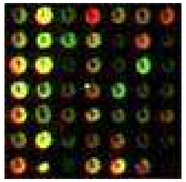
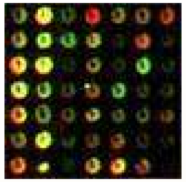
Mining formal concepts and alternative relevant bi-sets in gene expression data

Jérémy Besson Jean-François Boulicaut Céline Robardet

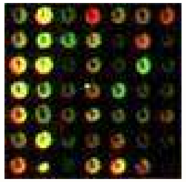


Introduction

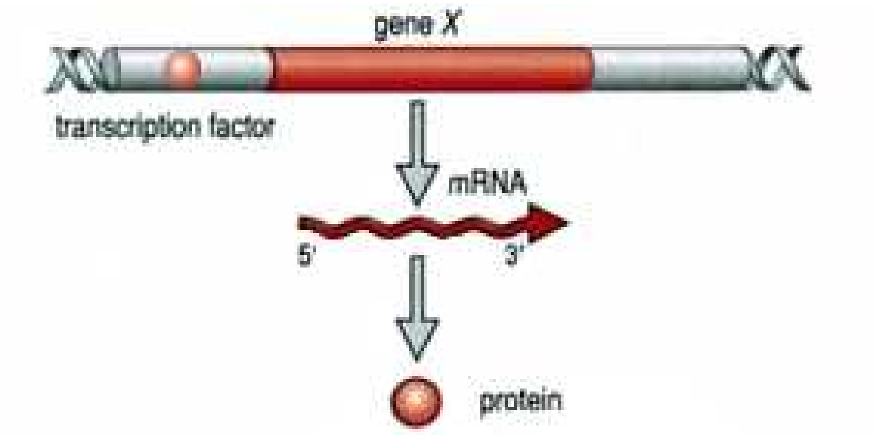


Overview

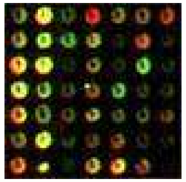
- Gene expression data
- Extracting large formal concepts on both dimensions
- How define relevant pattern in noisy data?
- DR bi-sets: new fault tolerant pattern type which extend formal concepts



Molecular biology paradigm



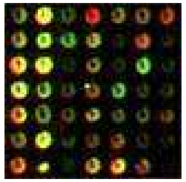
- Cell roles are determined by the proteins they produce
- Gene expression is regulated by complex mechanisms to adapt cells to their environment.
- Transcription Factors (TFs) bind on DNA upstream gene to induce or inhibit gene expression.



Boolean gene expression data

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0

- Encode gene over-expression in 4 biological experiments (number of gene copies over a threshold)

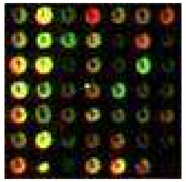


Boolean gene expression data

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0

- Encode gene over-expression in 4 biological experiments (number of gene copies over a threshold)
- Data enrichment by biological knowledge

$\Rightarrow$ Boolean table which encodes several relations



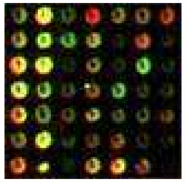
Boolean gene expression data

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0

G_4 and G_5 are over expressed in biological experiments E_3 and E_4. Transcription factors TF_1 and TF_2 can be at the origin of this co-expression.

Biological hypothesis:

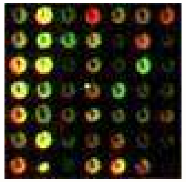
- G_4 and G_5 have an identical function
- G_4 and G_5 take part in the same regulatory pathway
- they are co-regulated by TF_1 and TF_2



Bi-sets as local patterns

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0

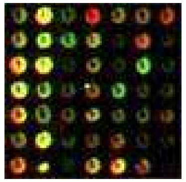
- 1-RECTANGLE: $\{\{E_2E_3\},\{G_1G_2\}\}$ is a 1-rectangle.



Bi-sets as local patterns

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0

- 1-RECTANGLE: $\{\{E_2E_3\},\{G_1G_2\}\}$ is a 1-rectangle.
- FORMAL CONCEPT: $\{\{E_2E_3E_4\},\{G_1G_2G_3G_4\}\}$ is a formal concept.



Bi-sets as local patterns

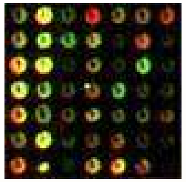
More formally...

- Boolean data encode relation **r** $\subseteq \mathcal{O} \times \mathcal{I}$
- Galois connection

$$(X,Y) \in 2^{\mathcal{O}} \times 2^{\mathcal{I}}$$
$$\phi(X) = \{i \in \mathcal{I} \mid \forall x \in X, (x,i) \in \mathbf{r}\}$$
$$\psi(Y) = \{o \in \mathcal{O} \mid \forall y \in Y, (o,y) \in \mathbf{r}\}$$

- (X,Y) is a FORMAL CONCEPT if $\psi(Y) = X$ and $\phi(X) = Y$

NB: X and Y are closed sets



Using user-defined constraints

Selection of relevant patterns (X, Y)

- Minimal frequency or minimal size constraints on objects or items
- Presence or absence of objects or items
- Minimal area constraint
- ...

Towards extraction feasibility

APRIORI specialization order: $\subseteq$, $\emptyset \rightarrow \mathcal{I}$

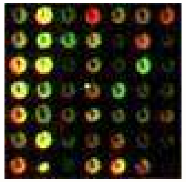
Anti-monotonic constraints

- ❖ $|X| > \tau_1$
- ❖ $x \in X$

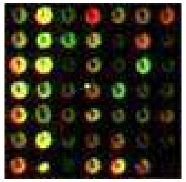
$\Rightarrow |X|$ decreases

- ❖ $|Y| < \tau_2$
- ❖ $y \notin Y$

$\Rightarrow |Y|$ increases



Extracting formal concepts under constraints



Why a novel algorithm?

Boolean contexts:

- dense
- whose two dimensions are large

We want to compute large formal concepts on both dimensions, which is often impossible by post-treatment

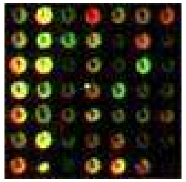
We wish to "push" constraints as:

- $|X| > \tau_1$ (minimal size constraint on objects)
- $|Y| > \tau_2$ (minimal size constraint on items)
- $|X| \times |Y| > \tau$ (minimal area constraint)

$\Rightarrow$ Specialization order $\preceq$

$(X, Y) \preceq (X', Y')$ iff $X \subseteq X'$ and $Y \subseteq Y'$

$(\mathcal{O}, \mathcal{I}) \rightarrow (\emptyset, \emptyset)$



D-MINER

Constraint-based mining of formal concepts in transactional data, PaKDD'04.

Constraint-based concept mining and its application to microarray data analysis, IDA'05.

	G_1	G_2	G_3	G_4
E_1	1	1	0	0
E_2	1	1	1	0
E_3	0	1	1	1

(X, Y) is a formal concept:

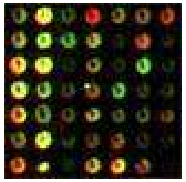
- $E_1 \in X$ then $G_3 \notin Y$ and $G_4 \notin Y$
- $G_4 \in Y$ then $E_1 \notin X$ and $E_2 \notin X$

} 1-rectangle ★

- $E_1 \notin X$ then $G_3 \in Y$ or $G4 \in Y$
- $G1 \notin Y$ then $E3 \in X$

} Maximality ★

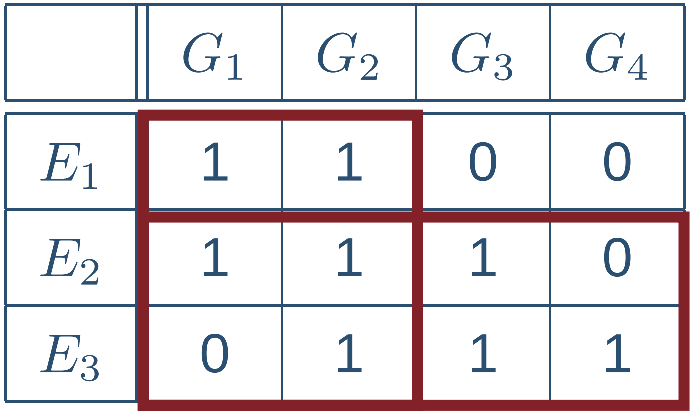
Enumeration on rows



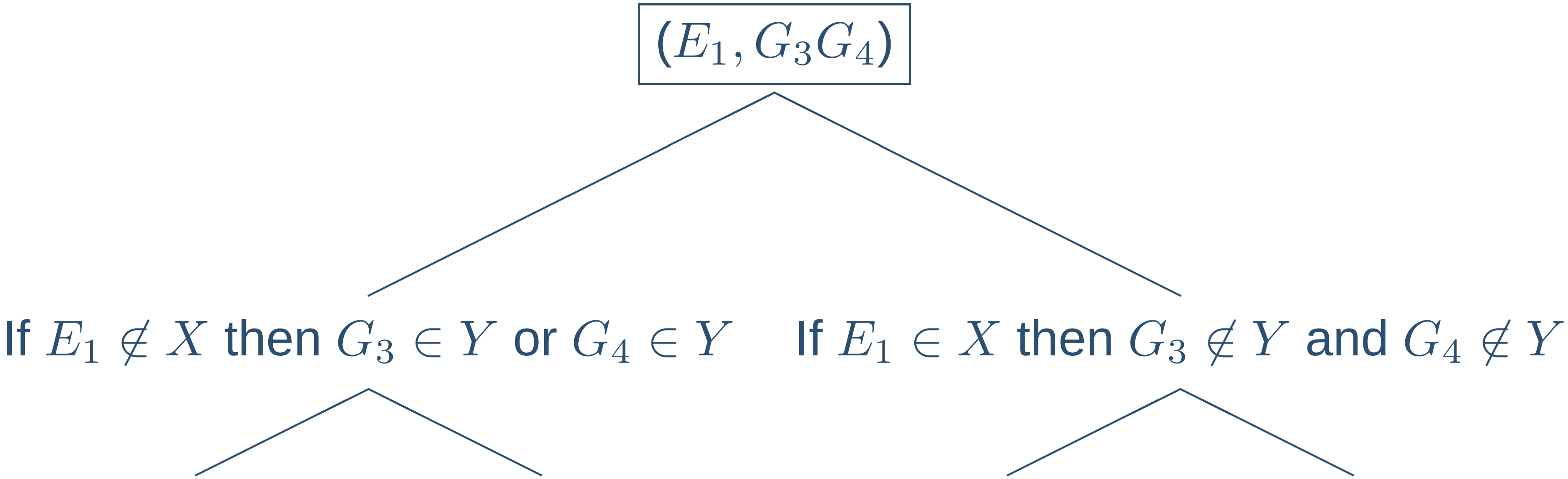
D-MINER

Constraint-based mining of formal concepts in transactional data, PaKDD'04.

