KATHOLIEKE UNIVERSITEIT LEUVEN

Identification of overlapping biclusters using Probabilistic Relational Models

Tim Van den Bulcke
Hui Zhao
Kristof Engelen
Bart De Moor
Kathleen Marchal

PMCB Workshop
Thursday, 26 July, 2007.

Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- Biclustering and biology
- Probabilistic Relational Models
- *Pro*Bic biclustering model
- Algorithm
- Results
- Conclusion

Overview

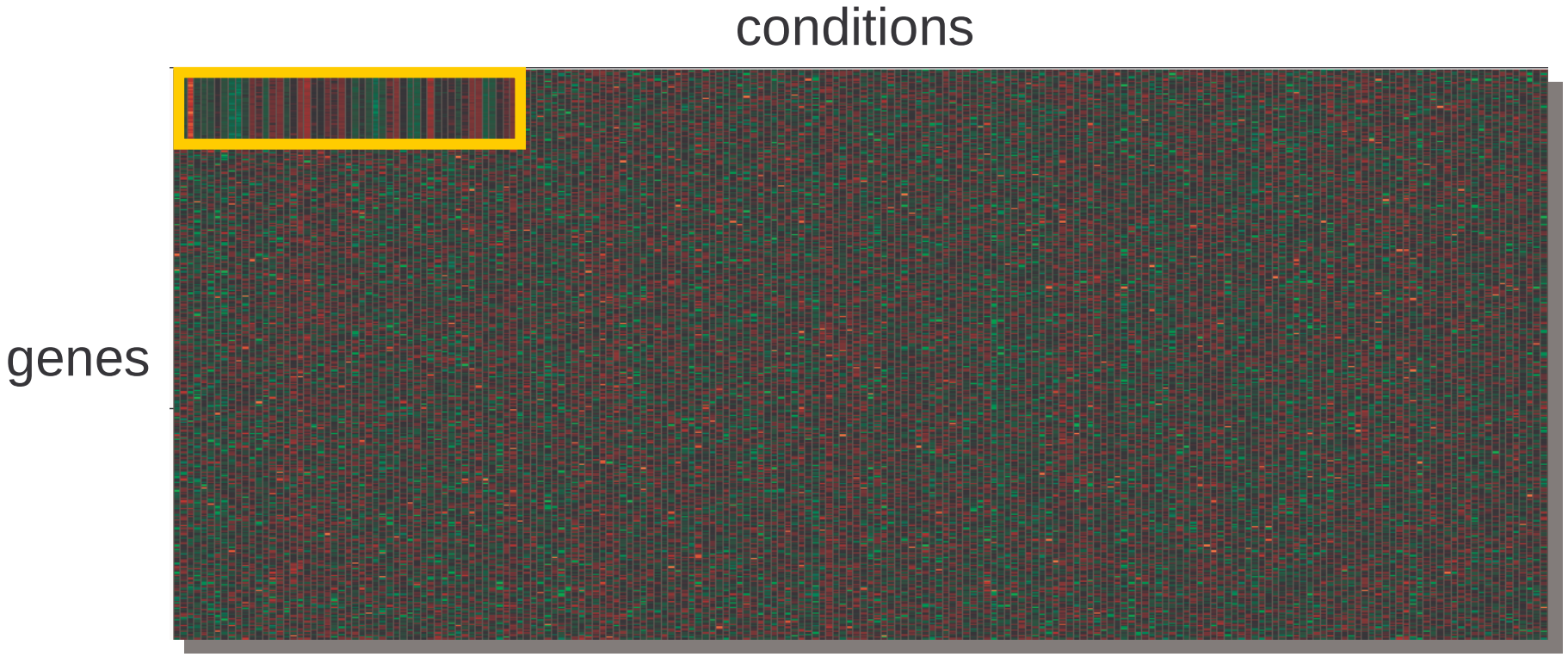
KATHOLIEKE UNIVERSITEIT LEUVEN

- Biclustering and biology
- Probabilistic Relational Models
- *Pro*Bic biclustering model
- Algorithm
- Results
- Conclusion

Biclustering and biology

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Definition in the context of gene expression data:**
  A **bicluster** is a subset of genes which show a similar expression profile under a *subset* of conditions.

