



Schwarz wald, March 13th, 2004

Constraint-based mining of concepts from gene expression data

Jérémy BESSON

INSA Lyon - LIRIS CNRS FRE 2672

INRA - INSERM U449

France





Our biological problem

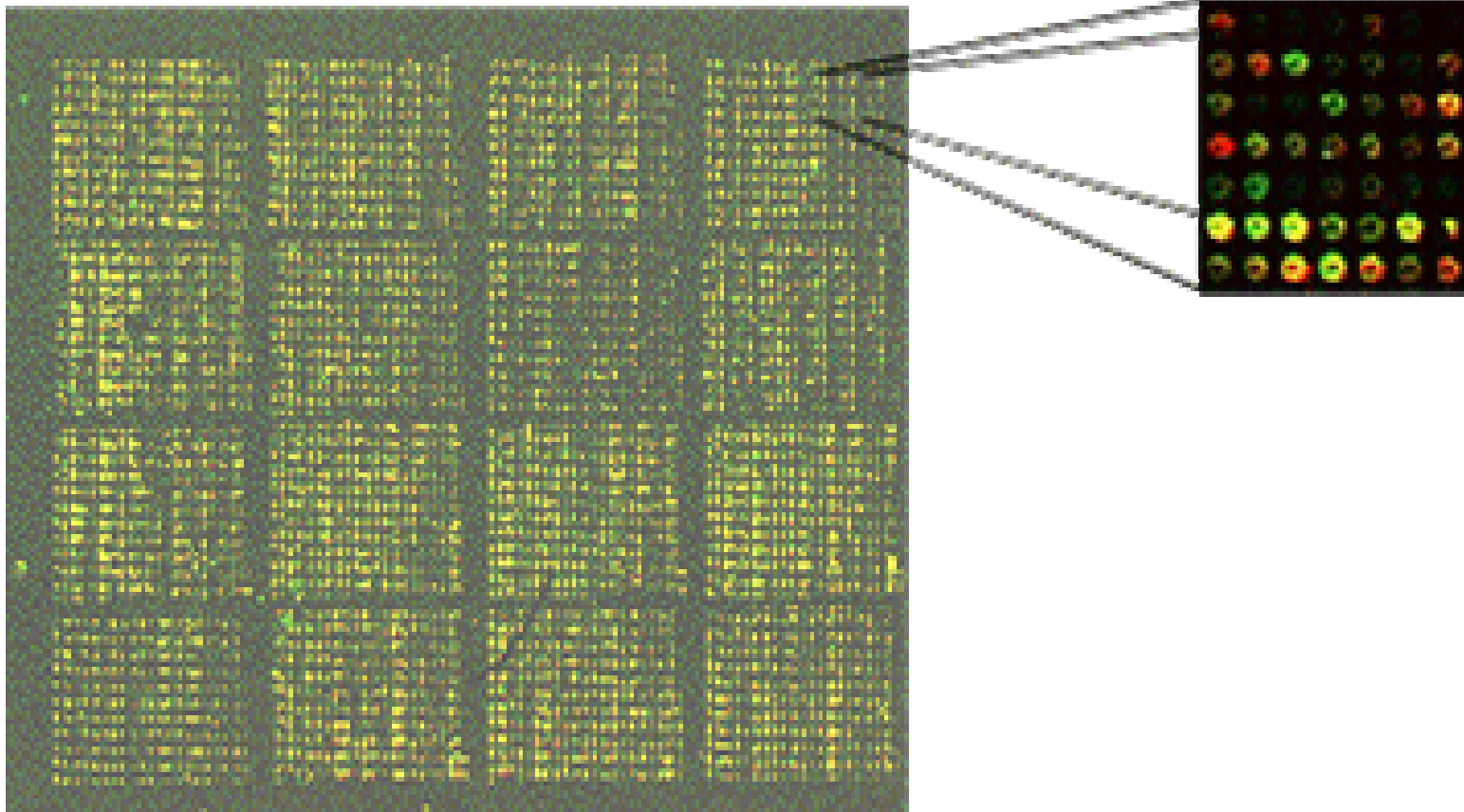
- ◆ **Biological questions:** find informations about
 - regulatory ways
 - gene functions
 - co-expressed genes and transcription modules

