



---

Hinterzarten, March 11th, 2004



# Condensed representations: a key concept for inductive databases

Jean-François Boulicaut  
INSA Lyon - LIRIS CNRS FRE 2672, France

**Other contributors:** Christophe Rigotti, Baptiste Jeudy,  
Artur Bykowski, Jérémy Besson, Cyrille Masson





---

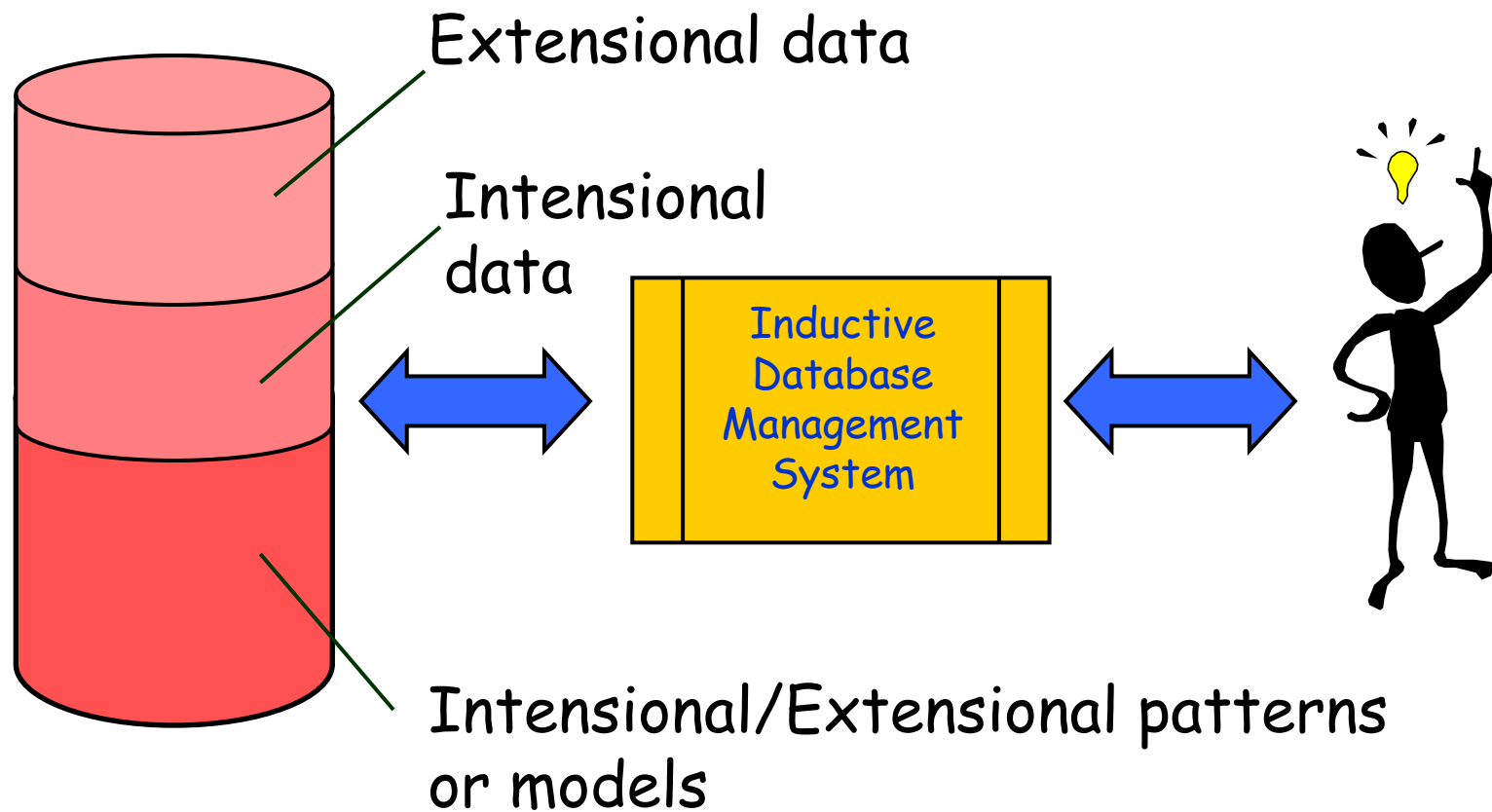
## Outline

- ◆ Introduction
    - The inductive database research group
    - *Gene expression data analysis*
  - ◆ Mining bi-sets from gene expression data
    - Step 1: Mining frequent closed sets
    - Step 2: A better use of Galois operators
    - Step 3: Constraint-based mining of bi-sets
  - ◆ Conclusion and perspectives
-