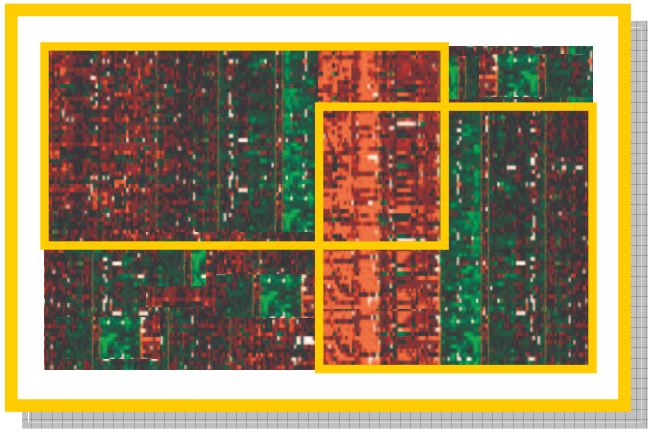


Biclustering and biology

KATHOLIEKE UNIVERSITEIT LEUVEN

Why bi-clustering?*

- Only a small set of the genes participates in a cellular process.
- A cellular process is active only in a subset of the conditions.
- A single gene may participate in multiple pathways that may or may not be coactive under all conditions.



* From: Madeira et al. (2004) *Biclustering Algorithms for Biological Data Analysis: A Survey*

Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- Biclustering and biology
- Probabilistic Relational Models
- *Pro*Bic biclustering model
- Algorithm
- Results
- Conclusion

Probabilistic Relational Models (PRMs)

KATHOLIEKE UNIVERSITEIT LEUVEN

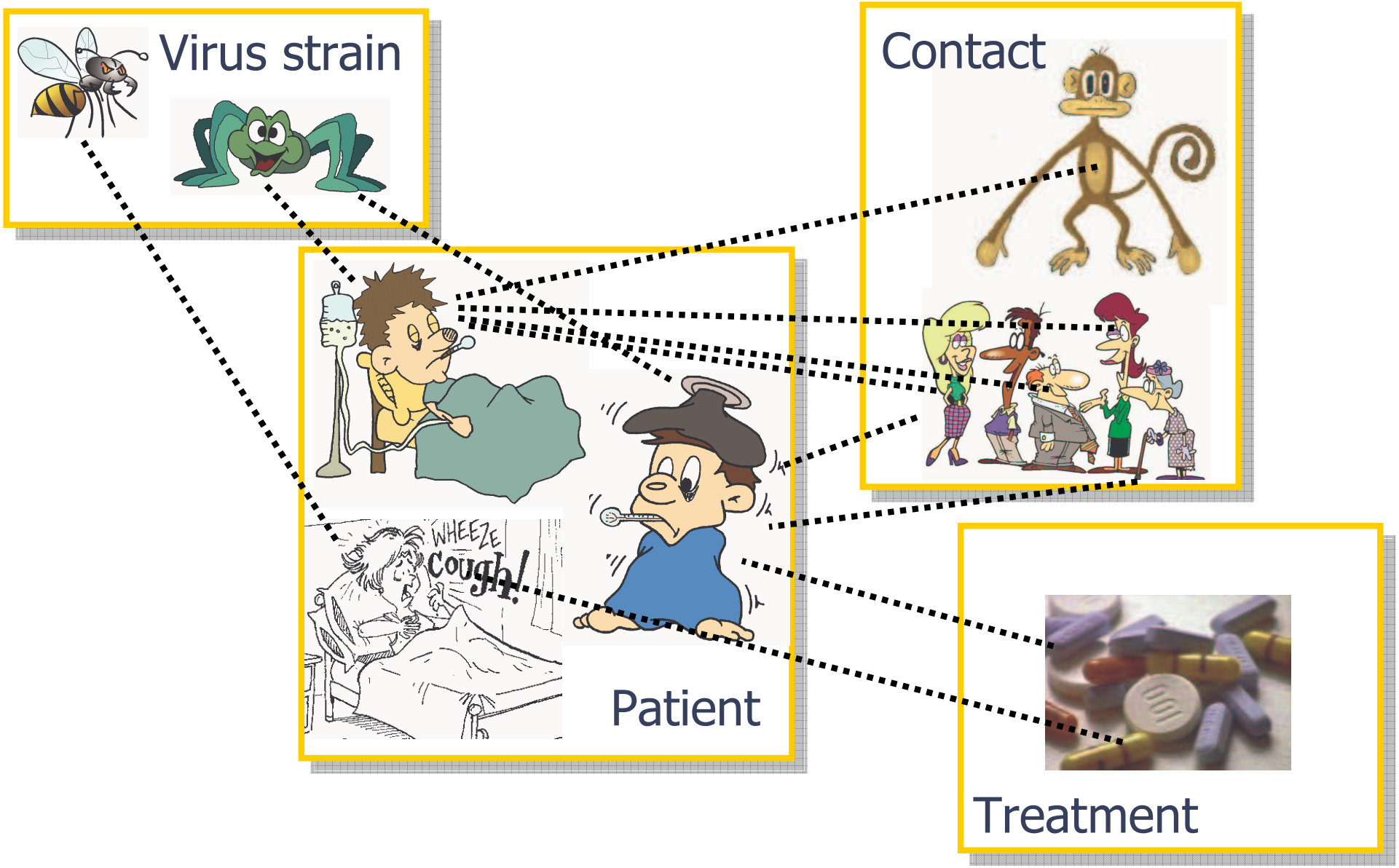
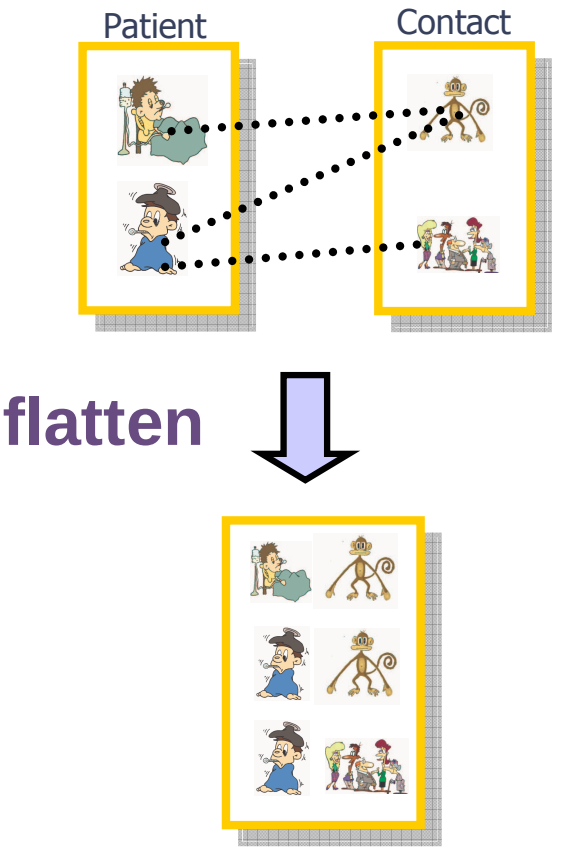