Constraint-based concept mining and its application to microarray data analysis, IDA'05.



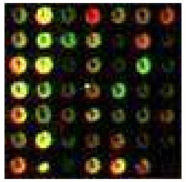
	G_1	G_2	G_3	G_4
E_1	1	1	0	0
E_2	1	1	1	0
E_3	0	1	1	1



(E_1, G_3G_4)

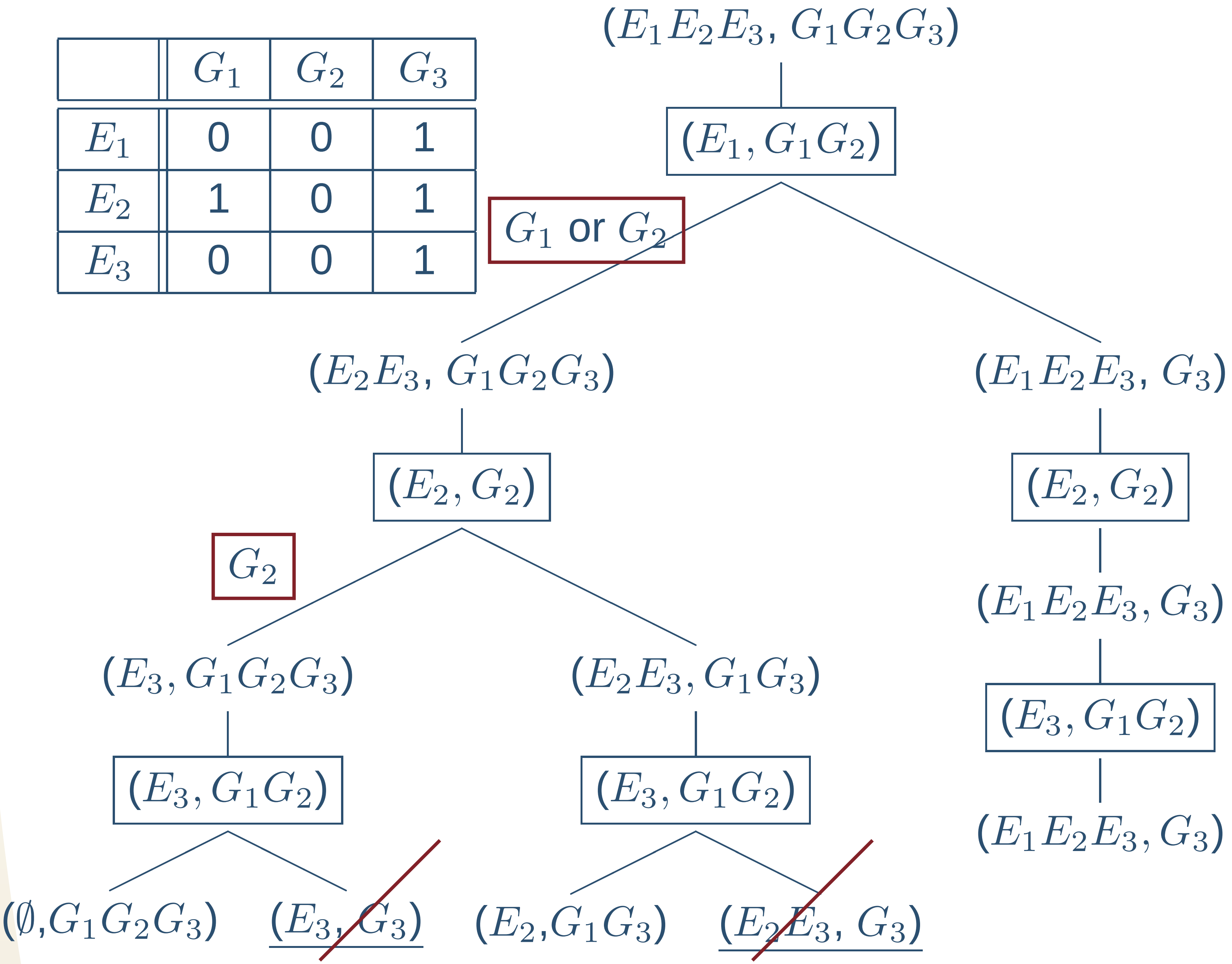
If $E_1 \notin X$ then $G_3 \in Y$ or $G_4 \in Y$

If $E_1 \in X$ then $G_3 \notin Y$ and $G_4 \notin Y$



D-MINER

	G_1	G_2	G_3
E_1	0	0	1
E_2	1	0	1
E_3	0	0	1



$(E_1E_2E_3,\ G_1G_2G_3)$

$\boxed{(E_1, G_1G_2)}$

G_1 or G_2

$(E_2E_3,\ G_1G_2G_3)$ $(E_1E_2E_3,\ G_3)$

$\boxed{(E_2, G_2)}$ $\boxed{(E_2, G_2)}$

G_2

$(E_1E_2E_3, G_3)$

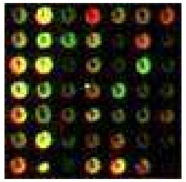
$(E_3, G_1G_2G_3)$ (E_2E_3, G_1G_3)

$\boxed{(E_3, G_1G_2)}$

$\boxed{(E_3, G_1G_2)}$ $\boxed{(E_3, G_1G_2)}$

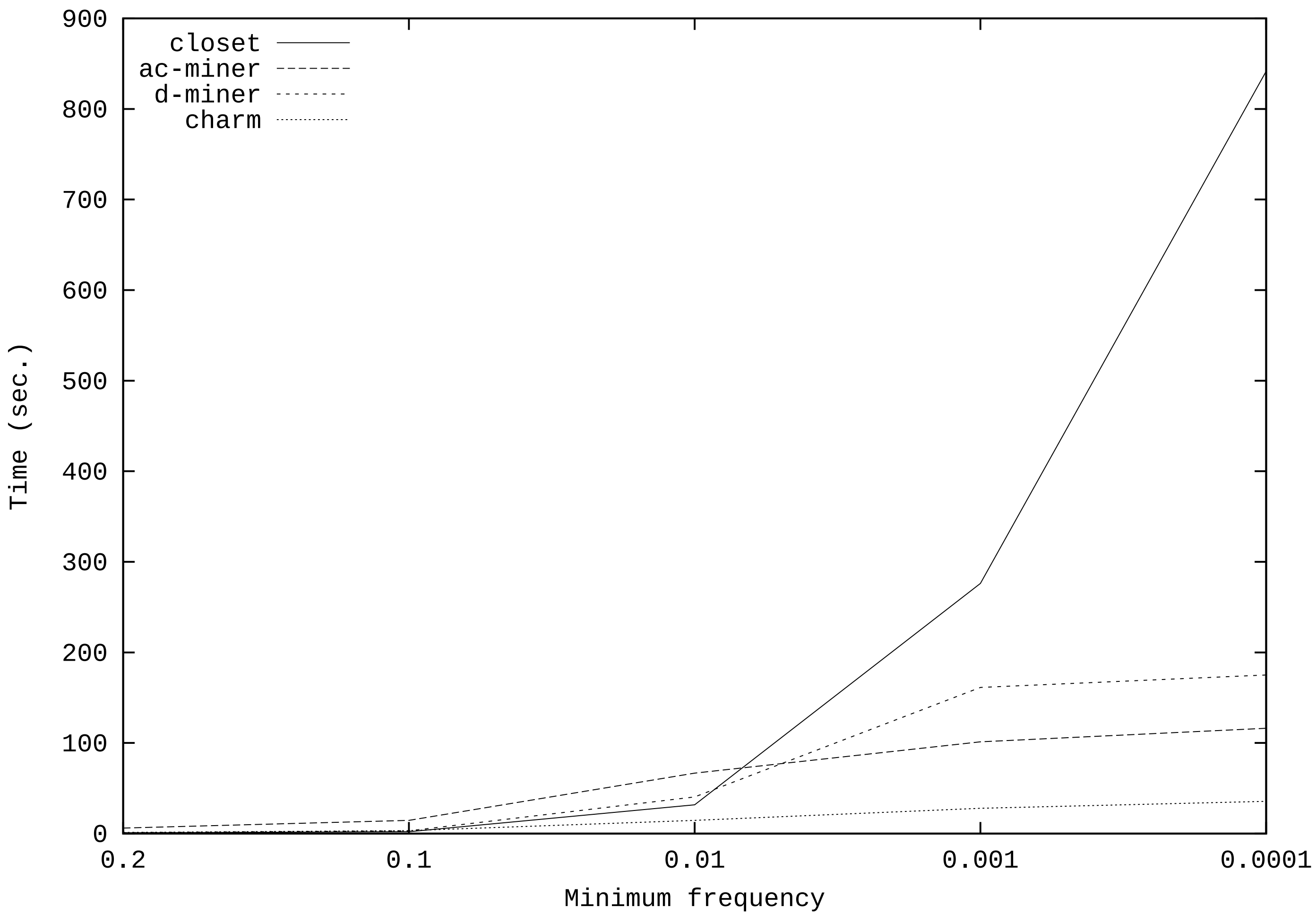
$(E_1E_2E_3, G_3)$

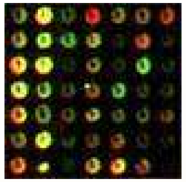
$(\emptyset, G_1G_2G_3)$ $\underline{\cancel{(E_3, G_3)}}$ (E_2, G_1G_3) $\underline{\cancel{(E_2E_3,\ G_3)}}$