---

# The Inductive Database framework



Imielinski & Mannila 96 (cacm)

---



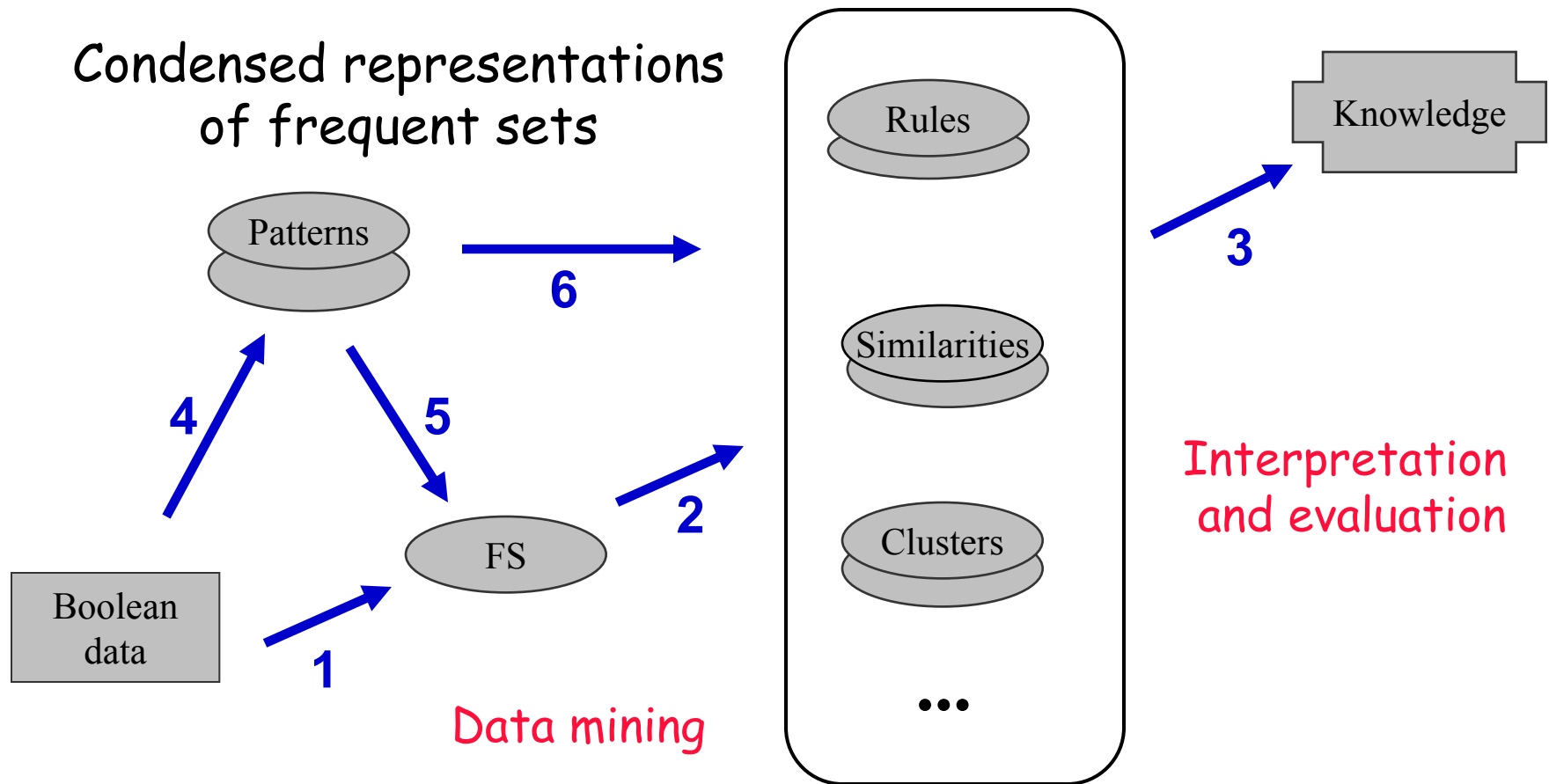
---

## Main topics

- ◆ Mining frequent sets in dense and correlated transactional data
    - ε-adequate representations w.r.t. frequency queries (condensed representations)
  - ◆ Constraint-based data mining
    - Itemsets and sequential patterns
  - ◆ Multiple uses of frequent sets
  - ◆ « Application » to gene expression data analysis
    - 3 Ph. D students co-supervised by biologists
-