in type 2 diabetes disease

- ◆ DNA Microarrays, bibliography, Bio-informatic Databases, ...
-



DNA microarray





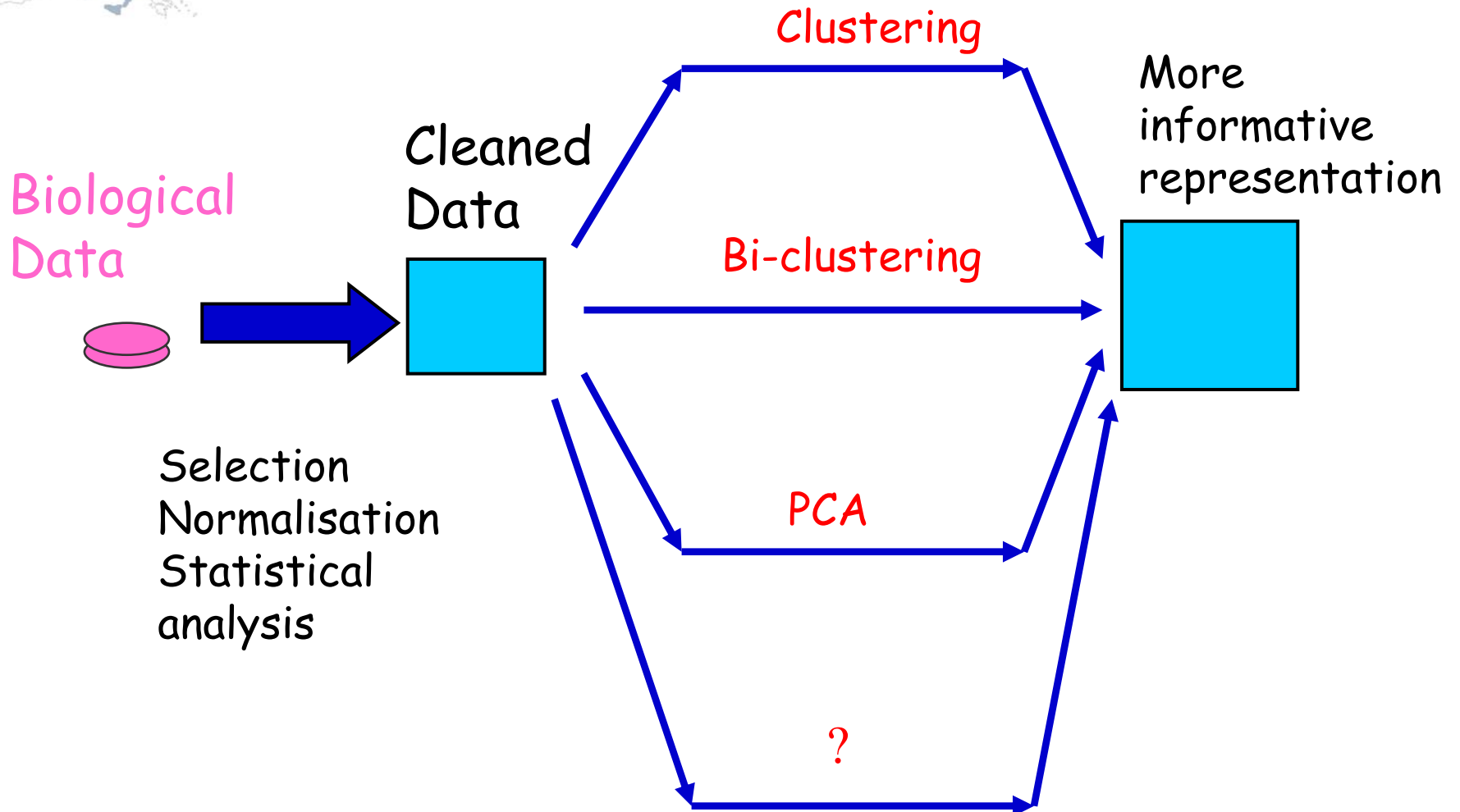
Our Data

- ◆ 5 microarrays for healthy persons and 5 microarrays for type 2 diabetic patients.
- ◆ Each spot (~gene) represents the influence of insulin (before and after an insulin clamp)