Experimental evaluation

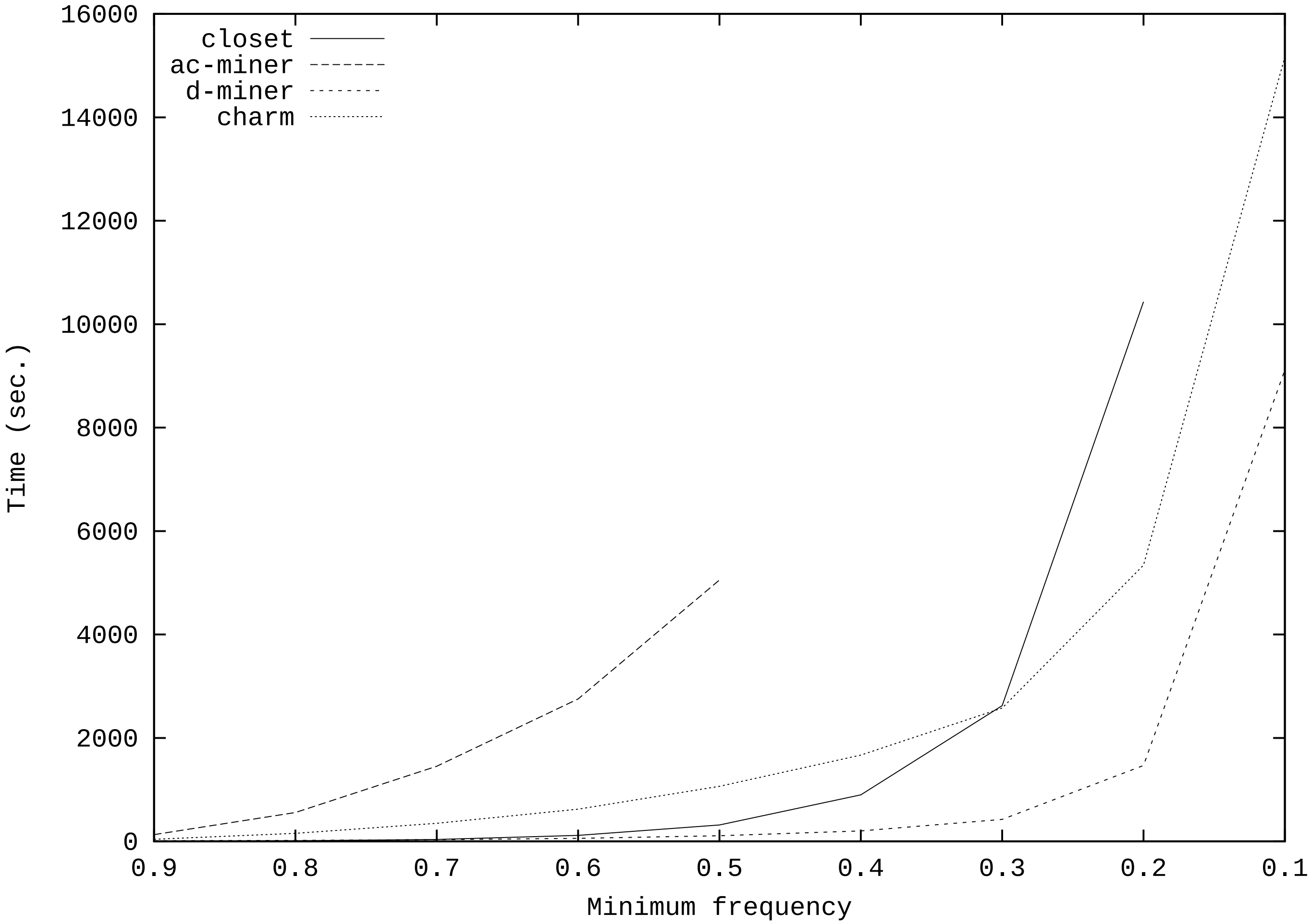
Mushroom (8 124 × 120)

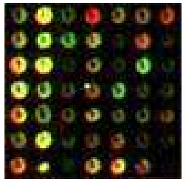




Experimental evaluation

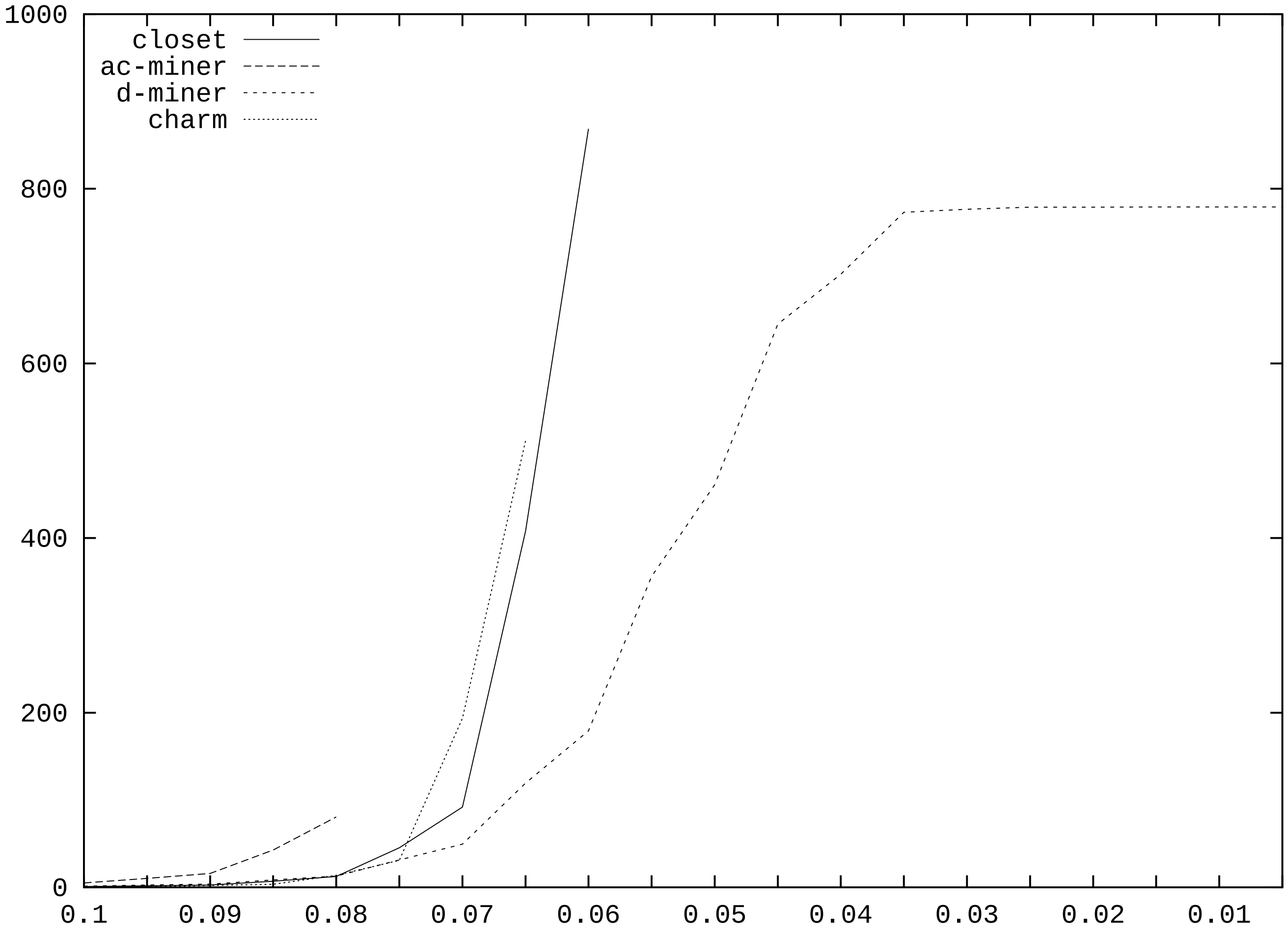
Connect 4 (67 557 × 149)

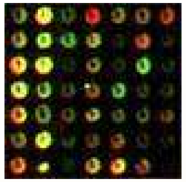




Experimental evaluation

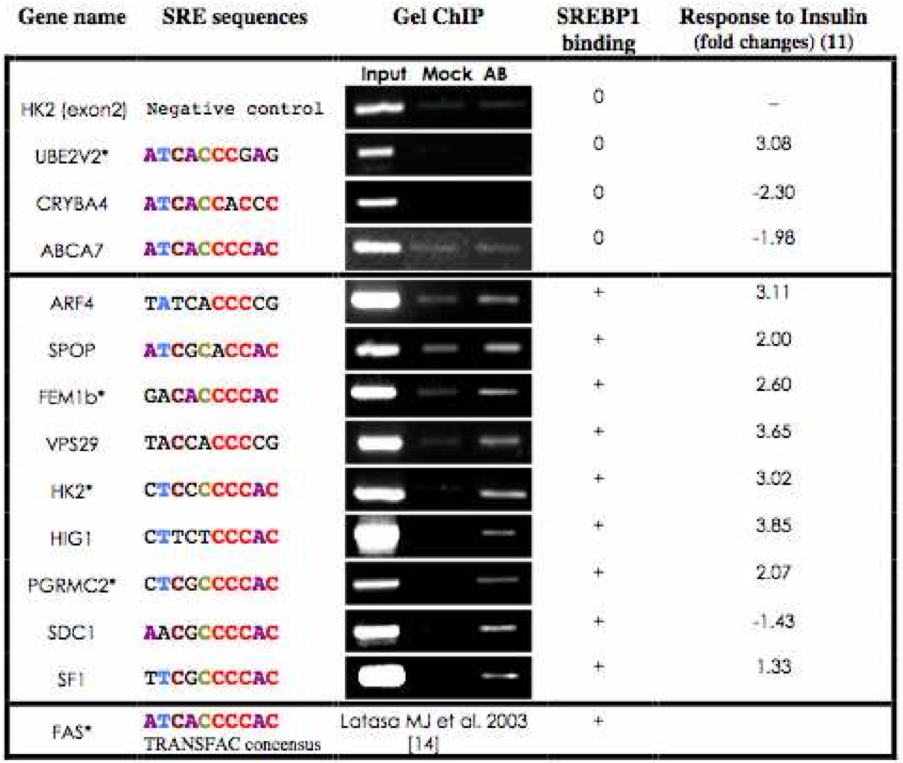
Biological Data (104 × 304)





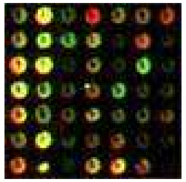
Biological evaluation

Resolving transcription network from microarray data with constraint-based formal concept mining revealed new target genes of SREBP1, submitted to PLOS



Gene name	SRE sequences	Gel ChIP Input Mock AB	SREBP1 binding	Response to Insulin (fold changes) (11)
HK2 (exon2)	Negative control		0	–
UBE2V2*	ATCACCCGAG		0	3.08
CRYBA4	ATCACCACCC		0	-2.30
ABCA7	ATCACCCCAC		0	-1.98
ARF4	TATCACCCCG		+	3.11
SPOP	ATCGCACCAC		+	2.00
FEM1b*	GACACCCCAC		+	2.60
VPS29	TACCACCCCG		+	3.65
HK2*	CTCCCCCCAC		+	3.02
HIG1	CTTCTCCCAC		+	3.85
PGRMC2*	CTCGCCCCAC		+	2.07
SDC1	AACGCCCCAC		+	-1.43
SF1	TTCGCCCCAC		+	1.33
FAS*	ATCACCCCAC TRANSFAC concensus	Latasa MJ et al. 2003 [14]	+	

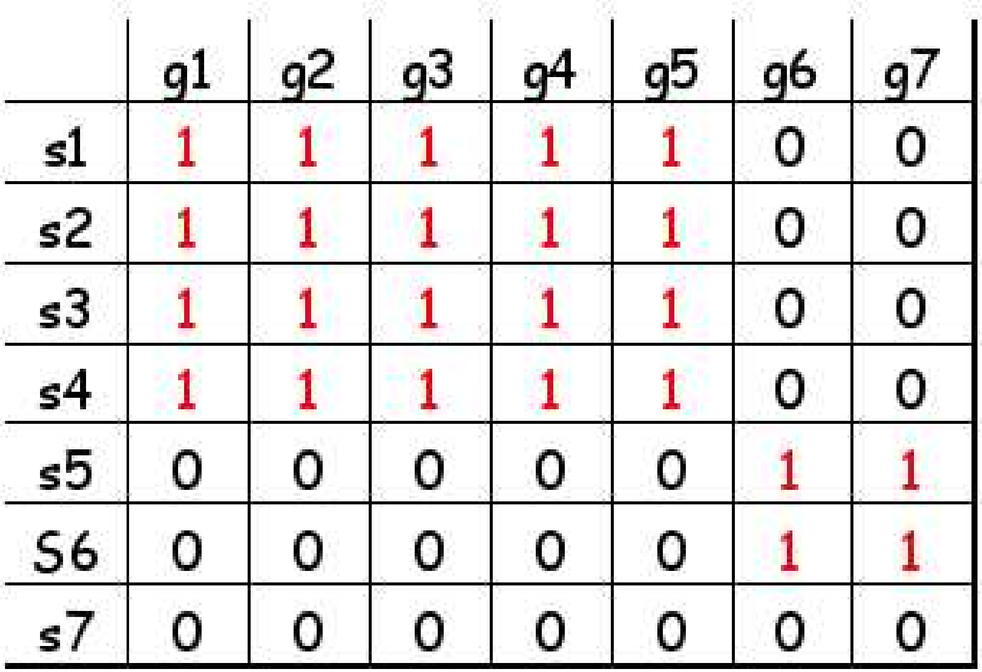
- Formal concepts containing TFs SREBP1, SP1 and NFY
- The largest concept contains 13 genes
- Chromatin immunoprecipitation shown that SREBP1 bind on the DNA promoter sequence of 10 of these genes.



