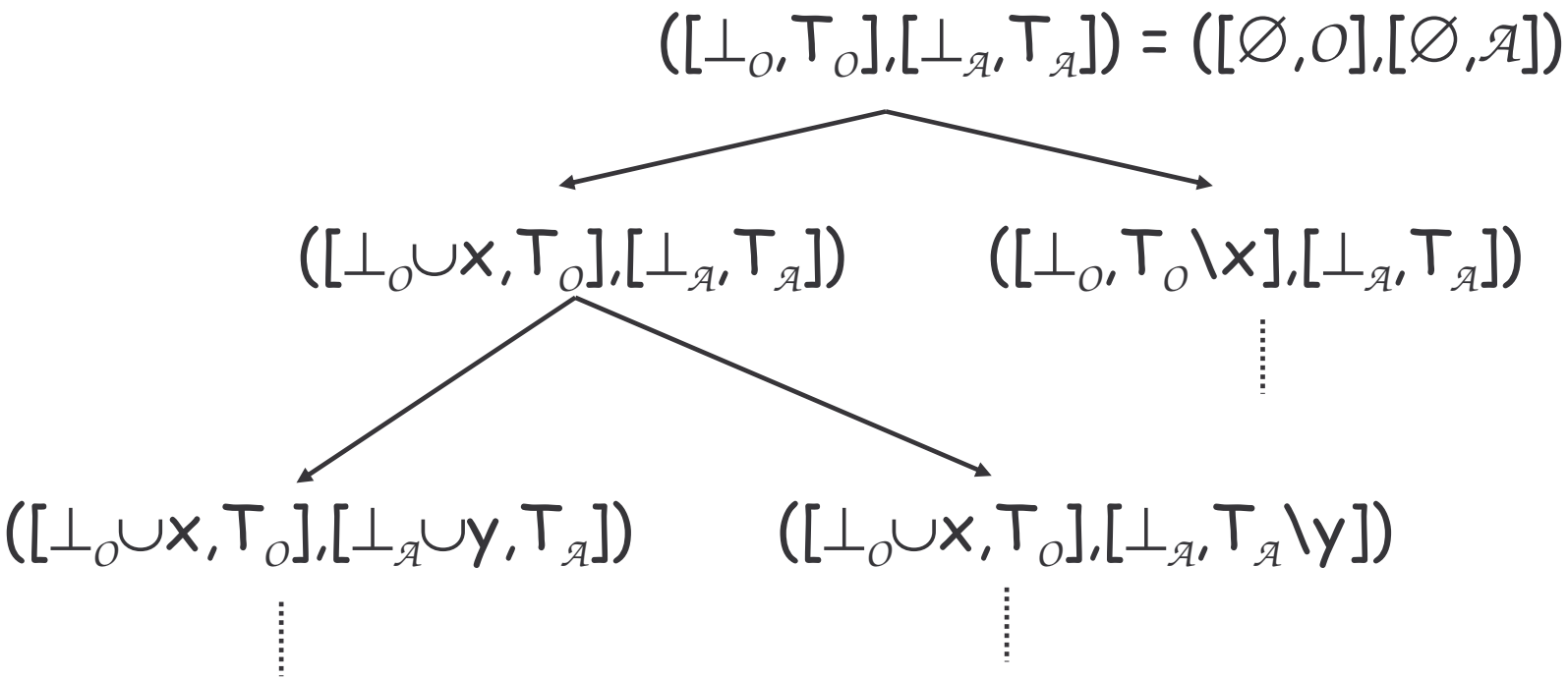
$b \notin \perp_{\mathcal{A}}$

$|X|*|Y| > \varepsilon$

$|T_o|*|T_{\mathcal{A}}| > \varepsilon$

$Moyval_o > \varepsilon \equiv \sum x \in T_o \; Val(x) \; / \; |\perp_o| > \varepsilon$

Enumération



Extraction de concepts formels sous contraintes :

D-Miner

Algorithme : D-Miner

- Extraire les concepts formels satisfaisants des contraintes monotones sur $(\mathcal{O},\subseteq)$, $(\mathcal{A},\subseteq)$ et $(\mathcal{O}\times\mathcal{A},\subseteq_{be})$

- Exemple :
  - ✓ {(X,Y) concepts tels que $|X| > 5 \wedge |Y| > 4$}
  - ✓ {(X,Y) concepts tels que $a \in X \vee |X| * |Y| > 10$ }

Extraction de concepts

	A_1	A_2	A_3	A_4
O_1	1	1		
O_2	1	1	1	
O_3		1	1	1

(X,Y) est un concept

- ✓ $O_1 \in X$ alors $A_3 \notin Y$ et $A_4 \notin Y$
- ✓ $A_1 \in Y$ alors $O_3 \notin X$

} 1-rectangle : C_{1r}

- ✓ $O_3 \notin X$ alors $A_1 \in Y$
- ✓ $A_4 \notin Y$ alors $O_1 \in X$ ou $O_2 \in X$

} Maximalité : C_{max}

Propriétés

- C_{1r} est une contrainte monotone suivant $\subseteq_{be}$
- $C_{max}(X,Y) \equiv$

  $\forall x \in \mathcal{O} \setminus X, \exists y \in Y$ tq $(x,y) \notin r$

  $\forall y \in \mathcal{A} \setminus Y, \exists x \in X$ tq $(x,y) \notin r$

- $C_{max}([\bot_O, T_O],[\bot_A, T_A]) \equiv$

  $\forall x \in \mathcal{O} \setminus T_O \; \exists y \in \bot_{\mathcal{A}}$ tq $(x,y) \notin r \wedge$

  $\forall y \in \mathcal{A} \setminus T_{\mathcal{A}} \; \exists x \in \bot_O$ tq $(x,y) \notin r$