Image: free interpretation from Segal et al. Rich probabilistic models

Probabilistic Relational Models (PRMs)

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Traditional approaches "flatten" relational data**
  - Causes bias
  - Centered around one view of the data
  - Loose relational structure
- **PRM models**
  - Extension of Bayesian networks
  - Combine advantages of probabilistic reasoning with relational logic



Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- Biclustering and biology
- Probabilistic Relational Models
- *Pro*Bic biclustering model
- Algorithm
- Results
- Conclusion

*Pro*Bic biclustering model: notation

KATHOLIEKE UNIVERSITEIT LEUVEN

- g: gene
- c: condition
- e: expression
- $g.B_k$: gene-bicluster assignment for gene *g* to bicluster *k*
- $c.B_k$: condition-bicluster assignment for condition *c* to bicluster *k*
- *e.Level:* expression level value
- *G, C, E (capital letters):* set of <u>all</u> genes, conditions, expression levels resp.
- $\mu_{g.B,\ c.B,\ c}$, $\sigma_{g.B,\ c.B,\ c}$: Normal distribution parameters for condition c, with gene-bicluster and condition-bicluster assignments *g.B* and c*.B*

*Pro*Bic biclustering model

- **Dataset instance**

Gene

ID	B1	B2
g1	? *(0 or 1)*	? *(0 or 1)*
g2	? *(0 or 1)*	? *(0 or 1)*

Condition

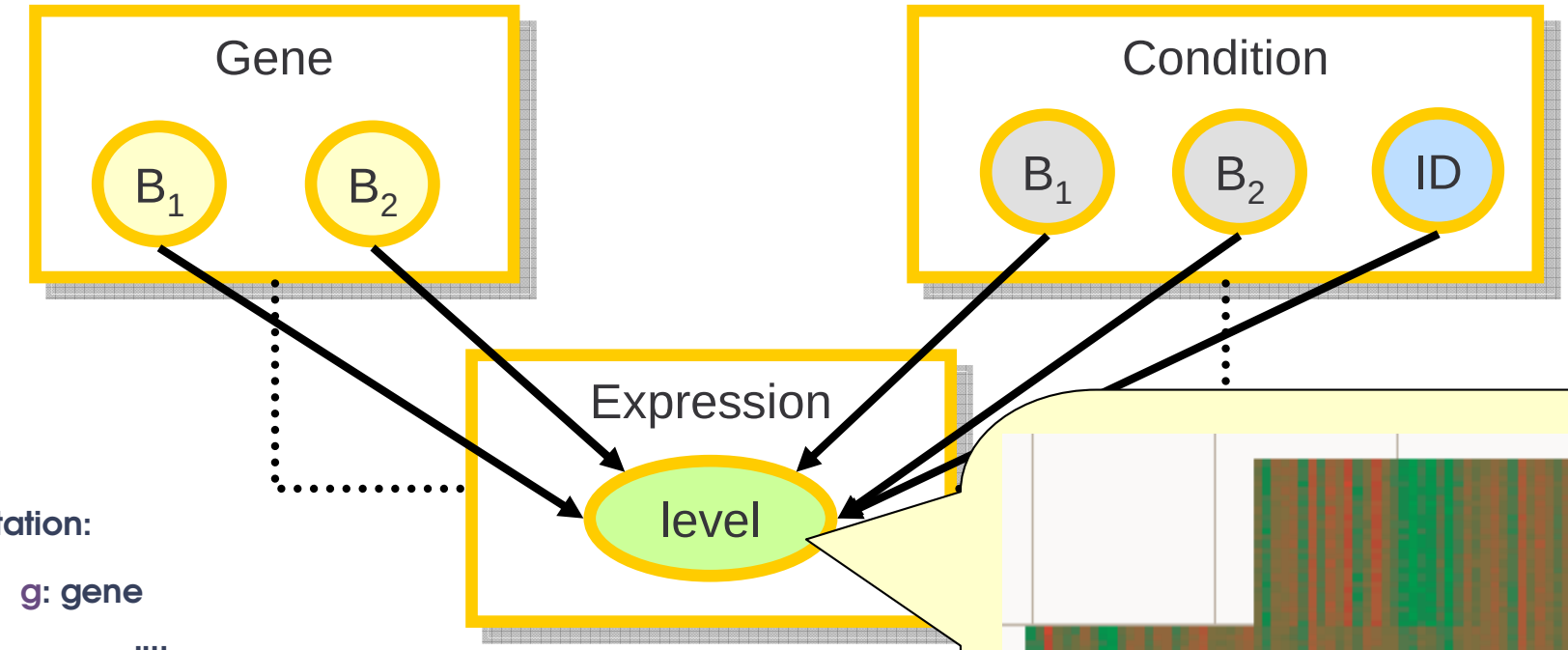
ID	B1	B2
c1	? *(0 or 1)*	? *(0 or 1)*
c2	? *(0 or 1)*	? *(0 or 1)*