Relevant bi-sets in noisy data

Improving a priori relevancy of extracted bi-sets

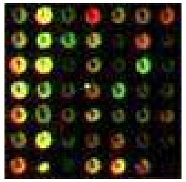
- Formal concepts: maximum sets of objects and items which are always in relation
- “Too strong” association ⇒ noise sensitivity.
- Large 1-rectangles with “few” exceptions should be useful.



	g1	g2	g3	g4	g5	g6	g7
s1	1	1	1	1	1	0	0
s2	1	1	1	1	1	0	0
s3	1	1	1	1	1	0	0
s4	1	1	1	1	1	0	0
s5	0	0	0	0	0	1	1
S6	0	0	0	0	0	1	1
s7	0	0	0	0	0	0	0

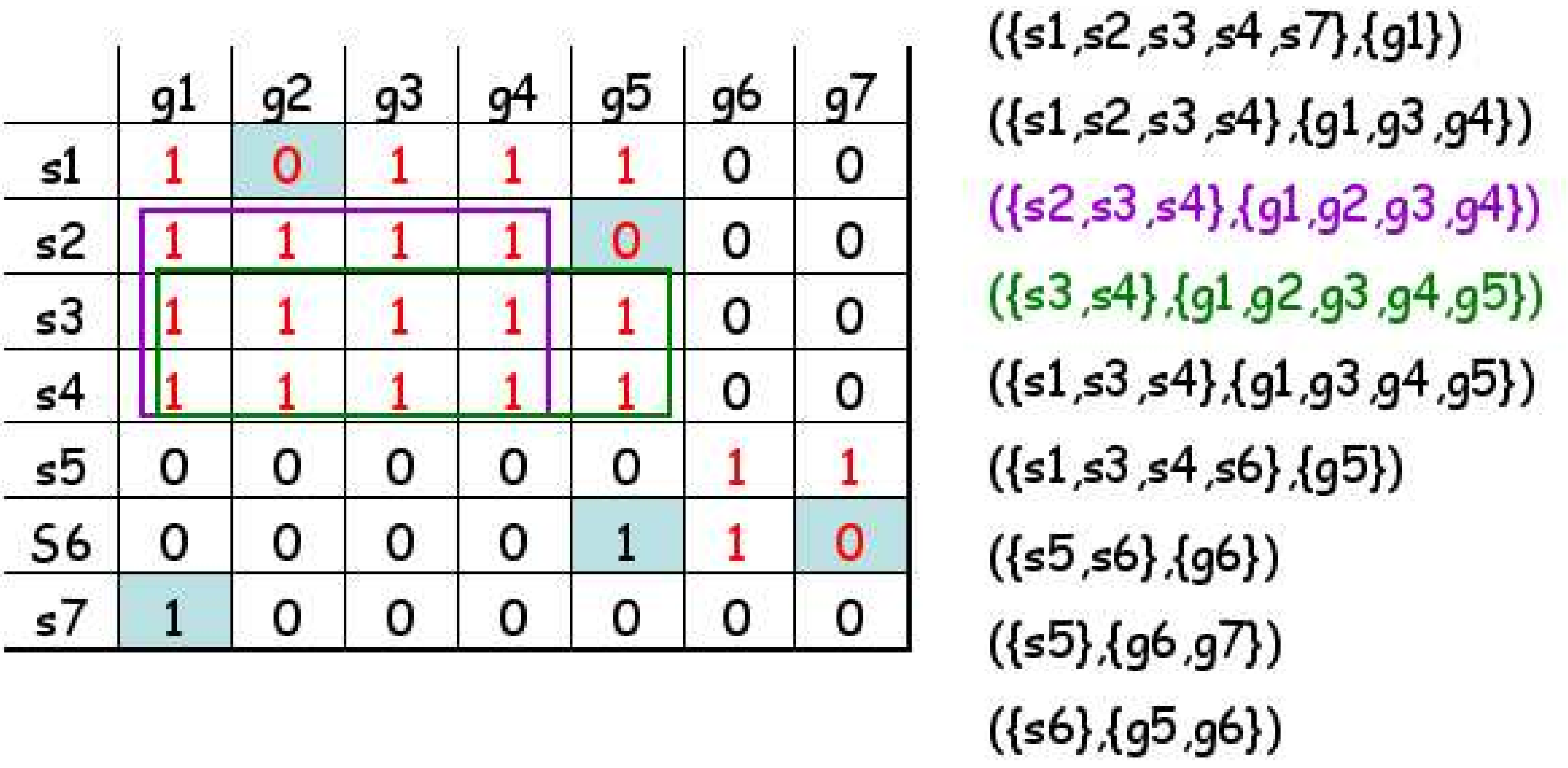
({s1,s2,s3,s4},{g1,g2,g3,g4,g5})

({s5,s6},{g6,g7})



Improving a priori relevancy of extracted bi-sets

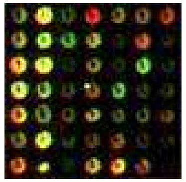
- Formal concepts: maximum sets of objects and items which are always in relation
- “Too strong” association ⇒ noise sensitivity.
- Large 1-rectangles with “few” exceptions should be useful.



	g1	g2	g3	g4	g5	g6	g7
s1	1	0	1	1	1	0	0
s2	1	1	1	1	0	0	0
s3	1	1	1	1	1	0	0
s4	1	1	1	1	1	0	0
s5	0	0	0	0	0	1	1
S6	0	0	0	0	1	1	0
s7	1	0	0	0	0	0	0

({s1,s2,s3,s4,s7},{g1})
({s1,s2,s3,s4},{g1,g3,g4})
({s2,s3,s4},{g1,g2,g3,g4})
({s3,s4},{g1,g2,g3,g4,g5})
({s1,s3,s4},{g1,g3,g4,g5})
({s1,s3,s4,s6},{g5})
({s5,s6},{g6})
({s5},{g6,g7})
({s6},{g5,g6})

10% of noise



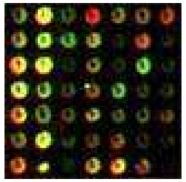
Improving a priori relevancy of extracted bi-sets

Desired properties:

- Complete extractions
- Bounded number of false values on rows and on columns
- Maximality on both dimensions
- Relevancy: there is no row outside the pattern which is identical to a pattern one (resp. column).
- Existence of functions $\phi : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{I}}$ and $\psi : 2^{\mathcal{I}} \rightarrow 2^{\mathcal{O}}$
- Galois connection

  1. $\forall v, w \in \mathcal{O},$ if $v \leq_{\mathcal{O}} w$ then $\phi(w) \leq_{\mathcal{I}} \phi(v),$
  2. $\forall i, j \in \mathcal{I},$ if $i \leq_{\mathcal{I}} j$ then $\psi(j) \leq_{\mathcal{O}} \psi(i),$
  3. $\forall v \in \mathcal{O}, \forall i \in \mathcal{I}, v \leq_{\mathcal{O}} \psi(\phi(v))$ and $i \leq_{\mathcal{I}} \phi(\psi(i))$

- Efficiency on real data



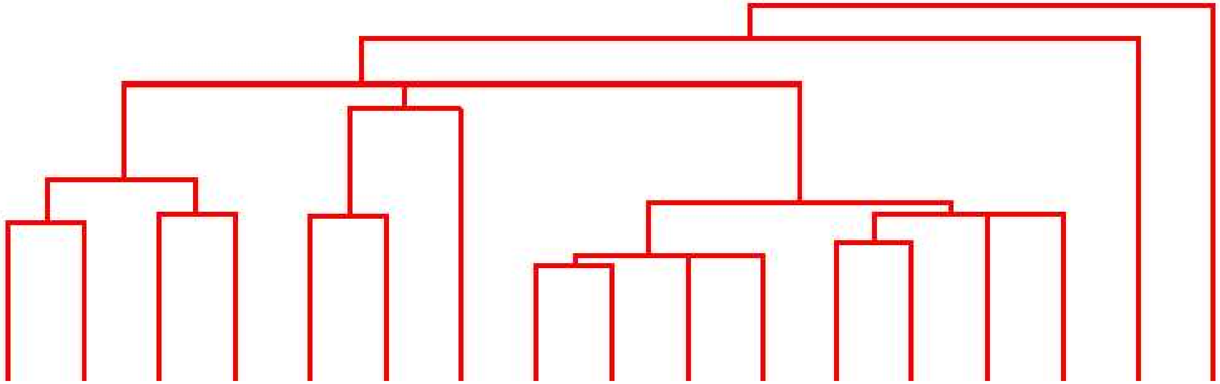
Improving a priori relevancy of extracted bi-sets

Team proposals:

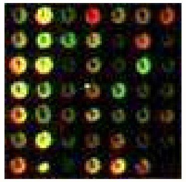
- Hierchical clustering of formal concepts: *Using classification and visualization on pattern databases for gene expression data analysis, PaRMa'04*
- δ-free sets and their pseudo closure: *Free-sets: a condensed representation of boolean data for the approximation of frequency queries, DMKD*
- Formal concept merging: *Mining Formal Concepts with a Bounded Number of Exceptions from Transactional Data, KDID'04*
- Relevant and dense bi-sets: *Mining relevant and dense bi-sets as an alternative to closed set extraction from noisy data, submitted to ICDM'05*