The classical process





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

0 Sets of situations

1 Sets of genes



Looking for biologically relevant sets

Sets of genes

- Sets of genes which are co-regulated (synexpression groups)
- Sets of genes which have the same transcription factors i.e., putative regulatory ways
- Sets of genes which are up-regulated for diabetic patients but not for healthy people

Associated sets of situations

Using formal concepts (Wille 82)



First approach

Tanks to Galois connection:

It is possible to compute concepts from closed sets of objects (situations) or from closed sets of items (genes)

In a $10 * 8000$ matrix:

a simple transposition allows to extract all concepts with any (frequent) closed set extractor (Riout et al. 2003)



Second approach

But it is not sufficient

- ◆ More and more data (in immediate future)
- ◆ Dense boolean contexts
- ◆ We need to enrich our data to increase biological relevancy (higher dimensionality)

In our enriched $100 * 300$ matrix:

impossible to extract all the concepts with known algorithms



Constraint-based mining

We need to use constraints on concepts (S,G)

- ◆ To reduce the search space
- ◆ To increase the biological relevancy

Monotonic constraints seem useful

$$|S| > \alpha \text{ and } |G| > \beta$$





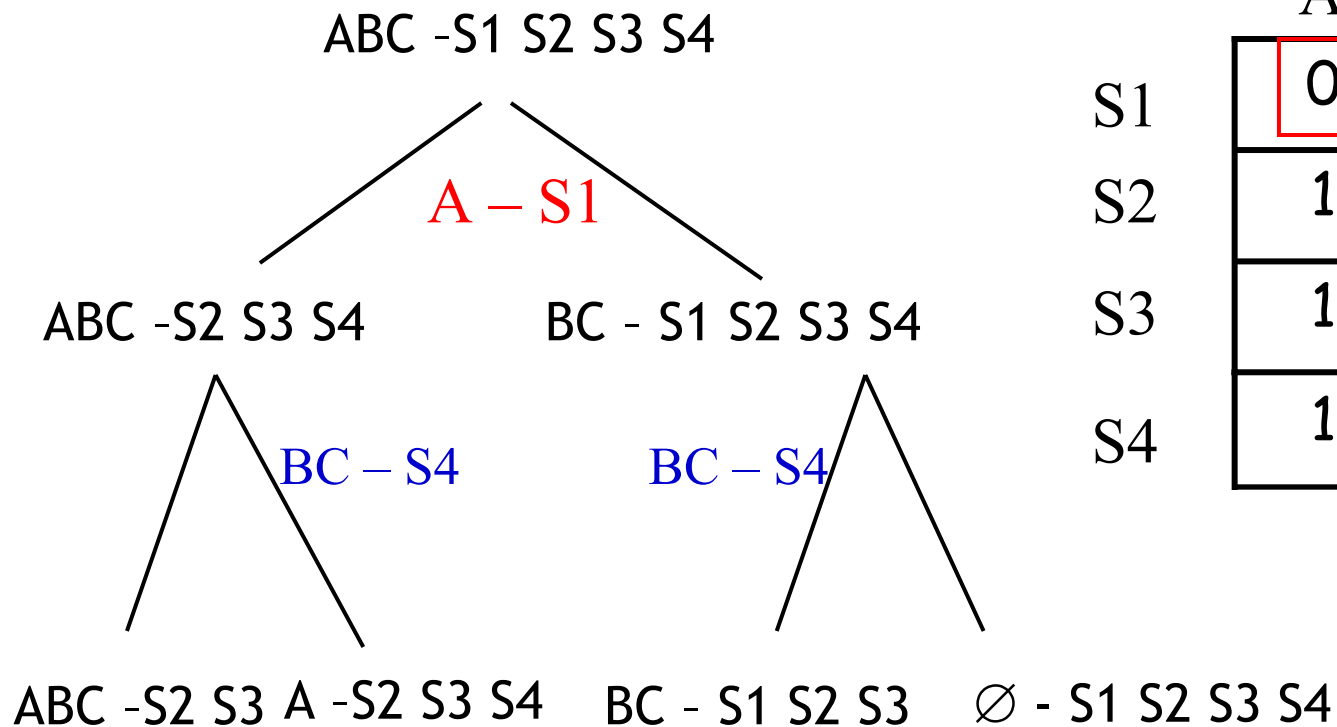
D-miner (MineSweeper): Basic idea

	G1	G2	G3	
S1	0	1	1	-----
S2	1	1	1	-----
S3	1	1	1	-----

The diagram shows a 3x3 grid of cells. The columns are labeled G1, G2, and G3. The rows are labeled S1, S2, and S3. The values in the cells are: (S1, G1) = 0, (S1, G2) = 1, (S1, G3) = 1, (S2, G1) = 1, (S2, G2) = 1, (S2, G3) = 1, (S3, G1) = 1, (S3, G2) = 1, (S3, G3) = 1. A red line starts at the top of the G2 column, goes down, then right across the top of the G2 and G3 cells, then down again. Dashed lines extend from the right and bottom of the grid.



An exemple



	A	B	C
S1	0	1	1
S2	1	1	1
S3	1	1	1
S4	1	0	0



Pseudo-algorithm

- ◆ Compute the cutters : H (O-rectangles)
- ◆ Start with (O, I)

Cut recursively the bi-sets (T, G) with the cutter
 $H_i = (t_{ci}, g_{ci})$ into:

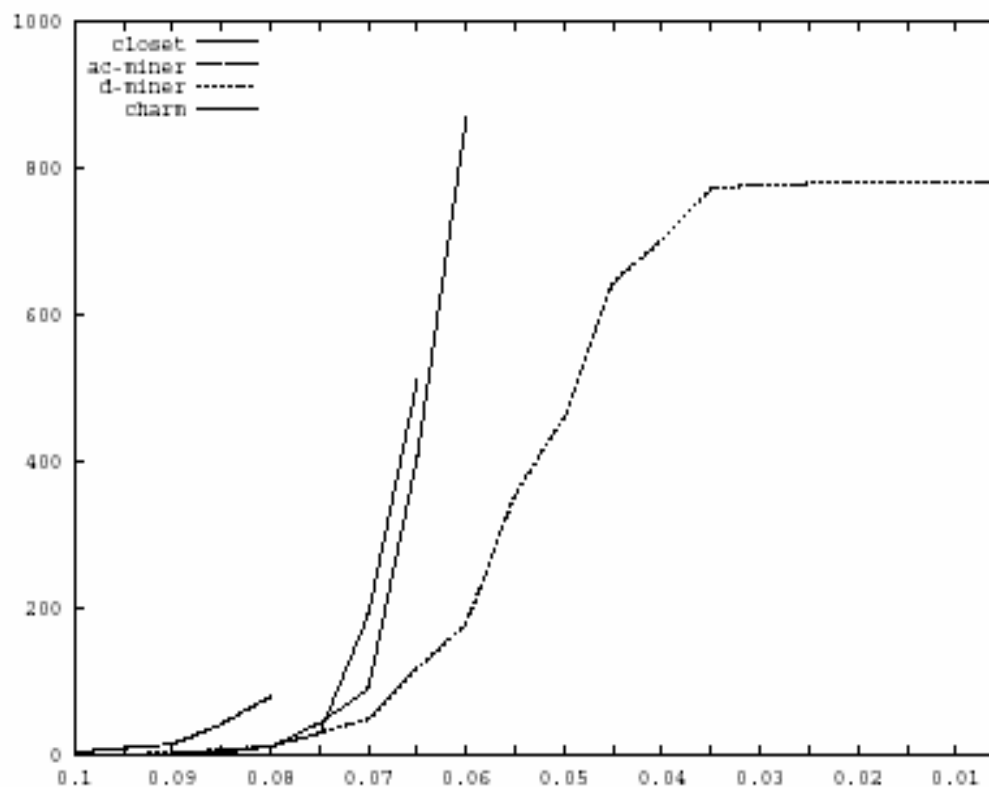
$(T \setminus t_{ci}, G)$

$(T, G \setminus g_{ci})$

} Push monotonic constraints
and
Uniqueness constraint



Experimental validation



PAKDD 2004 : Besson, Robardet, Boulicaut : Constraint-based mining of formal concepts in transcriptional data. To appear.