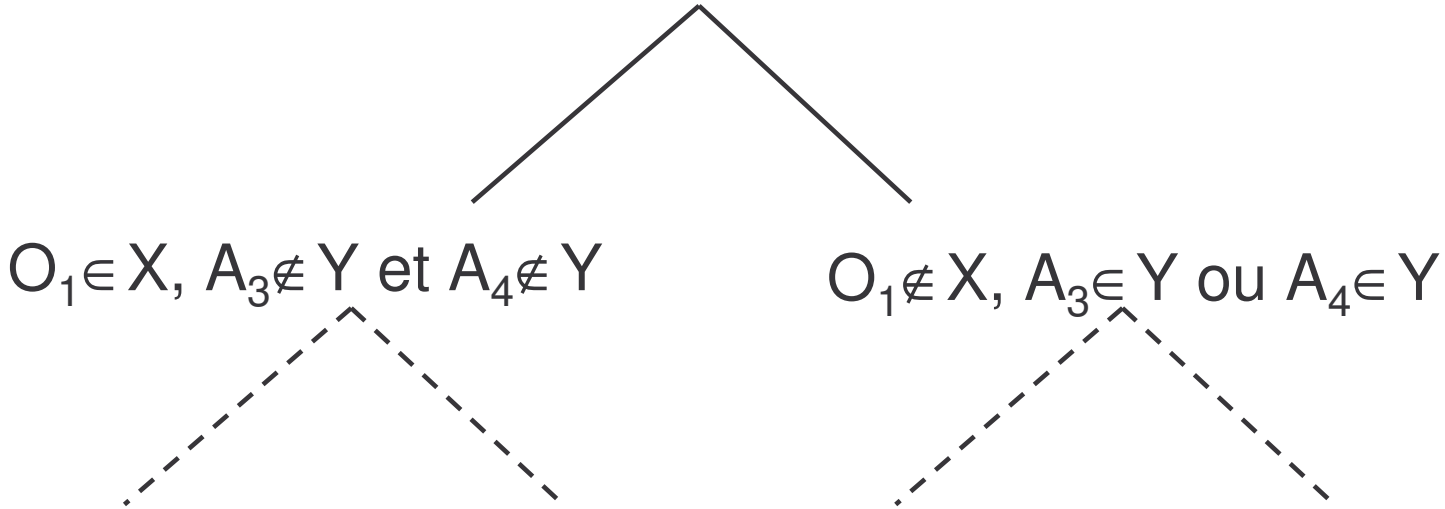
- Exemples
  - ✓ {(X,Y) tels que $C_{1r} \wedge C_{max} \wedge |X| > 5 \wedge |Y| > 4$}
  - ✓ {(X,Y) tels que $C_{1r} \wedge C_{max} \wedge (a \in X \vee |X| * |Y| > 10)$ }

D-Miner

- Contextes contenant beaucoup de concepts formels
- La contrainte C_{max} est coûteuse et peu efficace pour réduire l'espace de recherche.
- C_{max} est exploitée au minimum
  - ✓ Vérification de la consistance
  - ✓ Pas de propagation