Hierarchical clustering of formal concepts

Use a distance based on the element frequencies in formal concepts



	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0



Hierarchical clustering of formal concepts

Introduction

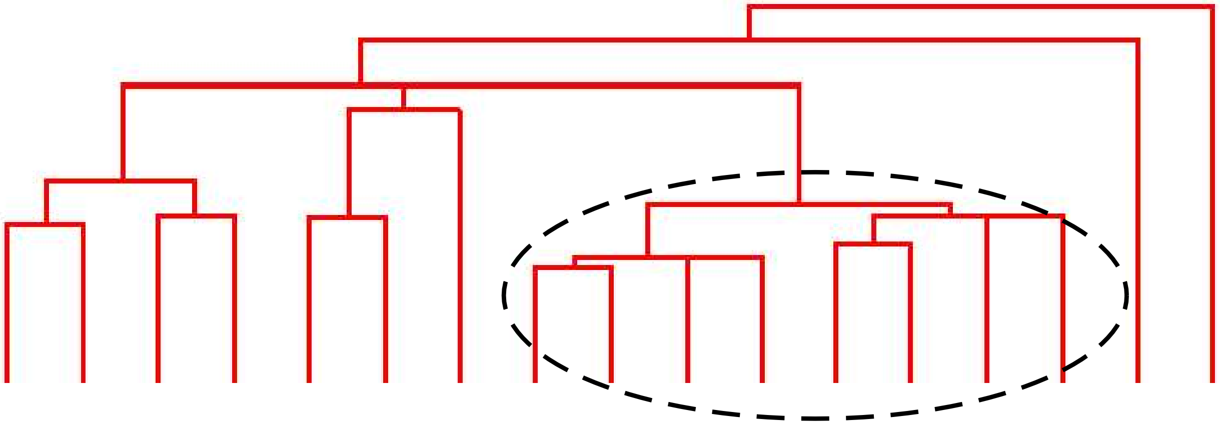
Extracting formal concepts under constraints

Relevant bi-sets in noisy data

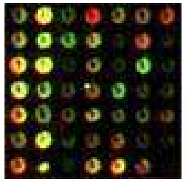
- Improving a priori relevancy of extracted bi-sets
- Hierarchical clustering of formal concepts
- Closed-sets, free-sets and δ-free sets
- Bi-sets from δ-free sets
- Merging formal concepts

Relevant and dense bi-sets in noisy data

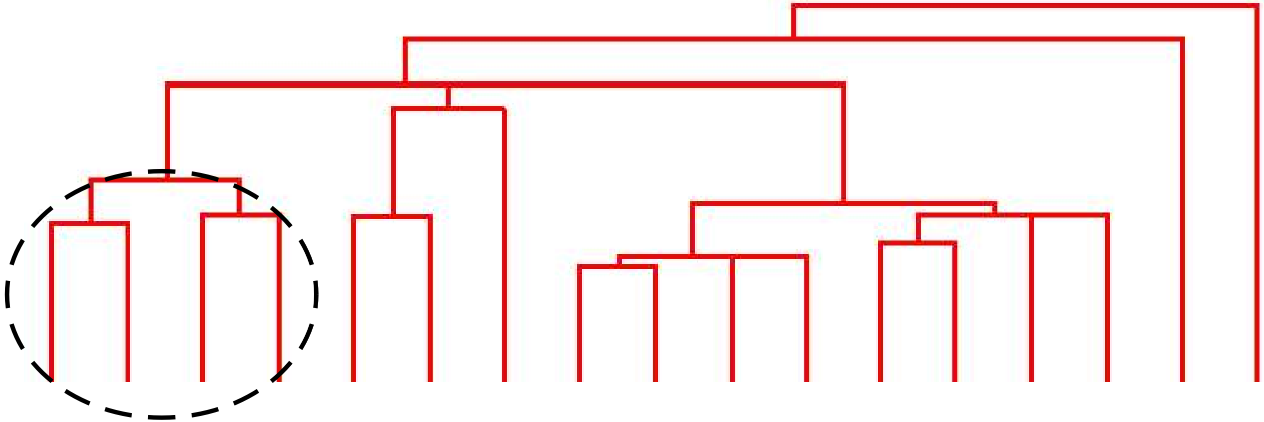
Conclusion



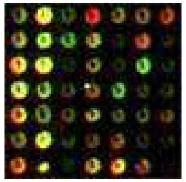
	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0



Hierarchical clustering of formal concepts



	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	0	0



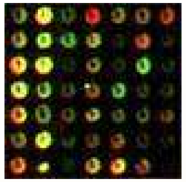
Hierarchical clustering of formal concepts

Advantages

- Visualization
- Easily computable

Drawbacks

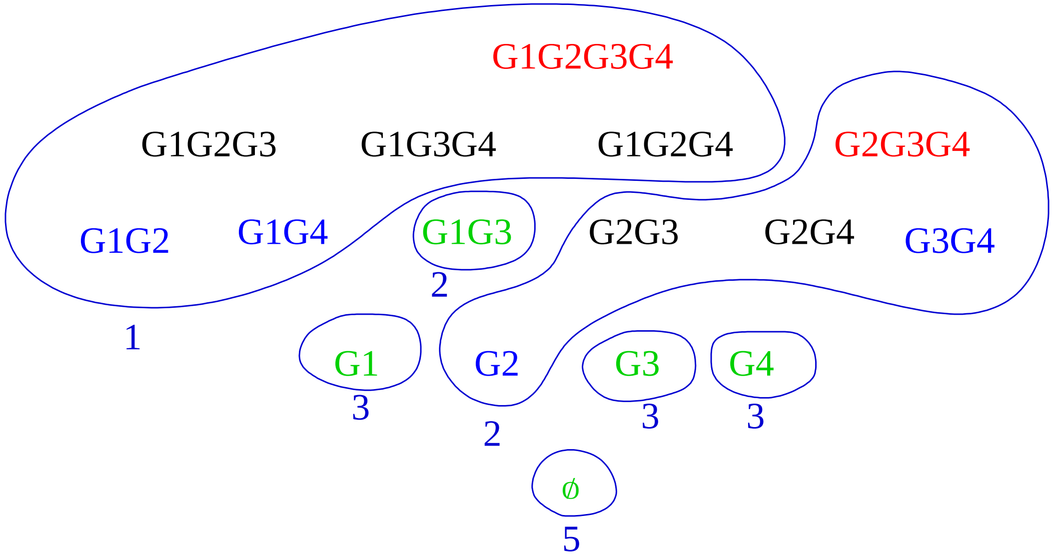
- No formal properties on the extracted patterns

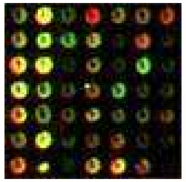


Closed-sets, free-sets and δ-free sets

- Close$_{item}(Y) = Y'$ where Y' is the greatest overset of Y such that $\psi(Y') = \psi(Y)$
- Close$_{item}(Y) = \phi(\psi(Y))$
- Equivalence relation: $(X, Y) \sim (X', Y') \Leftrightarrow$ Close$_{item}(Y) =$Close$_{item}(Y')$
- Equivalence classes
  - The **closed** itemset is the maximum element of this classe
  - The **free** itemsets are the minimal elements of this classe (also called key patterns)

G_1	G_2	G_3	G_4
1	0	1	0
1	1	1	1
0	1	1	1
0	0	0	1
1	0	0	0

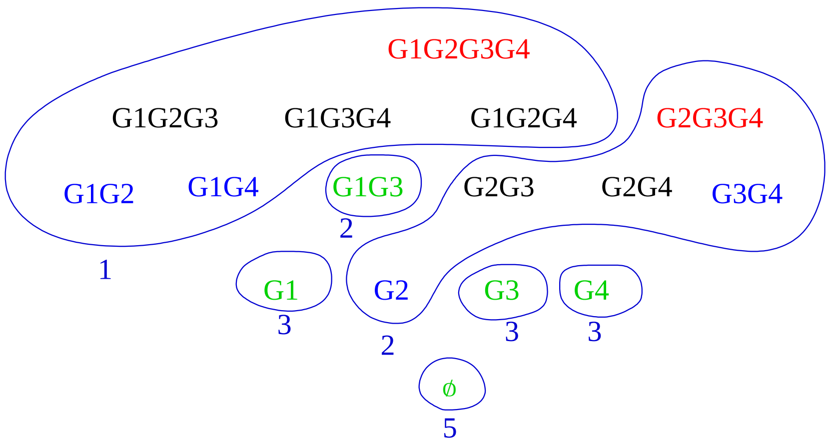


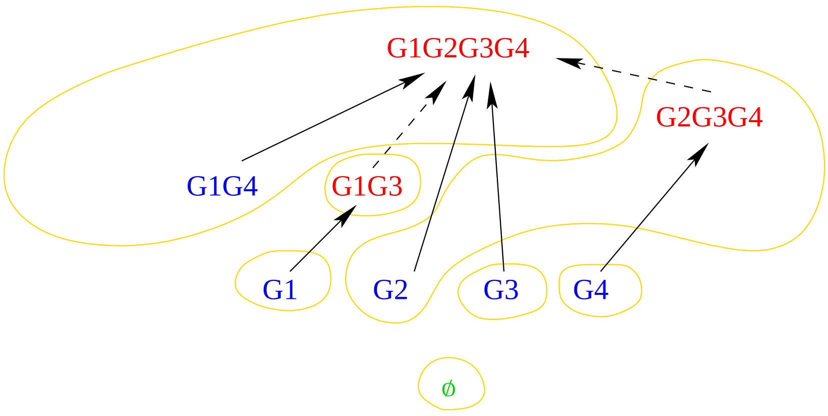


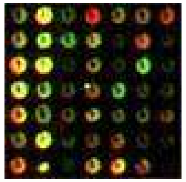
Closed-sets, free-sets and δ-free sets

- Y is free iff $\forall Z \subset Y, \psi(Z) \neq \psi(Y)$
- Generalization with δ a positif integer:
  Y is δ-free iff $\forall Z \subset Y$, $|\psi(Z)| - |\psi(Y)| > \delta$

G_1	G_2	G_3	G_4
1	0	1	0
1	1	1	1
0	1	1	1
0	0	0	1
1	0	0	0



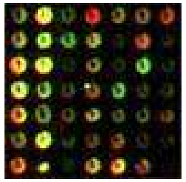




Bi-sets from δ-free sets

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

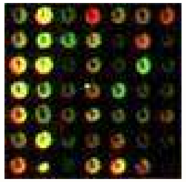
- A 1-free and its support: $\{\{E_2 E_3 E_4 TF_1\}, \{G_2\}\}$



Bi-sets from δ-free sets

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

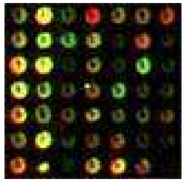
- A 1-free and its support: $\{\{E_2E_3E_4TF_1\},\{G_2\}\}$
- Its pseudo closure: $\{\{E_2E_3E_4TF_1\},\{G_1G_2G_3G_4G_5\}\}$