Expression

g.ID	c.ID	level
g1	c1	-2.4
g1	*c2*	*(missing value)*
g2	c1	1.6
g2	c2	0.5

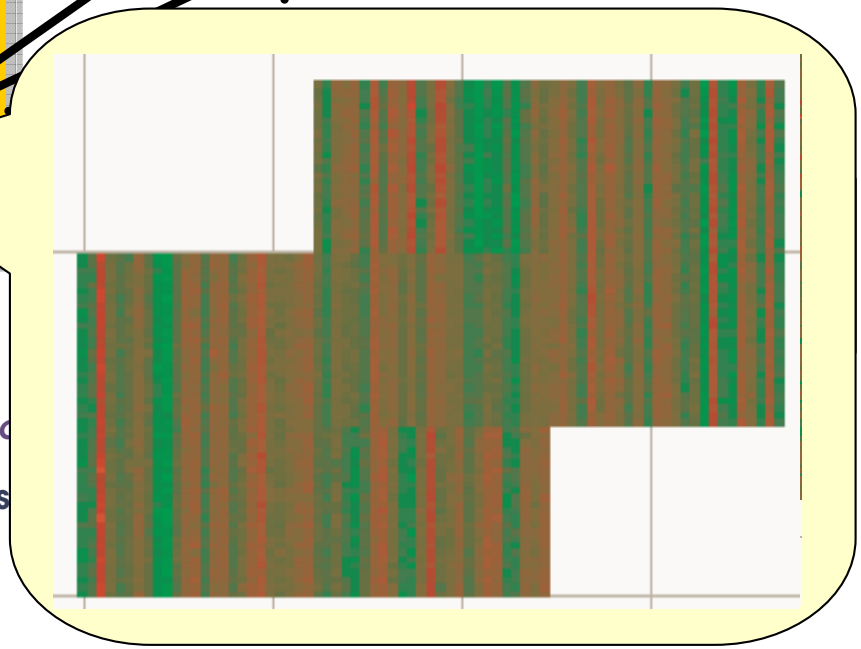
*Pro*Bic biclustering model

- **Relational schema and PRM model**



Notation:

- g: gene
- c: condition
- e: expression
- $g.B_k$: gene-bicluster assignment for gene *g* to bicluster *k* (*0 c*
- $c.B_k$: condition-bicluster assignment for condition *c* to biclus
- *e.Level:* expression level value (continuous, known)



*Pro*Bic biclustering model

KATHOLIEKE UNIVERSITEIT LEUVEN

Database instance

Gene

ID	B1	B2
g1	? *(0 or 1)*	? *(0 or 1)*
g2	? *(0 or 1)*	? *(0 or 1)*

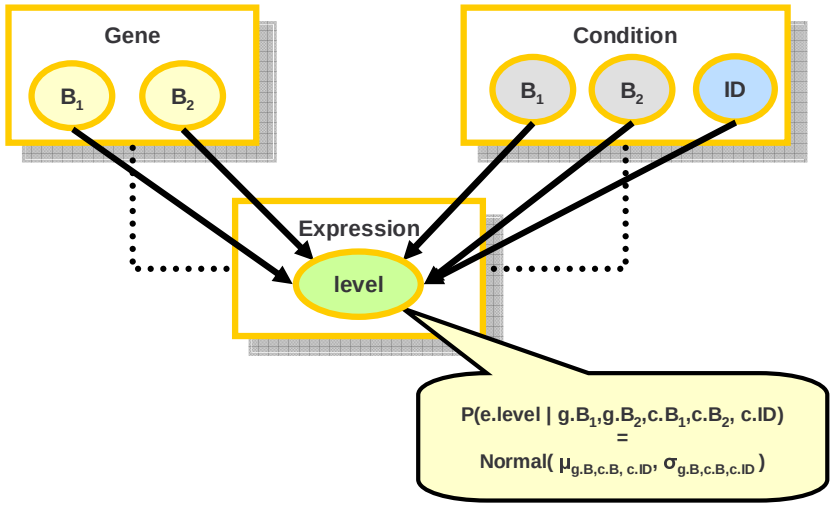
Condition

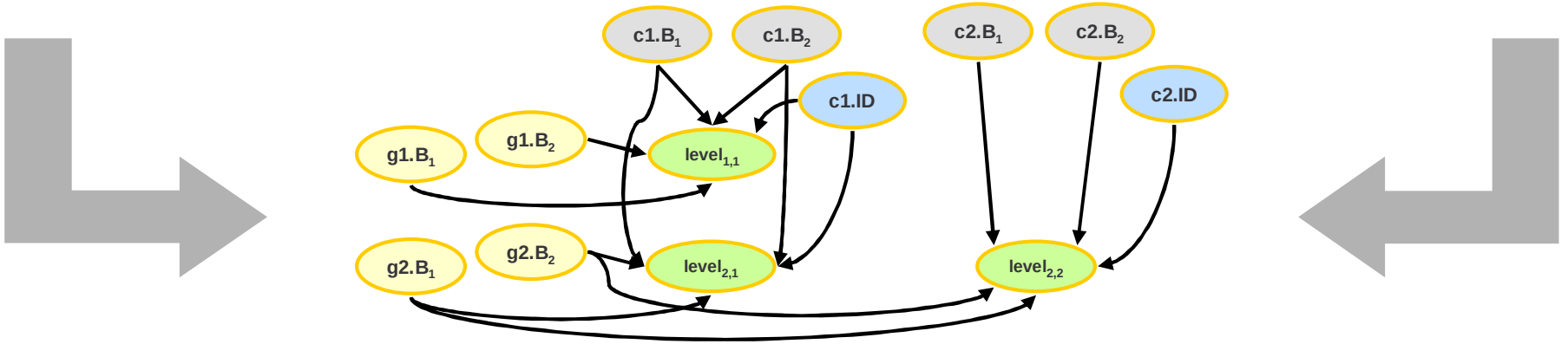
ID	B1	B2
c1	? *(0 or 1)*	? *(0 or 1)*
c2	? *(0 or 1)*	? *(0 or 1)*