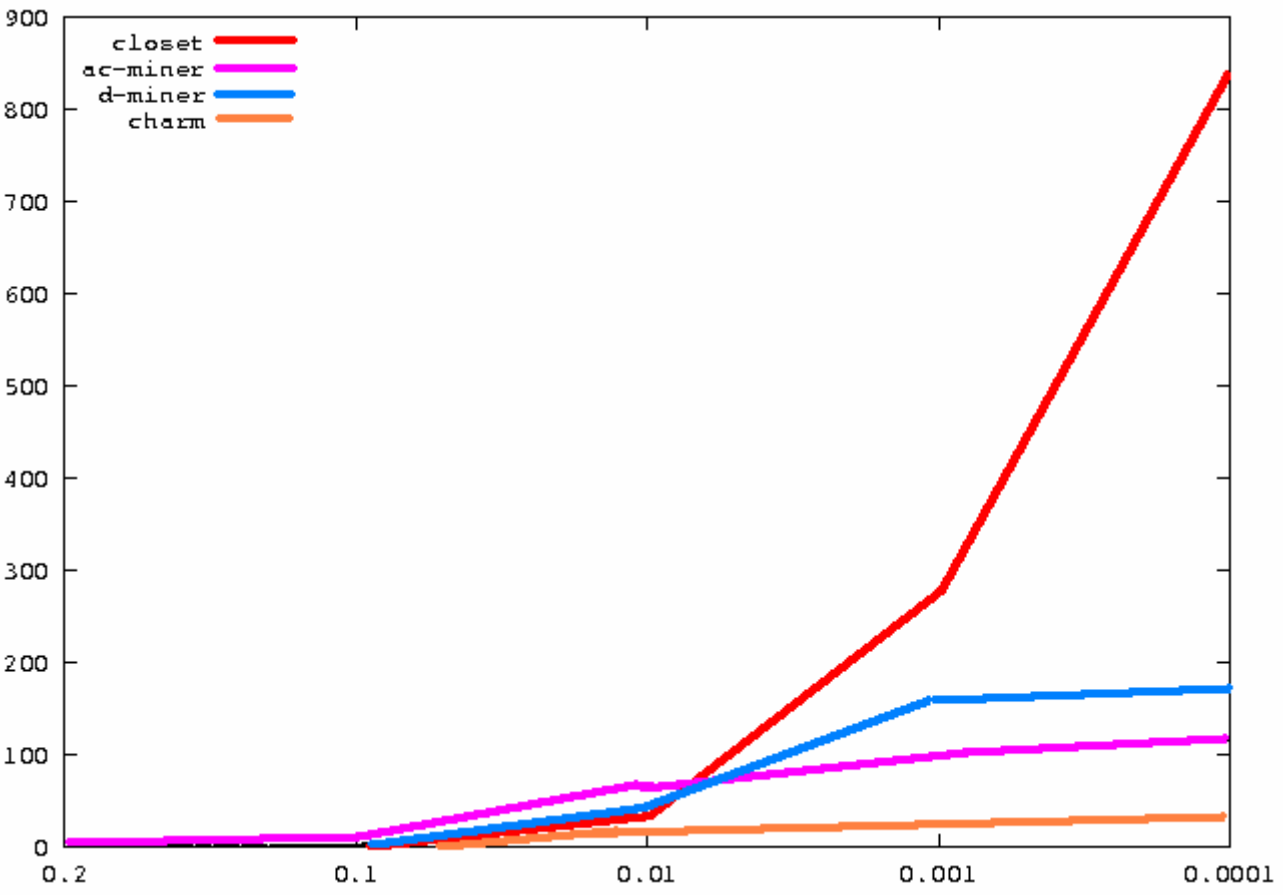
D-Miner

	A_1	A_2	A_3	A_4
O_1	1	1	0	0
O_2	1	1	1	0
O_3	0	1	1	1



Validation expérimentale mushroom

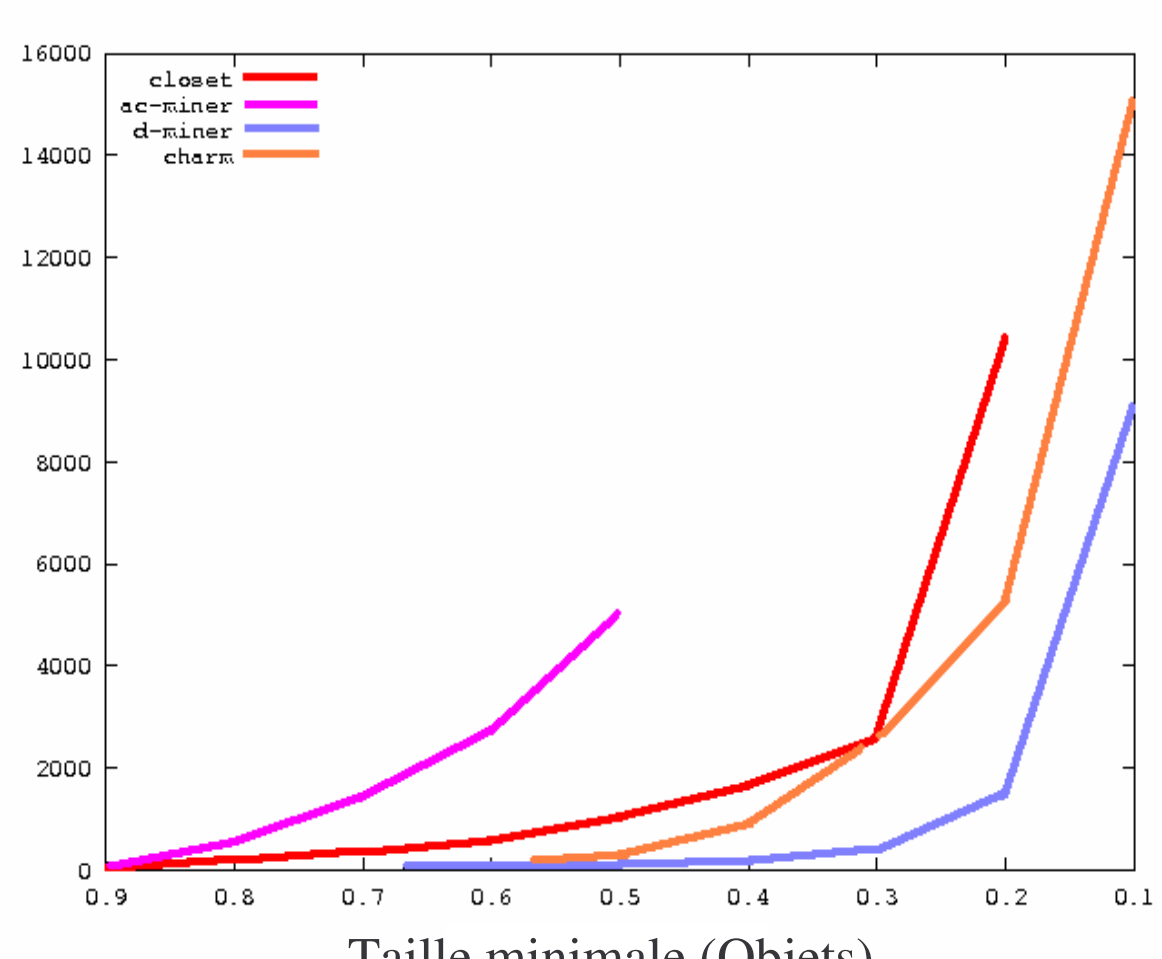
Temps (s)



Taille minimale (Objets)