Bi-sets from δ-free sets

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

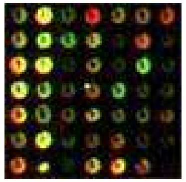
- A 1-free and its support: $\{\{E_2 E_3 E_4 TF_1\}, \{G_2\}\}$
- Its pseudo closure: $\{\{E_2 E_3 E_4 TF_1\}, \{G_1 G_2 G_3 G_4 G_5\}\}$
- Another 1-free and its support: $\{\{TF_1\ TF_2\}, \{G_5 G_6\}\}$



Bi-sets from δ-free sets

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

- A 1-free and its support: $\{\{E_2 E_3 E_4 TF_1\}, \{G_2\}\}$
- Its pseudo closure: $\{\{E_2 E_3 E_4 TF_1\}, \{G_1 G_2 G_3 G_4 G_5\}\}$
- Another 1-free and its support: $\{\{TF_1\ TF_2\}, \{G_5 G_6\}\}$
- Its pseudo closure: $\{\{TF_1\ TF_2\}, \{G_2 G_3 G_4 G_5 G_6 G_7\}\}$



Bi-sets from δ-free sets

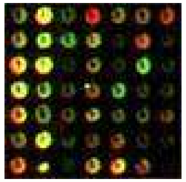
Advantages:

- Completeness
- Bounded number of false values on columns
- Maximality on columns
- Relevancy on columns
- Existence of function ϕ
- Efficient computation

Drawbacks:

- Number of false values unbounded on rows
- No maximality on rows
- No relevancy on rows
- No function ψ

$\Rightarrow$ Dissymmetrical approach



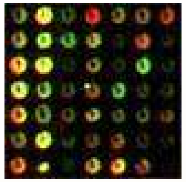
Merging formal concepts

Three constraints:

- $\mathcal{C}_{\alpha\beta}$: the number of false values on rows and columns is bounded (respectively by α and β).
- The bi-sets are *relevant:*
  - Either all identical items (w.r.t. the object set) are added when satisfiing $\mathcal{C}_{\alpha\beta}$
  - Or they are all excluded.
- The bi-sets are maximal.

Principle:

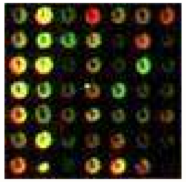
- Merging formal concepts



Merging formal concepts

Adaptation of algorithms for extracting maximal patterns (e.g., maximal itemsets)

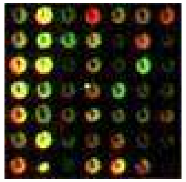
- Items: the extracted formal concepts
- Itemset: union of formal concepts
- Anti-monotonic constraint: $\mathcal{C}_{\alpha\beta}$



Merging formal concepts

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

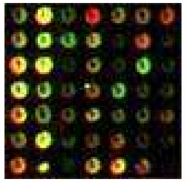
- Four formal concepts
  - $\{\{E_2 E_3 E_4 TF_1\}, \{G_3 G_4\}\}$



Merging formal concepts

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

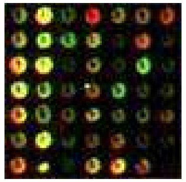
- Four formal concepts
  - $\{\{E_2E_3E_4TF_1\},\{G_3G_4\}\}$
  - $\{\{E_3E_4TF_1TF_2\},\{G_4G_5\}\}$



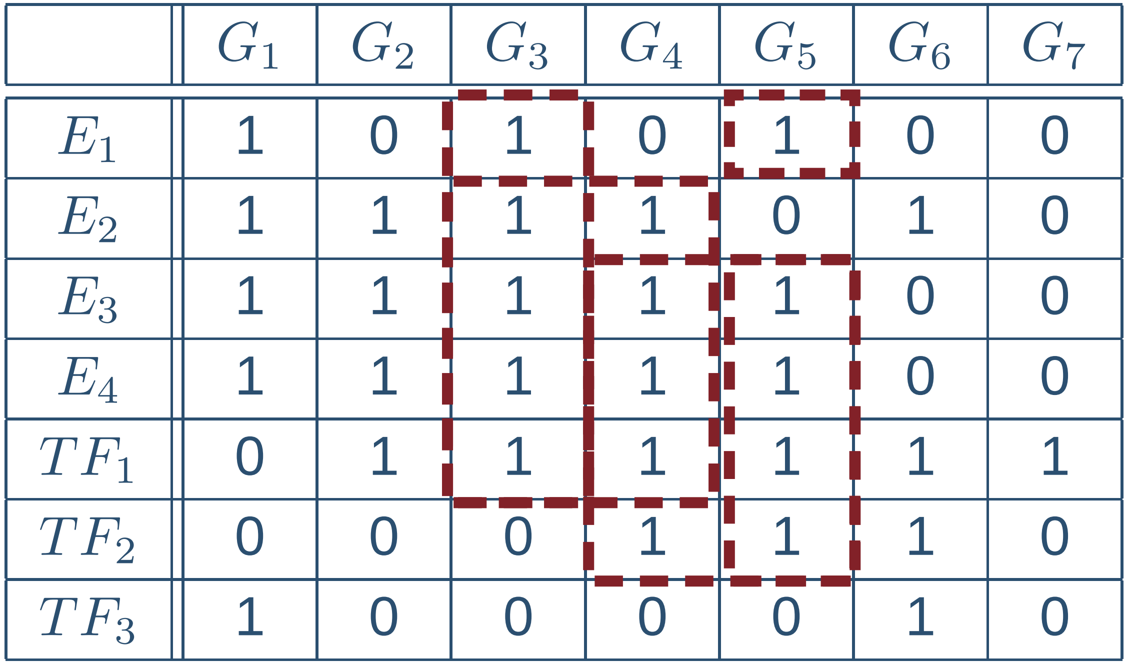
Merging formal concepts

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

- Four formal concepts
  - $\{\{E_2E_3E_4TF_1\},\{G_3G_4\}\}$
  - $\{\{E_3E_4TF_1TF_2\},\{G_4G_5\}\}$
  - $\{\{E_1E_2E_3E_4TF_1\},\{G_3\}\}$

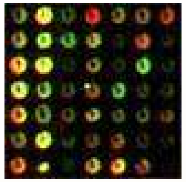


Merging formal concepts



	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

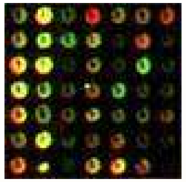
- Four formal concepts
  - $\{\{E_2E_3E_4TF_1\},\{G_3G_4\}\}$
  - $\{\{E_3E_4TF_1TF_2\},\{G_4G_5\}\}$
  - $\{\{E_1E_2E_3E_4TF_1\},\{G_3\}\}$
  - $\{\{E_1E_3E_4TF_1TF_2\},\{G_5\}\}$



Merging formal concepts

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

- Four formal concepts
  - $\{\{E_2E_3E_4TF_1\},\{G_3G_4\}\}$
  - $\{\{E_3E_4TF_1TF_2\},\{G_4G_5\}\}$
  - $\{\{E_1E_2E_3E_4TF_1\},\{G_3\}\}$
  - $\{\{E_1E_3E_4TF_1TF_2\},\{G_5\}\}$
- Merging: $\{\{E_1E_2E_3E_4TF_1TF_2\},\{G_3G_4G_5\}\}$



Merging formal concepts

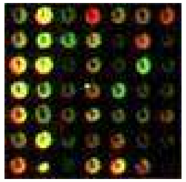
	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

Two patterns resulting from the merging process:

- $\{\{E_1E_2E_3E_4TF_1TF_2\},\{G_3G_4G_5\}\}$ ━━━━
- $\{\{E_1E_2E_3E_4TF_1\},\{G_1G_3G_5\}\}$ ╍╍╍╍

$\Rightarrow$ No Galois connection

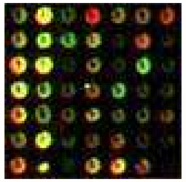
There are bi-sets (O,I) and (O',I') such that $O \subset O'$ and $I' \not\subset I$



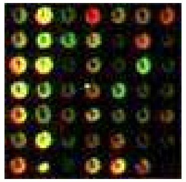
Merging formal concepts

Advantages:
- The number of false values is bounded on rows and columns
- Maximality on both dimensions
- Relevancy on rows and columns

Drawbacks:
- The extraction process is incomplete
- There is no function ϕ nor function ψ
- The extraction process is quite hard

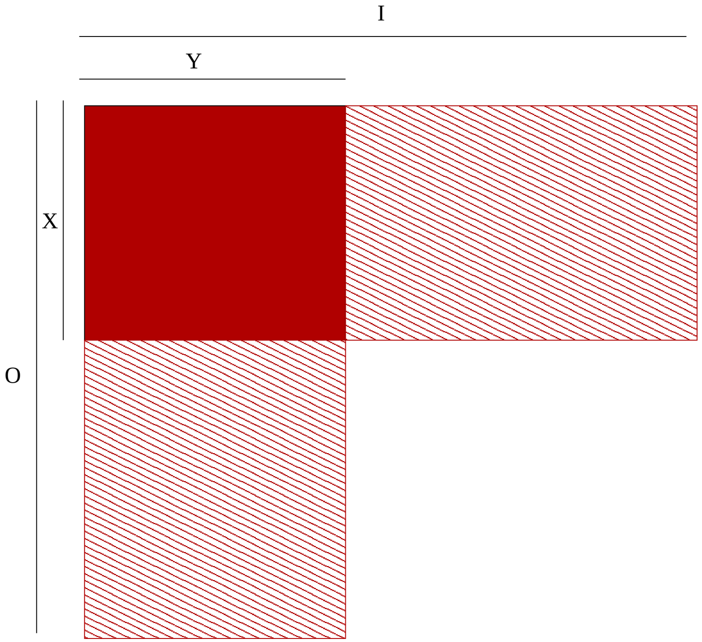


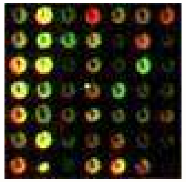
Relevant and dense bi-sets in noisy data



Principle

- Fault-tolerant bi-sets
  - Bi-sets with a bounded number of false values per row and column.
- Relevancy
  - The bi-set outside rows and columns contain more false values than the inside ones.