# Multiples uses of frequent itemsets





---

# Gene expression data for the computer scientist

## ◆ Expression matrices

$A_1$	$A_2$	$A_3$
45	21	12
78	44	22
98	7	23
23	13	56

E.g.

SAGE (74 x 822)

SAGE (90 x 12636)

DNA chips (6 x 1065)

DNA chips (10 x 8171)

---



---

## Typical mining tasks

- ◆ A priori interesting sets of genes
    - Similar expression profiles (« clustering »)
    - **Synexpression groups**
      - Sets of genes that are frequently co-regulated
    - **Transcription modules**
      - « Bi-sets »
      - « Genes 1, 12, 37, and 93 are co-regulated together only in situations 78, 79, 80, 81 and 82 »
      - Using the ITEM pattern domain !
-



---

## Gene expression boolean contexts

- ◆ Recording over-expression and/or under-expression and/or significant variation

$A_1$	$A_2$	$A_3$
1	0	0
1	1	1
1	0	1
0	1	1

Biological situations - Genes

$2^S$  Sets of situations

$2^A$  Sets of genes

---



---

## Applying itemset extraction?

- ◆ Find each set of genes that verifies constraint  $C$

$A_1$	$A_2$	$A_3$
1	0	0
1	1	1
1	0	1
0	1	1

$A_2 A_3$  [2/4, closed, ...]

$A_1 A_2$  [1/4, closed, ...]

Association rules can be derived

Bi-sets are interesting

$(\{A_2 A_3\}, \{S_2, S_4\})$

---



---

## A pattern domain

### ◆ Language

- Bi-sets  $\langle X, T \rangle$  with  $X \in 2^A$  and  $T \in 2^S$

### ◆ Evaluation functions

- E.g., Galois operators, frequency

### ◆ Primitive constraints

- E.g.,  $C_{\text{minfreq}}$ ,  $C_{\text{maxfreq}}$ ,  $C_{\text{close}}$ ,  $C_{\text{free}}$ , syntactic constraints