Expression

g.ID	c.ID	level
g1	c1	-2.4
g1	*c2*	*(missing value)*
g2	c1	1.6
g2	c2	0.5

PRM model





ground Bayesian network

*Pro*Bic biclustering model

- ***Pro*Bic posterior (~ likelihood x prior):**

$$posterior \propto \prod_{c \in set(c.ID)} \{ \prod_{(gb,cb) \in set(G.B,C.B)} P(\mu_{gb,cb,c}, \sigma_{gb,cb,c}) [\prod_{\substack{e \in E: \\ e.gene.B=gb, \\ e.cond.B=cb, \\ e.cond.ID=c}} P(e.L | g.B_1, g.B_2, c.B_1, c.B_2, c.ID)] \}$$

$$\prod_k \prod_c P(c.B_k) \qquad \prod_k \prod_g P(g.B_k)$$

Expression level prior (μ, σ)'s

Expression level conditional probabilities

Prior condition-to-bicluster assignment

Prior gene-to-bicluster assignment

Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

- Biclustering and biology
- Probabilistic Relational Models
- *Pro*Bic biclustering model
- Algorithm
- Results
- Conclusion

Algorithm: choices

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Different approaches possible**
- **Only approximative algorithms are tractable:**
  - MCMC methods (e.g. Gibbs sampling)
  - Expectation-Maximization (soft, hard assignment)
  - Variational approaches
  - simulated annealing, genetic algorithms, …
- **We chose a hard assignment Expectation-Maximization algorithm (E.-M.)**
  - Natural decomposition of the model in E.-M. steps
  - Efficient
  - Good convergence properties for this model
  - Extensible

Algorithm: Expectation-Maximization

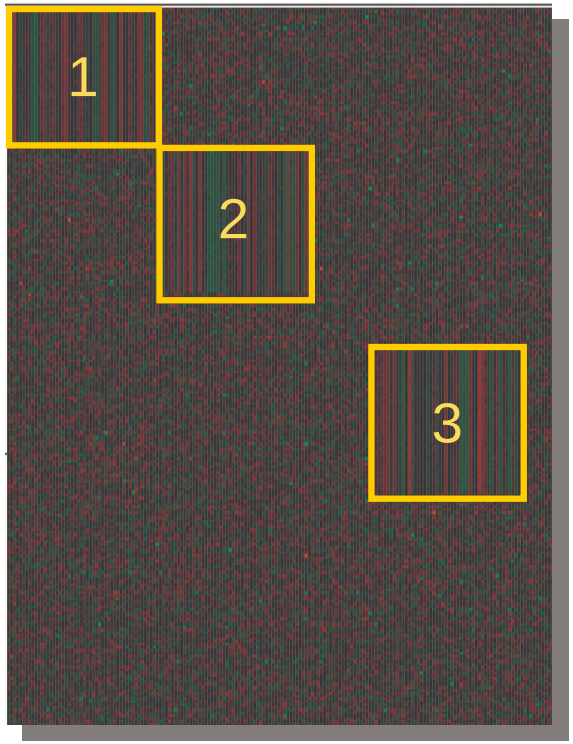
KATHOLIEKE UNIVERSITEIT LEUVEN

- **Maximization step:**
  - Maximize posterior w.r.t. μ, σ values (model parameters), given the current gene-bicluster and condition-bicluster assignments (=the hidden variables)
- **Expectation step:**
  - Maximize posterior w.r.t. gene-bicluster and condition-bicluster assignments, given the current model parameters
  - Two-step approach:
    - **Step 1:** max. posterior w.r.t. C.B, given G.B and μ, σ values
    - **Step 2:** max. posterior w.r.t. G.B, given C.B and μ, σ values