A generalization of formal concepts

Notations

- Number of false values of x on Y (where x is a row element and Y a column set):

$$\mathcal{Z}_l(x, Y) = \sharp\{y \in Y | (x, y) \notin \mathbf{r}\}$$

- Number of false values of y on X:

$$\mathcal{Z}_c(y, X) = \sharp\{x \in X | (x, y) \notin \mathbf{r}\}$$

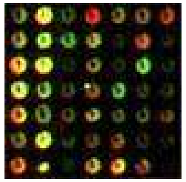
Formal concepts: a bi-set (X, Y) is a formal concept in $\mathbf{r}$ iff

- There are no false values inside the concept

$$\forall x \in X, \mathcal{Z}_l(x, Y) = 0, \;\; \forall y \in Y, \mathcal{Z}_c(y, X) = 0$$

- There is at least one false value on each outside rows and columns

$$\forall x \in \mathcal{S} \setminus X, \mathcal{Z}_l(x, Y) \geq 1, \;\; \forall y \in \mathcal{G} \setminus Y, \mathcal{Z}_c(y, X) \geq 1$$



A generalization of formal concepts

A bi-set (X, Y) is a relevant and dense bi-set in $\mathbf{r}$ iff

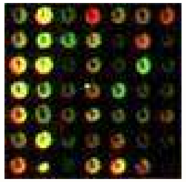
$\mathcal{C}_d$: The number of false values are bounded inside the pattern

$$\forall x \in X, \mathcal{Z}_l(x, Y) \leq \alpha, \quad \forall y \in Y, \mathcal{Z}_c(y, X) \leq \alpha'$$

$\mathcal{C}_r$: The external elements contain at least δ false values in addition compared to the interior elements

$$\forall e \in \mathcal{O} \setminus X, \forall x \in X, \mathcal{Z}_l(e, Y) \geq \mathcal{Z}_l(x, Y) + \delta$$

$$\forall e' \in \mathcal{I} \setminus Y, \forall y \in Y, \mathcal{Z}_c(e', X) \geq \mathcal{Z}_c(y, X) + \delta'$$

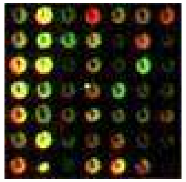


DR-*bi-sets on our toy example*

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
E_1	1	0	1	0	1	0	0
E_2	1	1	1	1	0	1	0
E_3	1	1	1	1	1	0	0
E_4	1	1	1	1	1	0	0
TF_1	0	1	1	1	1	1	1
TF_2	0	0	0	1	1	1	0
TF_3	1	0	0	0	0	1	0

$\alpha = \alpha' = 1$ and $\delta = \delta' = 1$

$\{\{E_1 E_2 E_3 E_4 TF_1 TF_2\}, \{G_3 G_4 G_5\}\}$

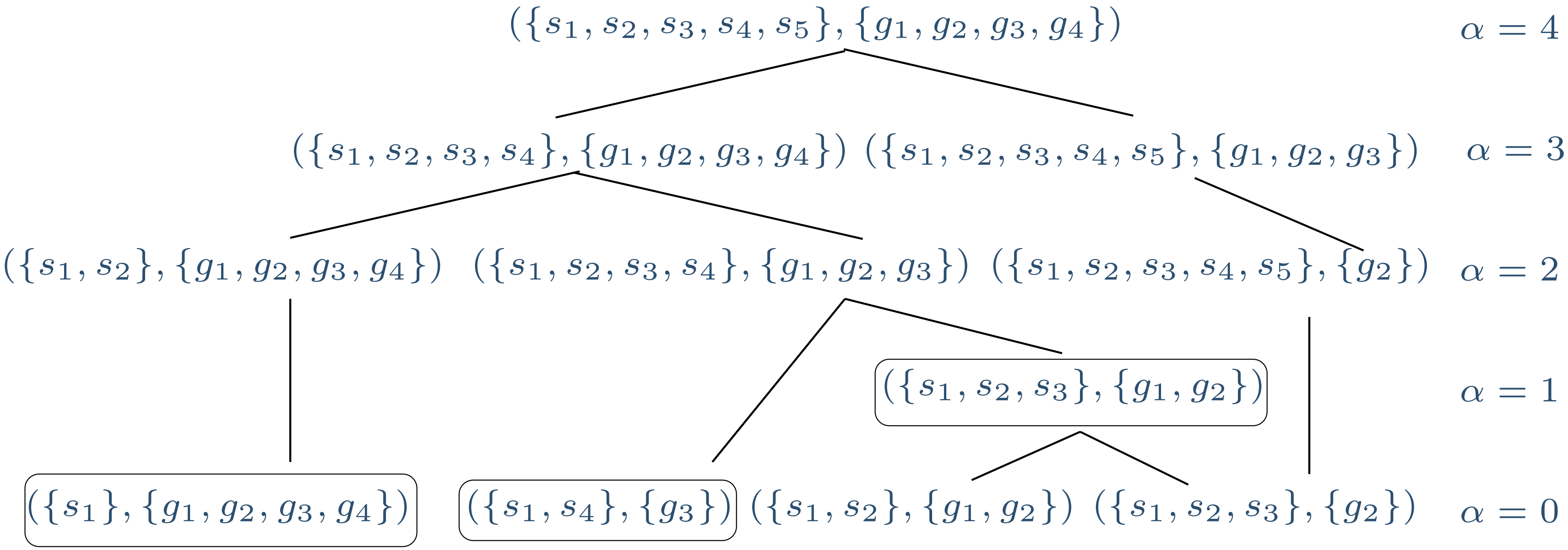


A collection of bi-sets satisfying $\mathcal{C}_d$ and $\mathcal{C}_r$

	g_1	g_2	g_3	g_4
s_1	1	1	1	1
s_2	1	1	0	0
s_3	0	1	0	0
s_4	0	0	1	0
s_5	0	0	0	0

$\alpha = 5$ and $\alpha' = 4$

$\delta = 1$ and $\delta' = 1$



$(\{s_1, s_2, s_3, s_4, s_5\}, \{g_1, g_2, g_3, g_4\})$ $\alpha = 4$

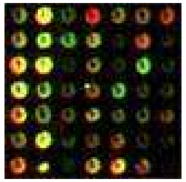
$(\{s_1, s_2, s_3, s_4\}, \{g_1, g_2, g_3, g_4\})$ $(\{s_1, s_2, s_3, s_4, s_5\}, \{g_1, g_2, g_3\})$ $\alpha = 3$

$(\{s_1, s_2\}, \{g_1, g_2, g_3, g_4\})$ $(\{s_1, s_2, s_3, s_4\}, \{g_1, g_2, g_3\})$ $(\{s_1, s_2, s_3, s_4, s_5\}, \{g_2\})$ $\alpha = 2$

$(\{s_1, s_2, s_3\}, \{g_1, g_2\})$ $\alpha = 1$

$(\{s_1\}, \{g_1, g_2, g_3, g_4\})$ $(\{s_1, s_4\}, \{g_3\})$ $(\{s_1, s_2\}, \{g_1, g_2\})$ $(\{s_1, s_2, s_3\}, \{g_2\})$ $\alpha = 0$

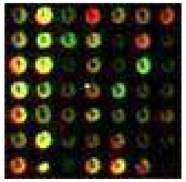
DR-bi-sets: maximal bi-sets which satisfy $\mathcal{C}_r$ and $\mathcal{C}_d$.



DR-*bi-sets*

- When $\alpha = \alpha' = 0$ and $\delta = \delta' = 1$, DR-bi-sets are the formal concepts.
- When $\alpha > 0, \alpha' > 0, \delta > 0, \delta' > 0$ and selecting maximal bi-sets: DR-bi-sets are fault tolerant patterns.

- DR-bi-sets increase with α and α'
- There exists two functions which link sets of DR-bi-sets ($\delta > 0$)
- These functions are decreasing (α fixed)

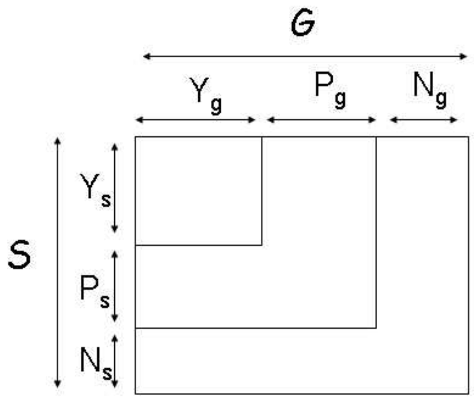


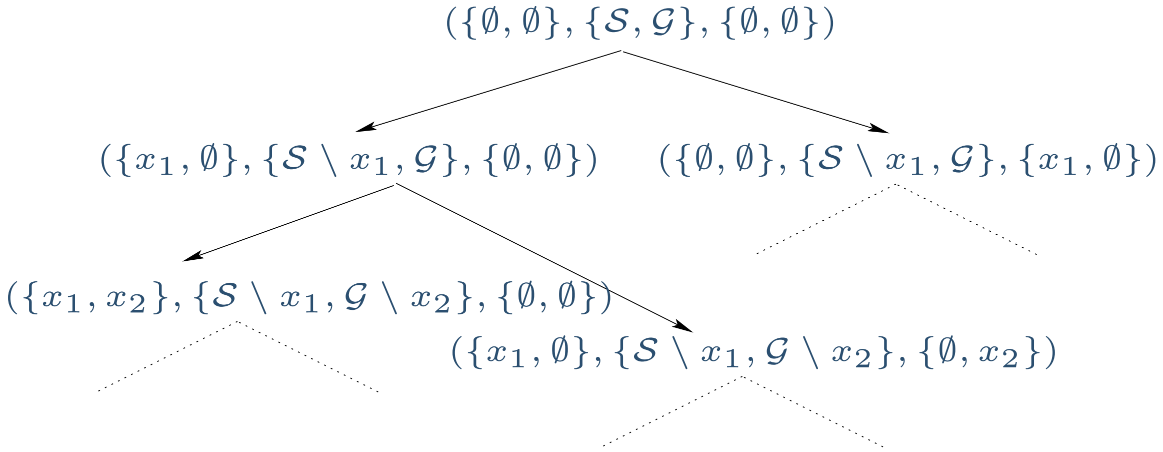
A complete algorithm: DR-MINER

Adapting DUAL-MINER principle

A candidate (Y, P, N) is such that:

- $Y = (Y_S, Y_G)$ (Yes bi-set) contains the elements which belong to the bi-set and all its children.
- $N = (N_S, N_G)$ (No bi-set) contains the elements which do not belong to the bi-set and its children.
- $P = (P_S, P_G)$ (Potential bi-set) contains the elements which have not been processed so far.