### ◆ Queries

- Combinations of primitive constraint
-



---

# Galois evaluation functions

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

$f(T,r)$  set of genes shared by situations in T

$g(X,r)$  set of situations shared by genes in X

$$f(\{1,2\}) = \{A,C\}$$

$$g(\{A,B\}) = \{1,4,6\}$$

$$Fr(X,r) = |g(X,r)|$$

---



---

# Galois closure operators

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

$h(X,r) = f(g(X,r),r)$  closure for a set of genes

$h'(T,r) = g(f(T,r),r)$  closure for a set of situations

$$h(\{A,B\}) = f(\{1,4,6\}) = \{A,B,C\}$$

$$h(\{A,B,C\}) = f(\{1,4,6\}) = \{A,B,C\}$$

$$h'(\{1,2\}) = g(\{A,C\}) = \{1,2,3,4,6\}$$

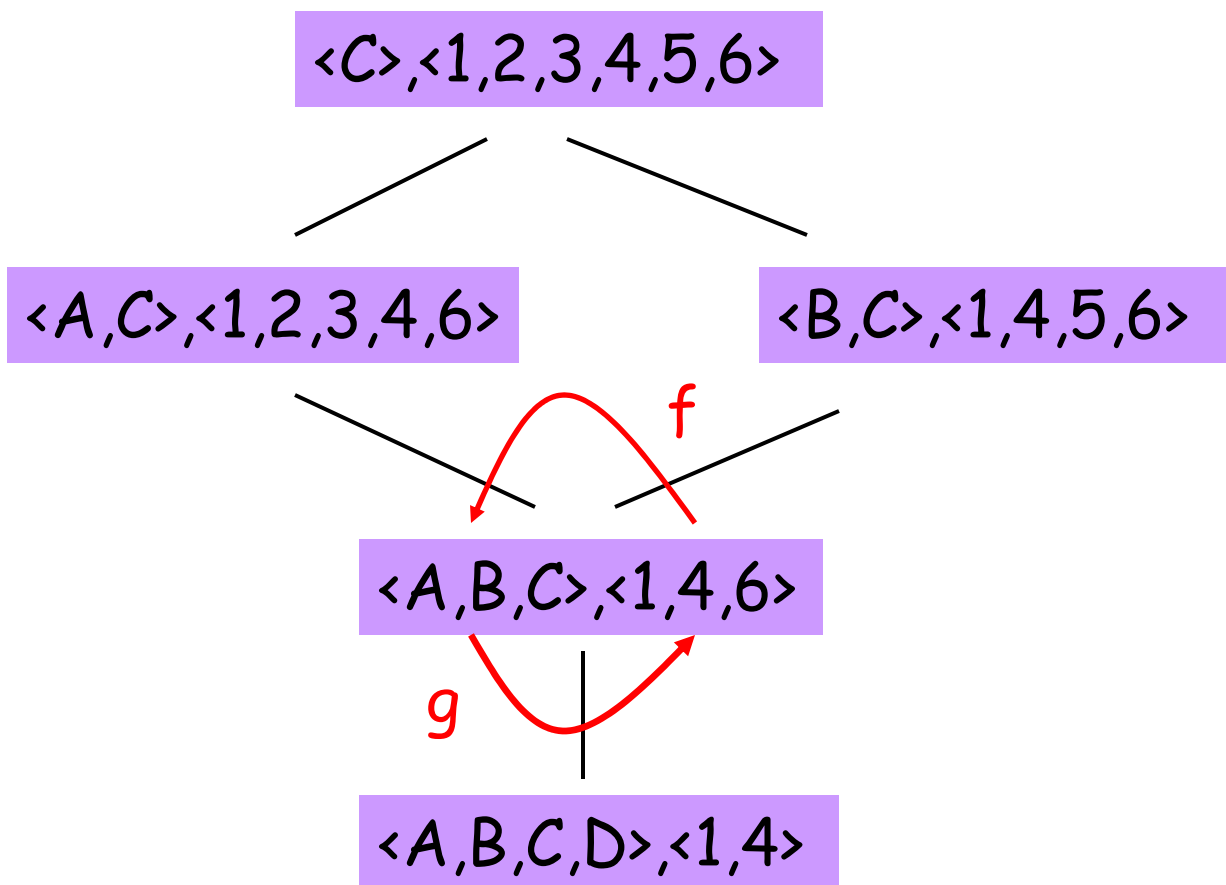
$$h'(\{1,4,6\}) = g(\{A,B,C\}) = \{1,4,6\}$$

---



## Concept lattices (Wille 82)

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0





---

## Constraints based on closures

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

$$C_{\text{close}}(X,r)$$

$$h(X,r) = X$$

$\{A,B,C\}$  is closed

$\{C,D\}$  is not closed

If  $C_{\text{free}}(X,r)$  then  $C_{\text{close}}(h(X),r)$

$C_{\text{free}}$  is anti-monotonic

(Boulicaut et al. 2000)

---



---

## Free sets

Assume

r

ABCDE  
ABCD  
ACD  
ABE  
CD  
CE

BD BC

DE

ABC ABD BCD

ACE BCE ADE BDE CDE

ABCD

ABCE ABDE ACDE BCDE

See also key patterns  
Bastide et al. 2000

**ABCDE**

---



---

## Typical queries

- ◆ Data manipulation and discretization operators
- ◆ Inductive queries returning bi-sets  $(X, T)$

- $C_{\text{minfreq}}(X, r)$   $\langle X, g(X, r) \rangle$
- $C_{\text{minfreq}}(X, r) \wedge C_{\text{close}}(X, r)$   $\langle X, g(X, r) \rangle$
- $C_{\text{minfreq}}(X, r1) \wedge C_{\text{maxfreq}}(X, r2)$   $\langle X, g(X, r1) \rangle$
- $C_{\text{close}}(X, r)$   $\langle X, g(X, r) \rangle$

« Genes 1, 12, 37, and 93 are over-expressed together only in situations 78, 79, 80, 81 and 82 »

- $C_{s1}(X, r) \wedge C_{s2}(T, r) \wedge C_{\text{close}}(X, r) \wedge C_{\text{close}}(T, r) \wedge T = g(X, r)$
-