Validation expérimentale connect4

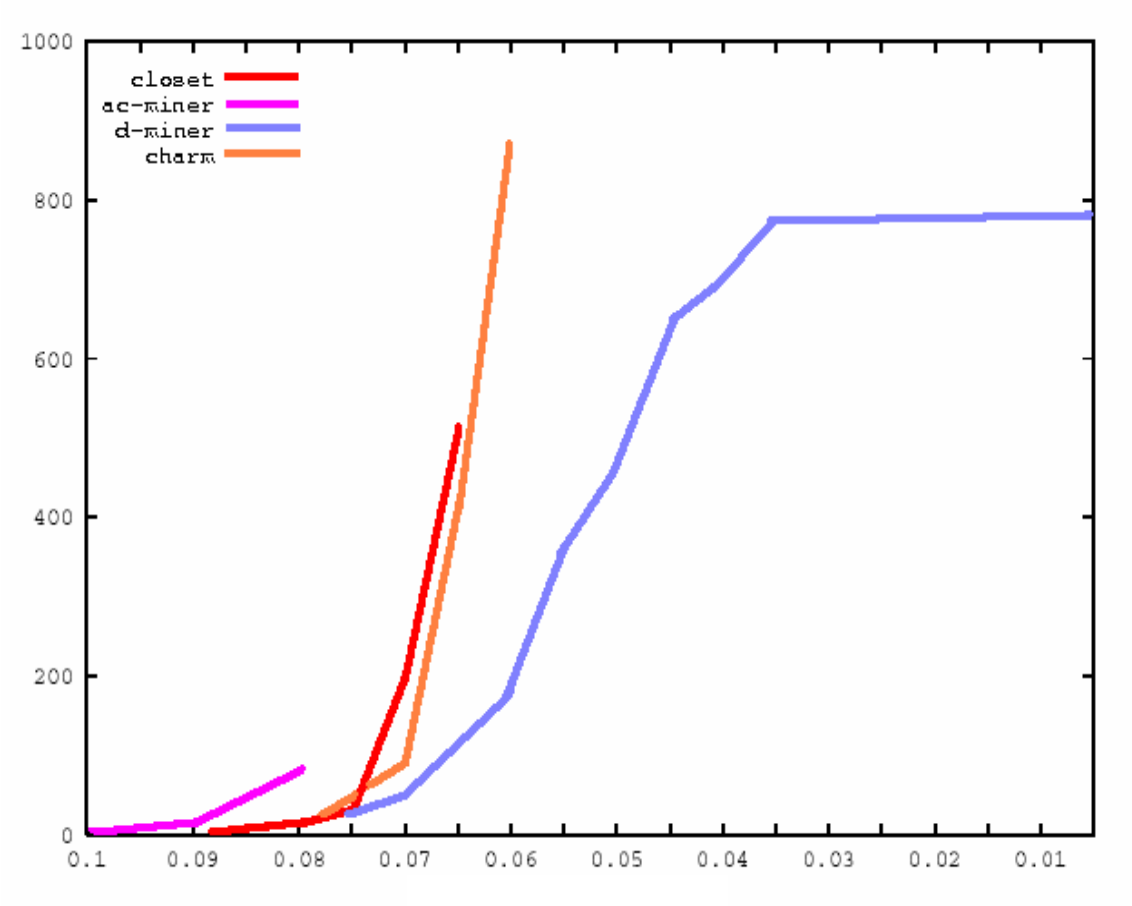
Temps (s)



Taille minimale (Objets)

Validation expérimentale données biologiques

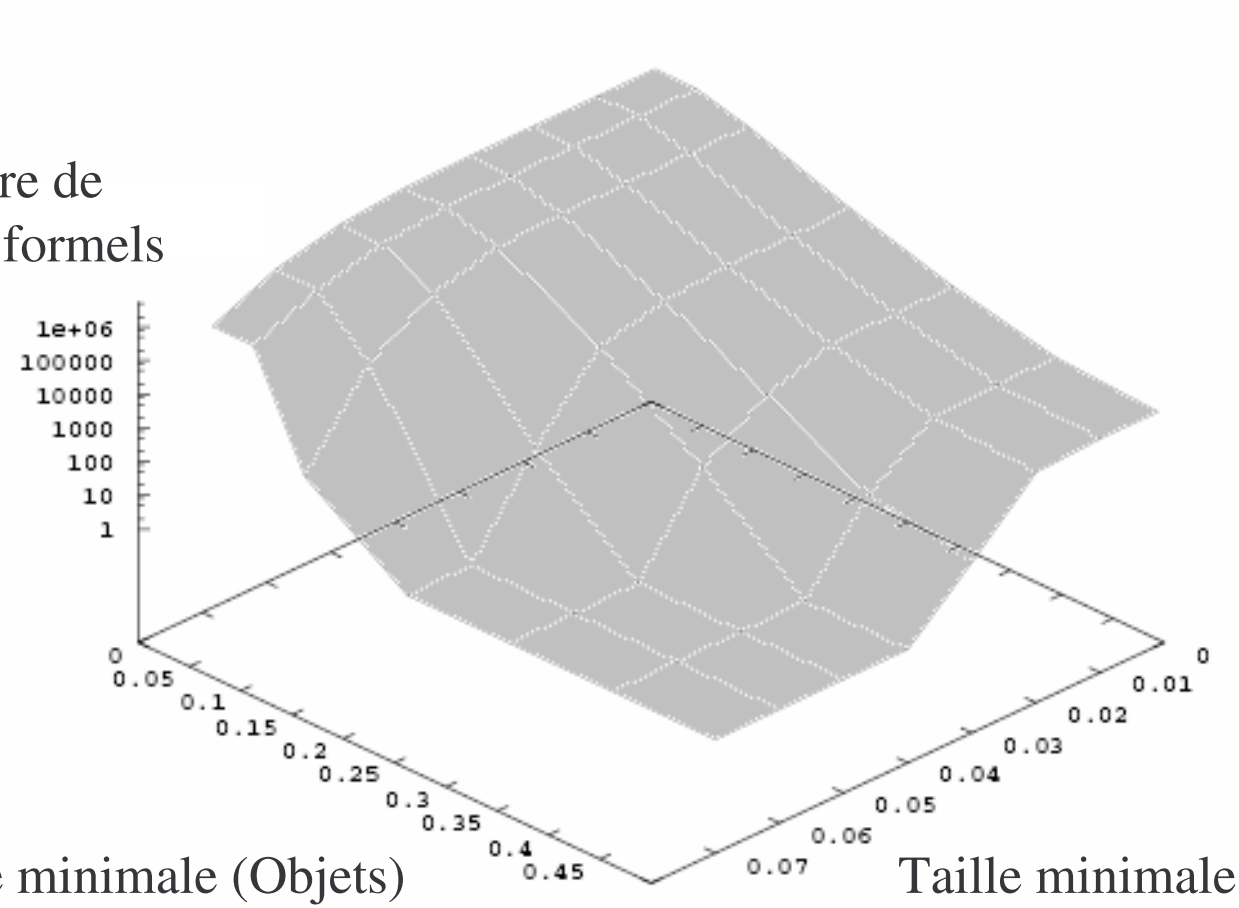
Temps (s)