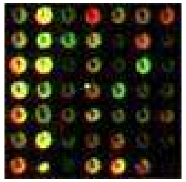
$(\{\emptyset,\emptyset\},\{\mathcal{S},\mathcal{G}\},\{\emptyset,\emptyset\})$

$(\{x_1,\emptyset\},\{\mathcal{S}\setminus x_1,\mathcal{G}\},\{\emptyset,\emptyset\})$

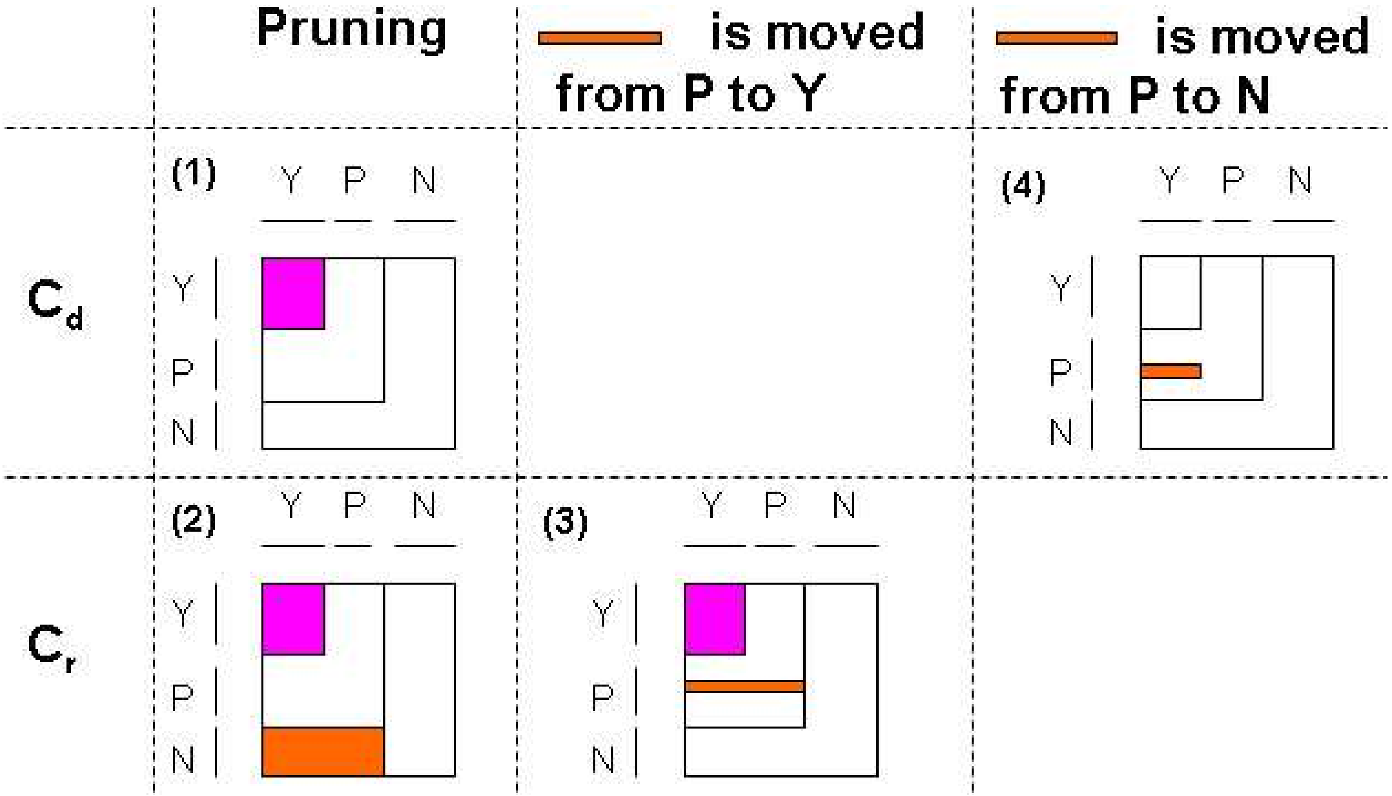
$(\{\emptyset,\emptyset\},\{\mathcal{S}\setminus x_1,\mathcal{G}\},\{x_1,\emptyset\})$

$(\{x_1,x_2\},\{\mathcal{S}\setminus x_1,\mathcal{G}\setminus x_2\},\{\emptyset,\emptyset\})$

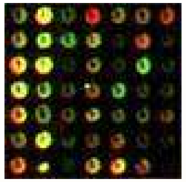
$(\{x_1,\emptyset\},\{\mathcal{S}\setminus x_1,\mathcal{G}\setminus x_2\},\{\emptyset,x_2\})$



Constraint checking

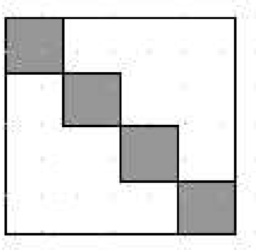


- All anti-monotonic constraints can be propagated as $\mathcal{C}_d$.
- All monotonic constraints can be used to prune the search space if they are not satisfied on $Y \cup P$.



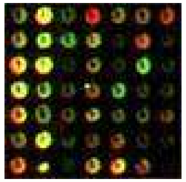
Experimental validation

Artificial dataset



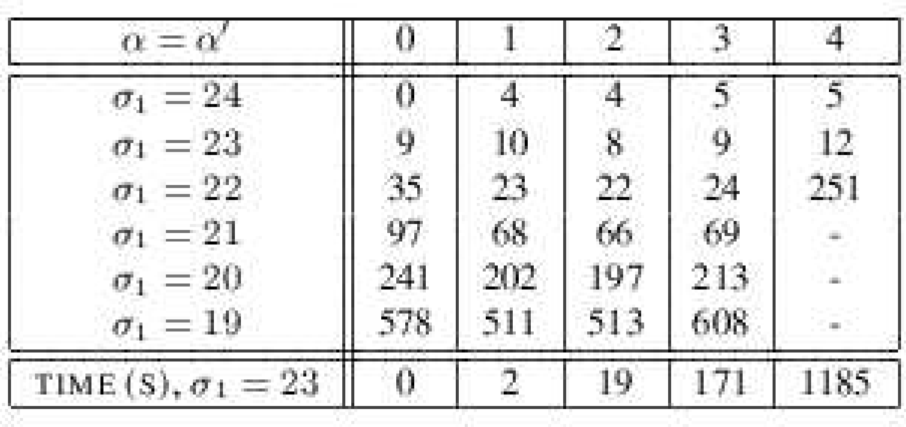
DR-bi-sets containing at least 4 elements on both dimensions, and with $\delta = 3$.

α		0		1		2		3	
	♯ concepts	Mean	sd	Mean	sd	Mean	sd	Mean	sd
5%	51.63	0	0	0.55	0.51	3	0	4	0
10%	141.53	0	0	0	0	1.6	0.6	2.8	0.41
15%	248.63	0	0	0	0	0	0	0.85	0.49
20%	309.05	0	0	0	0	0	0	1.1	0.45



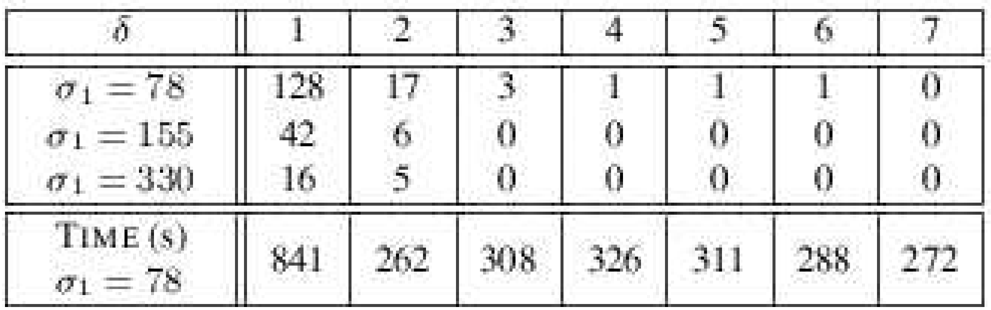
Experimental validation

Biological data (CAMDA’04) (46*3719)
Compute collections of DR-bi-sets containing at least 3 columns and σ_1 rows. $\delta = 1$.

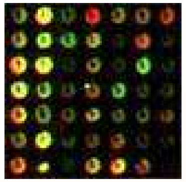


$\alpha = \alpha'$	0	1	2	3	4
$\sigma_1 = 24$	0	4	4	5	5
$\sigma_1 = 23$	9	10	8	9	12
$\sigma_1 = 22$	35	23	22	24	251
$\sigma_1 = 21$	97	68	66	69	-
$\sigma_1 = 20$	241	202	197	213	-
$\sigma_1 = 19$	578	511	513	608	-
TIME (S), $\sigma_1 = 23$	0	2	19	171	1185

UCI Internet Advertisements (3279*1555) Compute collections of DR-bi-sets containing at least σ_1 columns and $\alpha = 1$.

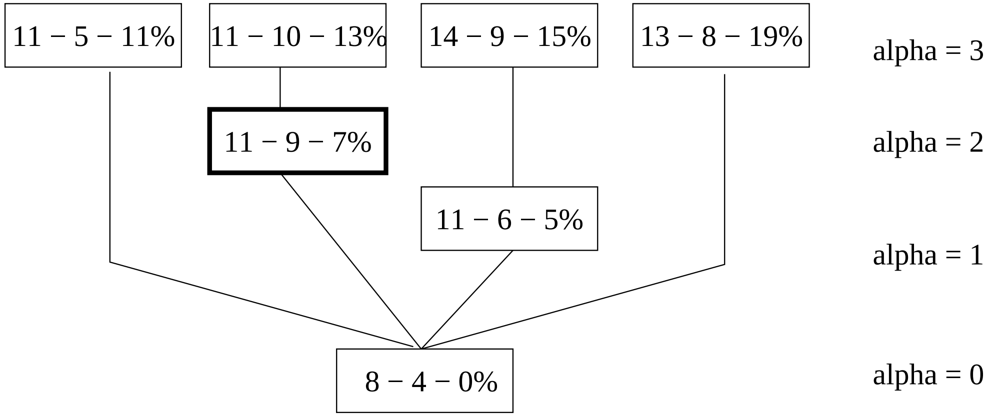


δ	1	2	3	4	5	6	7
$\sigma_1 = 78$	128	17	3	1	1	1	0
$\sigma_1 = 155$	42	6	0	0	0	0	0
$\sigma_1 = 330$	16	5	0	0	0	0	0
TIME (S) $\sigma_1 = 78$	841	262	308	326	311	288	272



Experimental validation

- Biological data: CAMDA’04 dataset (46*3719)
- Formal concept extension:
  - ❖ A formal concept containing 8 biological situations relating to the “ring” phase and 4 genes among which 3 are known to be active in this phase.



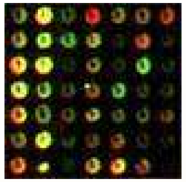
11 − 5 − 11%
11 − 10 − 13%
14 − 9 − 15%
13 − 8 − 19%
alpha = 3

11 − 9 − 7%
alpha = 2

11 − 6 − 5%
alpha = 1

8 − 4 − 0%
alpha = 0

Each triplet represents the number of experiments, genes and relative weak density of 0 values.



DR-*bi-sets properties*

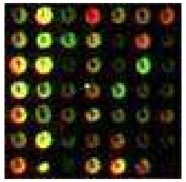
Advantages:

- Completeness
- Bounded number of false values on rows and columns
- Relevancy of the extracted patterns
- Functions

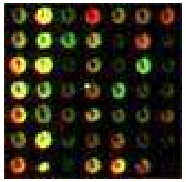
Drawbacks:

- Quite hard to compute

An incomplete use concerns formal concept extension.



Conclusion



Conclusion and future work

- Formal concepts are relevant to extract transcription modules
- D-MINER enables to compute large formal concepts on both dimensions
- Formal concepts are not robust to noise
- DR-bi-sets are well defined to handle data noise
  - They are hard to compute with DR-MINER
  - The computing efficiency can be increased by adding a stringent conjonction of monotonic and anti-monotonic constraints