Towards a biological validation?

The 17 genes in {Hs.2002, Hs.184, ..., Hs.302649} are associated to 8 transcription factors {c-Ets-, MZF1, ..., AML-1a} for 4 biological situations that concern healthy subjects

This concept is a priori interesting because it contains genes which are either up-regulated or down-regulated after insulin stimulation based on the homology of their promotor DNA sequences (common TF can regulate these genes either positively or negatively).



Using extracted concepts

How to use concepts to support biological knowledge discovery?

Short term:

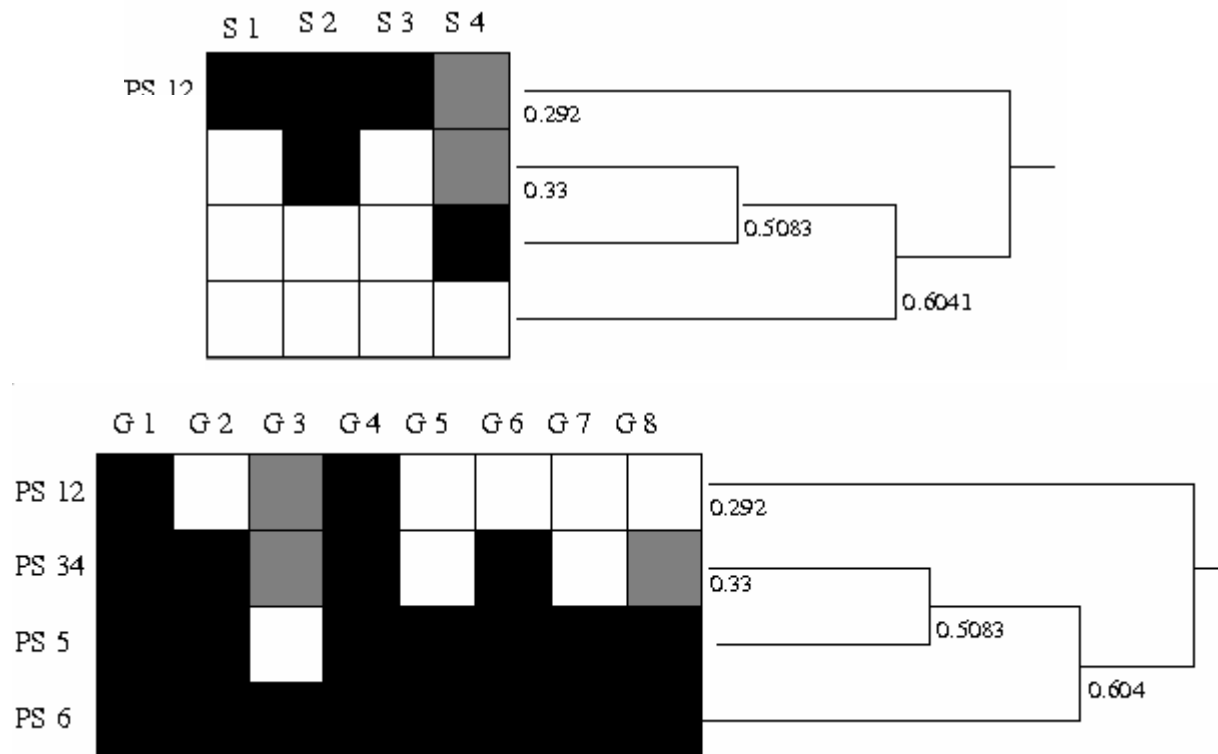
- Post-processing tools (e.g., browsing/selecting concepts, clustering concepts, visualization)

Long term:

- Towards inductive databases
-



Hierarchical clustering of concepts



PARMA 2004 : Robardet, Pensa, Besson, Boulicaut : using classification and visualisation on pattern databases for gene expression data analysis



Perspectives