Taille minimale (Objets)

Validation expérimentale données biologiques

Nombre de concepts formels



Taille minimale (Objets)

Taille minimale (Attributs)

Propriétés

- Algorithme juste et complet

- Complexité en temps avec n=|$\mathcal{O}$| et m=|$\mathcal{A}$|
  - ✓ Délai dans le pire cas : $O(n^2 * m)$
  - ✓ Délai en moyenne : $(n - \log_2(|\mathcal{C}|) + 1)\ O(n*m)$

Extraction de bi-ensembles denses et pertinents :

DR-Miner

Vision idéale

- Deux phénomènes réels
  - ✓ ($\{O_1,\ldots, O_4\},\{A_1,\ldots, A_5\}$)
  - ✓ ($\{O_5,O_6\},\{A_6,A_7\}$)

- Exemples
  - ✓ Groupes de gènes qui ont le même comportement transcriptionnel dans différentes conditions
  - ✓ Groupes de facteurs de transcription qui s'accrochent sur la région promotrice de gènes

Données idéales

Deux phénomènes réels :
$(\{O_1,\ldots, O_4\},\{A_1,\ldots, A_5\})$ et $(\{O_5,O_6\},\{A_6,A_7\})$

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
O_1	1	1	1	1	1		
O_2	1	1	1	1	1		
O_3	1	1	1	1	1		
O_4	1	1	1	1	1		
O_5						1	1
O_6						1	1
O_7							

Données réelles

Deux phénomènes réels :
$(\{O_1,\ldots, O_4\},\{A_1,\ldots, A_5\})$ et $(\{O_5,O_6\},\{A_6,A_7\})$

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
O_1	1	0	1	1	1		
O_2	1	1	1	1	0		
O_3	1	1	1	1	1		
O_4	1	1	1	1	1		
O_5						1	1
O_6					1	1	0
O_7	1						