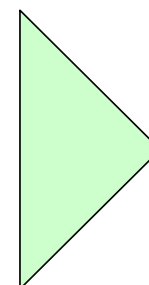
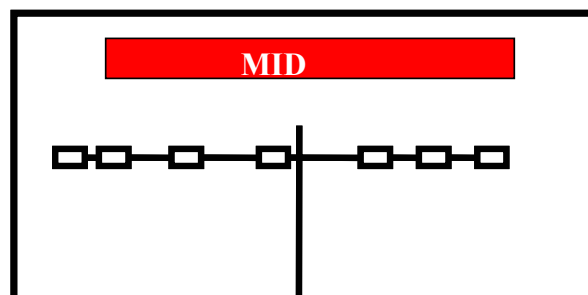
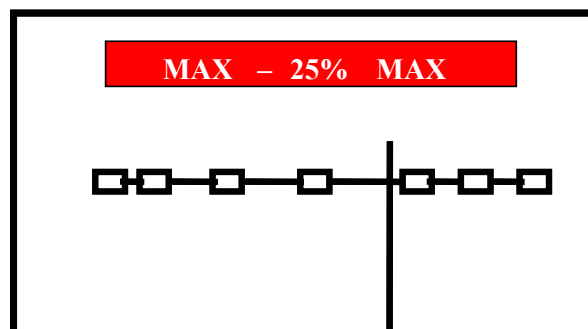
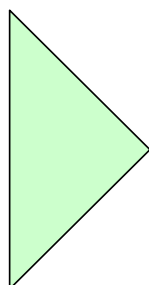
---

## Main phases of our research (1)

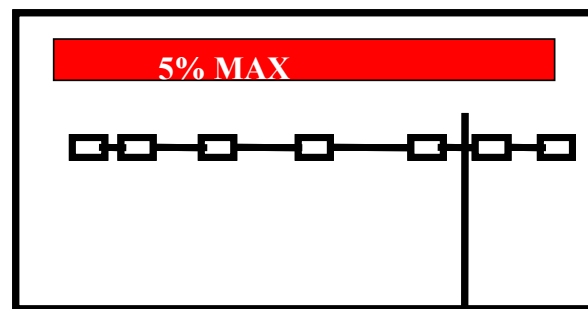
- ◆ Looking for synexpression groups
    - Mining frequent sets of genes from expression matrices was **impossible** with Apriori-like algorithms (End of 2001)
    - **What about condensed representations of frequent itemsets?**
      - It has worked well on SAGE 74x822 (mining frequent free and closed sets with ac-miner, publication in Genome Biology, december 2002)
-



Expression  
values



Boolean  
data





---

## Using ac-miner

- ◆ Min-Ex implementation (Boulicaut et al. 2000, 2003)

$X, Y : n$        $X$  is a free set of genes

$$X \cup Y = h(X, r) = h(X \cup Y, r)$$

$$\text{Fr}(X, r) = \text{Fr}(X \cup Y, r) = n$$

Two simple interpretations

$X \cup Y$  is a (frequent) closed set in  $r$

$X \Rightarrow Y$  is a « logical » association rule in  $r$

---





---

## An example of a useful association?

G-protein-coupled receptor  
related to chemokine receptors

splicing factor

**G protein-coupled receptor and KH type splicing  
regulatory protein KSRP => KIAA0340 gene (true in  
8 types of cells).**

member of the RAS gene  
superfamily

---



---

## Main phases of our research (2)

- ◆ Looking for synexpression groups
    - Mining larger expression matrices or typical microarray data with just a few situations was impossible with ac-miner (End of 2002)
    - **What about the classical properties of Galois connection?**
      - We can actually mine every concept (every closed set) in many gene expression data sets, e.g., in SAGE 90x12636 but also in microarray data (Riout et al. wdmkd '03, kdid '03)
      - Concepts are a priori interesting bi-sets
-



---

## The idea

- ◆ When looking for (frequent) closed sets of genes

Use closed set mining algorithms to extract closed sets on the smaller dimension (matrix transposition since generally we have much less situations than genes) ... when feasible

... otherwise, theoretical framework of constraint transposition Rioult et al. 2003 (wdmkd) ... ongoing

NB. Any efficient algorithm for computing (frequent) closed sets could be used.