$(\{O_1,O_2,O_3,O_4,O_7\},\{A_1\})$

$(\{O_1,O_2,O_3,O_4\},\{A_1,A_3,A_4\})$

$(\{O_2,O_3,O_4\},\{A_1,A_2,A_3,A_4\})$

$(\{O_3,O_4\},\{A_1,A_2,A_3,A_4,A_5\})$

$(\{O_1,O_3,O_4\},\{A_1,A_3,A_4,A_5\})$

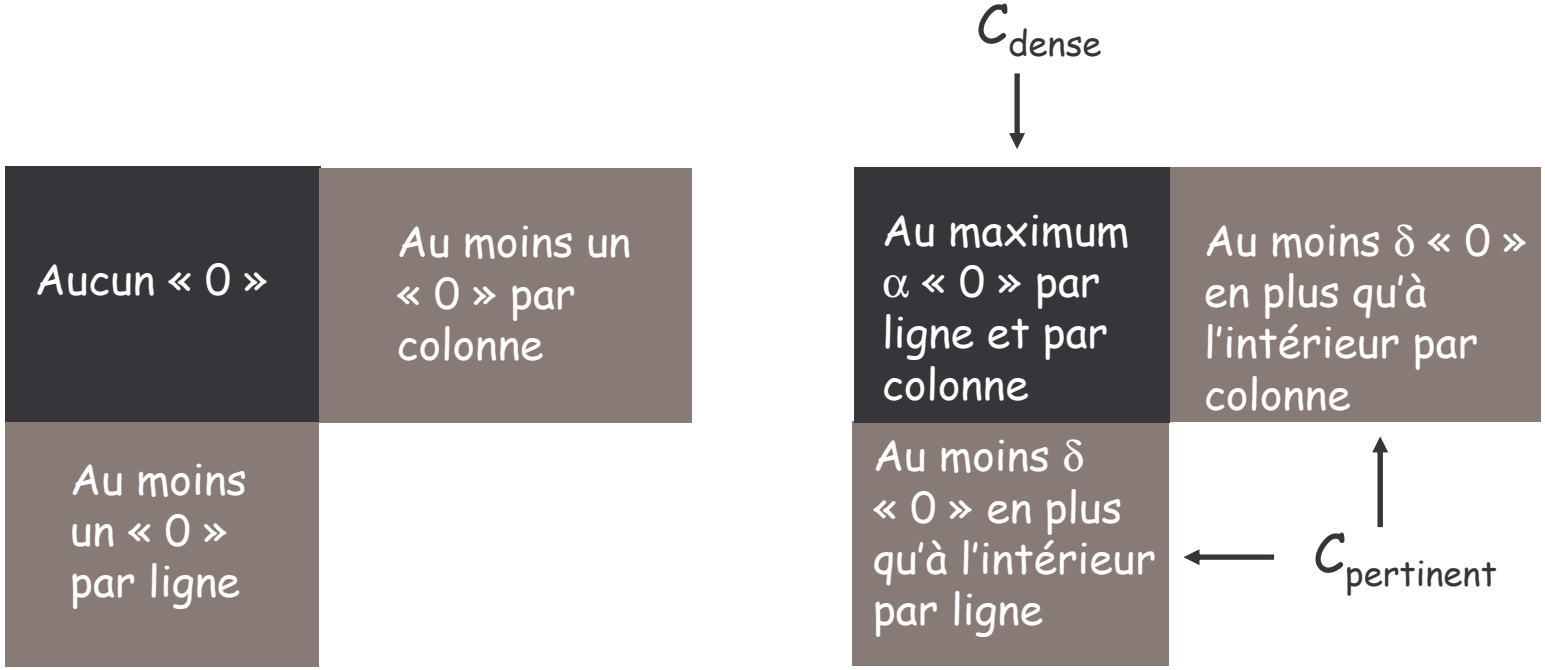
$(\{O_1,O_3,O_4,O_6\},\{A_5\})$

$(\{O_5,O_6\},\{A_6\})$

$(\{O_5\},\{A_6,A_7\})$

$(\{O_6\},\{A_5,A_6\})$

Motifs denses et pertinents



Les contraintes C_{dense} et $C_{pertinent}$ peuvent être exploitées activement durant l'extraction

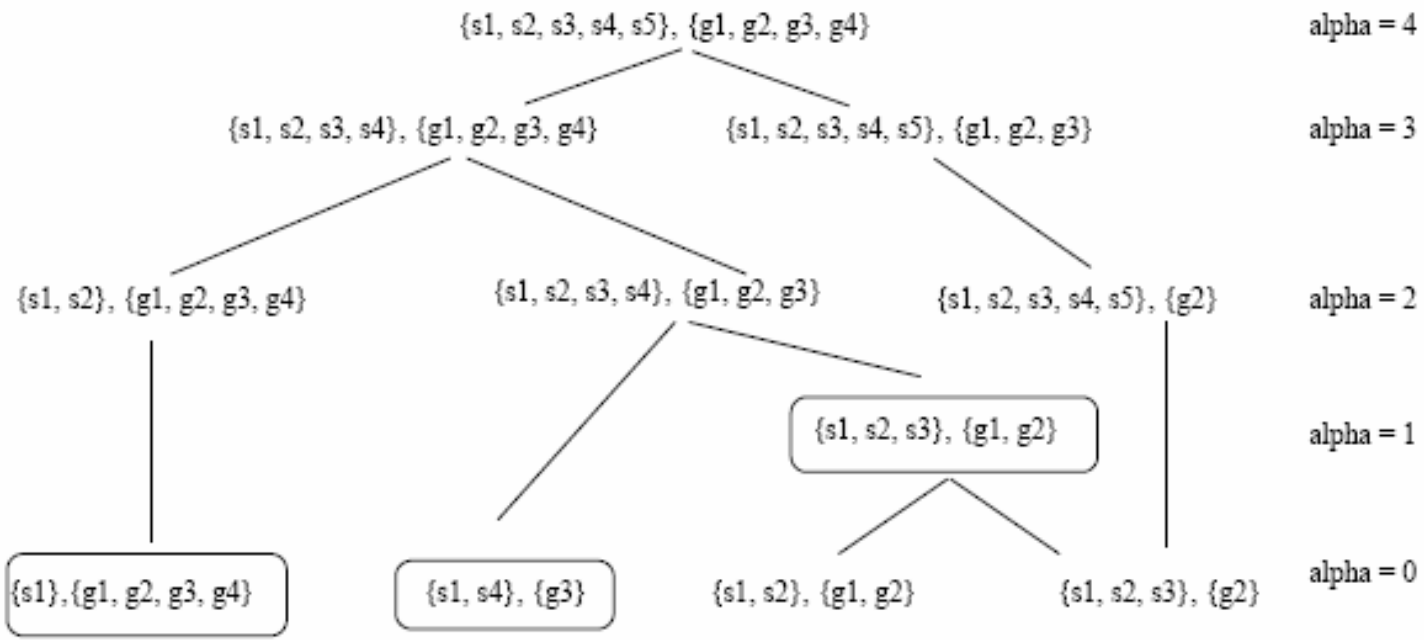
Exemple

Deux phénomènes réels :
$(\{O_1,\ldots, O_4\},\{A_1,\ldots, A_5\})$ et $(\{O_5,O_6\},\{A_6,A_7\})$

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
O_1	1	0	1	1	1		
O_2	1	1	1	1	0		
O_3	1	1	1	1	1		
O_4	1	1	1	1	1		
O_5					0	1	1
O_6					1	1	0
O_7	1						

$\alpha = \delta = 1$

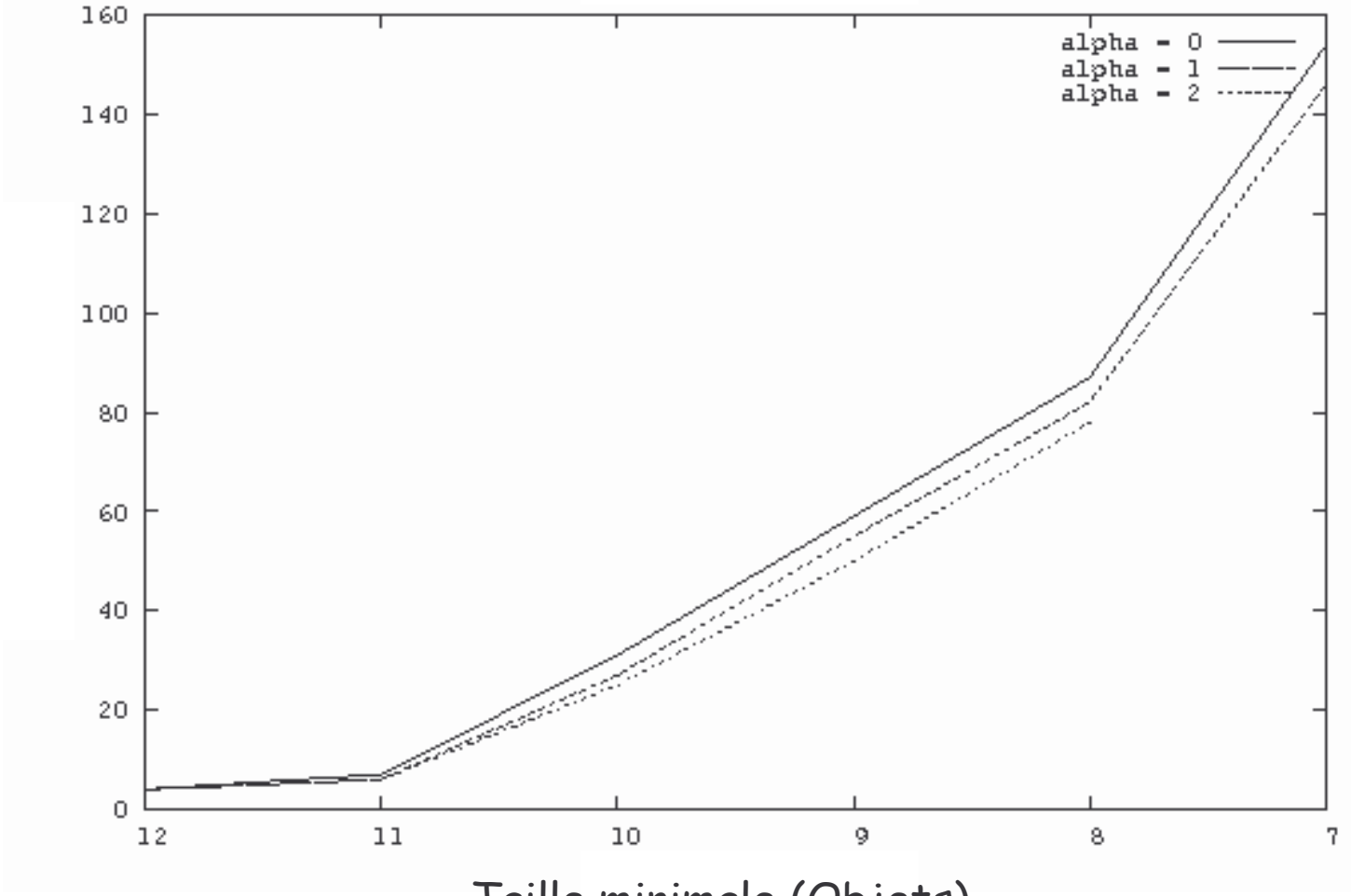
Motifs denses et pertinents



	g1	g2	g3	g4
s1	1	1	1	1
s2	1	1	0	0
s3	0	1	0	0
s4	0	0	1	0
s5	0	0	0	0

Validation expérimentale
Internet benchmark

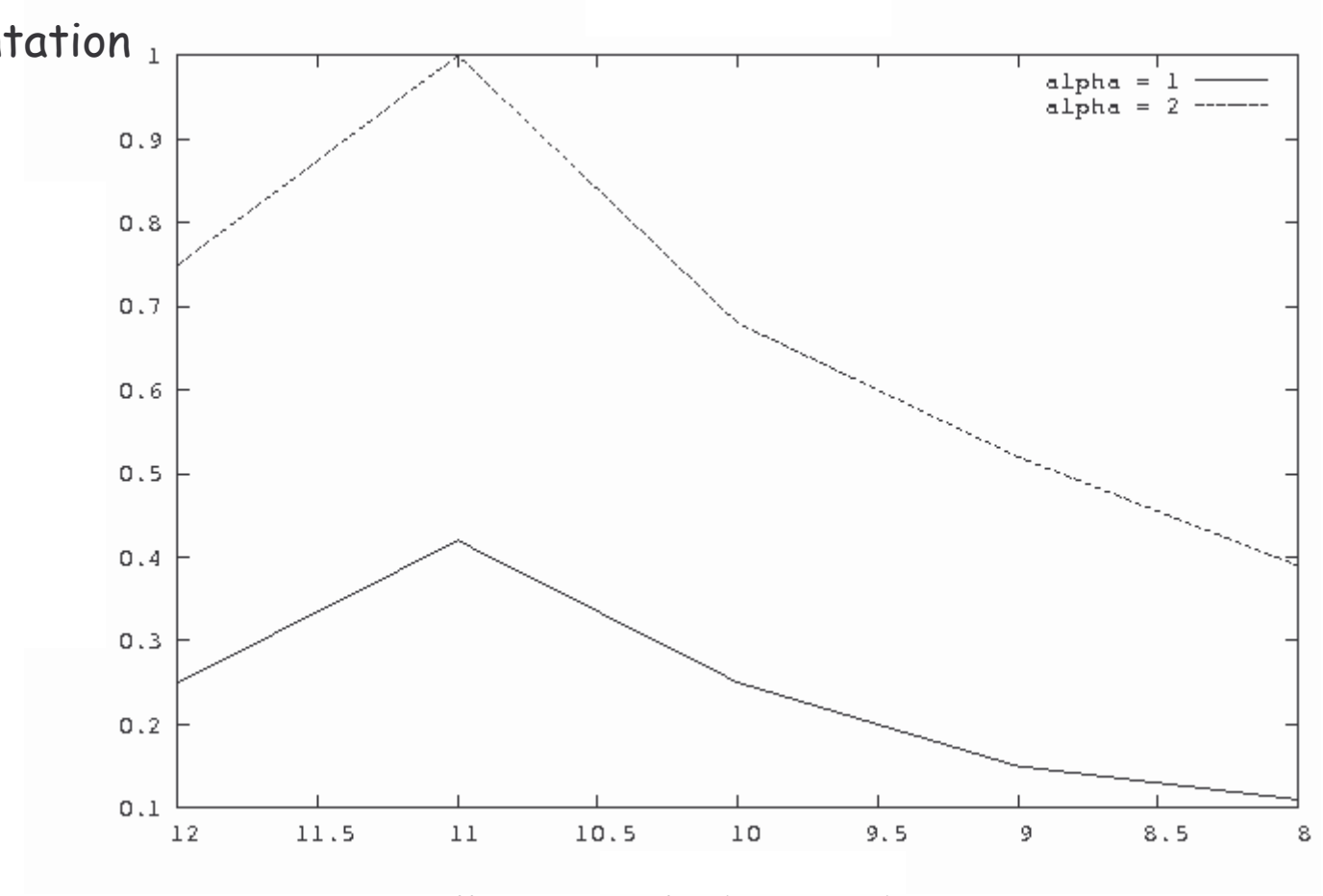
Nombre de motifs



Taille minimale (Objets)

Validation expérimentale
Internet benchmark

Pourcentage d'augmentation



Taille minimale (Objets)

Motifs denses et pertinents

- Extraction juste et complète
- Maximaux sur les deux dimensions
- Nombre d'exceptions borné par ligne et par colonne
- Fonctions entre les ensembles d'objets et les ensembles d'attributs, décroissantes pour α fixé
- Plus pertinents dans les données réelles

Application à l'insulino-résistance

Question biologique

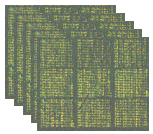
Quels sont les groupes de facteurs de transcription qui s'accrochent sur la région promotrice de gènes impliqués dans la réponse à l'insuline ?

	G_1	G_2	G_3	G_4
FT_1	1	1		1
FT_2	1	1		
FT_3	1		1	1
FT_4		1	1	1

G_1 et G_2 sont régulés par l'insuline et les facteurs de transcription TF_1 et TF_2 peuvent expliquer cette co-expression

Pré-traitement des données

Quels sont les groupes de facteurs de transcription qui s'accrochent sur la région promotrice de gènes impliqués dans la réponse à l'insuline ?



Selection et prétraitement

Données prétraitées

SAM

$G_1\ G_2\ G_4$ ...

Gènes régulés par l'insuline

(JBC 2003)

TFSEARCH

	G_1	G_2	G_3	G_4
FT_1	1	1		1
FT_2	1	1		1
FT_3	1		1	1
FT_4		1	1	1

5 Puces à ADN pangénomiques : niveau d'expression de ≈29000 gènes chez des personnes saines avant et après injection d'insuline

Extraction de concepts formels

	G_1	G_2	G_3	G_4
FT_1	1	1		1
FT_2	1	1		1
FT_3	1		1	1
FT_4		1	1	1

156 facteurs de transcription

344 gènes

Extraction de concepts formels

	G_1	G_2	G_3	G_4
FT_1	1	1		1
FT_2	1	1		1
FT_3	1		1	1
FT_4		1	1	1

156 facteurs de transcription

344 gènes

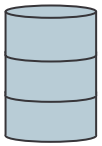
Concepts formels (X,Y)

> 5 Millions