---



---

## Direct/transposed extraction

### - Direct extraction ( $r$ )

- Computation of free sets of genes and their closures, i.e., the closed sets of genes ( $X$ )
- Associated closed sets of situations can be provided ( $g(X, r) = T$ ).
  - E.g., always intractable in SAGE 90x12636

### - Transposed extraction ( ${}^{\dagger}r$ )

- Computation of free sets of situations and their closures, i.e., the closed sets of situations ( $T$ )
  - Associated closed set of genes can be provided ( $g(T, {}^{\dagger}r) = f(T, r) = X$ ).
-



---

## Solver

- ◆ Trivial extension of **ac-miner** (frequency threshold 1)

$\langle X \rangle, \langle Y \rangle, \langle Z \rangle : n$

$X$  a free set (columns)

$X \cup Y = h(X, r) = h(X \cup Y, r)$

closed set (columns)

$Z = g(X, r)$  a closed set (lines)

We get a condensed representation of the frequent (closed) sets (MUFS)

Some free sets are available!

---



---

## Experiments with SAGE data (1)

	Discretization	Density	Nb free sets	Nb closed sets
M	ENE	82.8	intractable	intractable
<sup>r</sup> M	ENE	82.8	intractable	intractable
M	Mid-Ranged	12.2	13 580 544	80 068
<sup>r</sup> M	Mid-Ranged	12.2	209 829	80 068
M	Max - 25% Max	3.8	35 934	1 386
<sup>r</sup> M	Max - 25% Max	3.8	3 211	1 386
M	5% Max	4.8	72 630	1 808
<sup>r</sup> M	5% Max	4.8	3 362	1 808

Table 1. Results for  $M = 74 \times 822$

---



---

## Experiments with SAGE data (2)

	Discretization	Density	Nb free sets	Nb closed sets
M	ENE	34.5	intractable	intractable
<sup>r</sup> M	ENE	34.5	intractable	intractable
M	Mid-Ranged	4.8	intractable	intractable
<sup>r</sup> M	Mid-Ranged	4.8	324 565	196 130
M	Max - 25% Max	2.2	intractable	intractable
<sup>r</sup> M	Max - 25% Max	2.2	21 603	9 150
M	5% Max	4.7	intractable	intractable
<sup>r</sup> M	5% Max	4.7	54 762	31 766

Table 2. Results for  $M = 90 \times 12\ 636$

---



---

## Main phases of our research (3)

### ◆ Mining formal concepts

- Mining extended boolean contexts with transcription factors (or a large number of situations) remains intractable (Summer 2003)
  - Problems when none of the dimensions is « small enough » or when the density is high
  - **What about constraint-based mining of concepts?**
    - Ongoing work: the D-Miner algorithm
-



---

## Conclusion and perspectives (1)

- ◆ Concept **post-processing** is ongoing
    - Post-processing has to be supported (e.g., the use of available biological databases like *GO*)
      - Ph.D. S. Blachon - Ph.D. J. Besson
    - Querying huge collections of itemsets, concepts, association rules, etc. (Ph.D. C. Masson)
    - Visualization techniques (Ph.D. R. Pensa)
      - **Short term goal: providing free access (WWW) to pattern bases (and querying tools) to the scientific community**
-



---

## Conclusion and perspectives (2)

- ◆ Designing **query languages** for inductive databases on gene expression data (data + bi-sets) and prototypical KDD scenarios
  - ◆ Combining **itemset** and **string** pattern domains
    - Genes are sequences
    - Biological situations are sometime ordered (time)
  - ◆ Providing domain **specific inductive databases for molecular biology** seems to be a major issue for future research
-



---

## Acknowledgements

- Olivier Gandrillon (CGMC, University Lyon 1)
  - Sylvain Blachon (CGMC, University Lyon 1)
  - Jérémy Besson (LIRIS and INRA U449)
  - Sophie Rome (INRA U449)
  - Cyrille Masson (LIRIS)
  - Christophe Rigotti (LIRIS)
  - Ruggero Pensa (LIRIS)
  - Céline Robardet (LIRIS and now Prisma)
  - Artur Bykowski and Baptiste Jeudy (former Ph.D students)
  - Bruno Crémilleux (GREYC, University of Caen)
  - François Rioult (GREYC, University of Caen)
-