Extraction de concepts formels

	G_1	G_2	G_3	G_4
FT_1	1	1		1
FT_2	1	1		1
FT_3	1		1	1
FT_4		1	1	1

Concepts formels (X,Y) tq
SREBP1∈X



> 3.6 Millions

SREBP1 (Sterol-responsive-element binding protein 1) est connu pour être impliqués dans la réponse transcriptionnelle de l'insuline

Extraction de concepts formels

	G_1	G_2	G_3	G_4
FT_1	1	1		1
FT_2	1	1		1
FT_3	1		1	1
FT_4		1	1	1

Concepts formels (X,Y) tq
SREBP1∈X ∧ SP1∈X ∧ NF-Y∈X

1.477

SP1 et NF-Y ont une action conjointe avec SREBP1

Validation biologique

- 13 gènes
  - ✓ SPOP, ABCA7, FEM1B, HK2, MAPRE1, MORF4F4L2, ARF4, SF1, VSP29, CRYBA4, HIG1, SDC1 et PGRMC2
- 6 facteurs de transcription
  - ✓ SREBP, SP1, NF-Y, GATA-1, GATA-2 et AML-1a
- ChIP (DNA chromatin immunoprecipitation)
  - ✓ 90% des gènes ont effectivement un site de fixation actif pour SREBP1 quand SP1 et NF-Y sont présents.
  - ✓ Un témoin négatif : pas de site de fixation pour SP1 et NF-Y $\Rightarrow$ pas de SREBP
- Nouveaux gènes cibles de SREBP1, SP1 et NF-Y

Conclusion et perspectives

Extraction de bi-ensembles

- Extractions complètes de bi-ensembles sous contraintes
  - ✓ D-Miner (extraction de concepts formels)
  - ✓ DR-Miner (extraction de motifs tolérants au bruit)

- Extraction de bi-ensembles sous contraintes
  - ✓ Nouvelles contraintes, e.g., statistiques
  - ✓ Nouveaux types de motifs, e.g., bi-ensembles dans des données numériques, autres modèles de bruit
  - ✓ Etude de l'efficacité en pratique des stratégies de propagation des contraintes et caractérisation des jeux de données

Perspectives bioinformatiques

- Des motifs aux réseaux de régulation
  - ✓ Complémentarité des types de motifs
    - Ensembles de gènes et motifs dans leurs séquences promotrices
    - Motifs pour la capture des aspects dynamiques de la régulation (e.g. , données d'expression cinétiques)
    - Motifs locaux vs. motifs globaux
- Vers des requêtes inductives complexes combinant plusieurs types de motifs au service de multiples tâches d'analyse