Algorithm: Expectation-Maximization

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Expectation step 1: condition-bicluster assignment**
  - Independent per condition
  - Evaluate function for every condition and for every bicluster assignment
    e.g. 200 conditions, 30 biclusters: 200 * 2^{30} = 200 billion ~ a lot
  - But can be performed very efficiently:
    - Partial solutions can be reused among different bicluster assignments
    - Only evaluate potential good solutions: use ***Apriori*-like approach**.
    - Avoid background evaluations



Algorithm: initialization

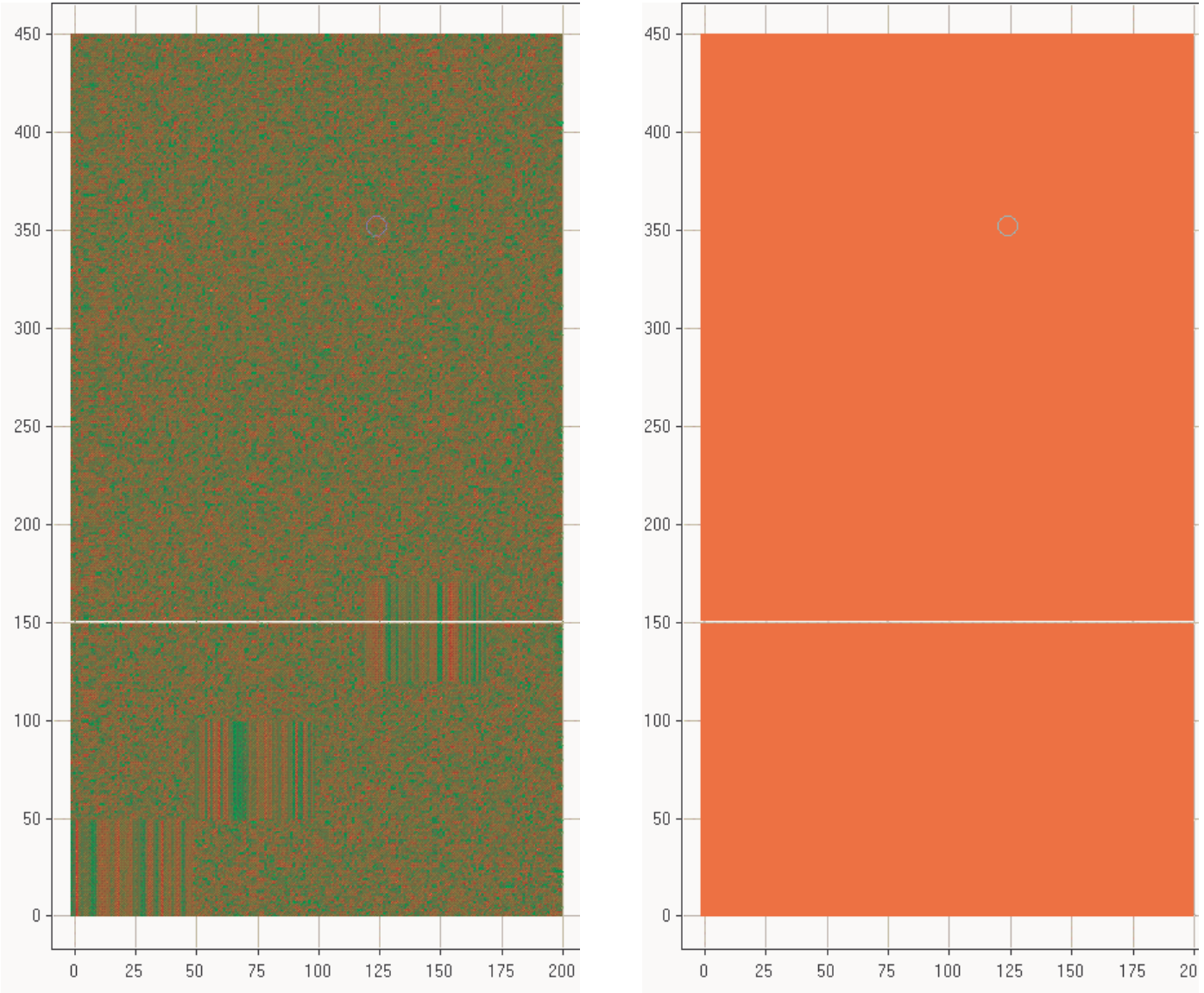
KATHOLIEKE UNIVERSITEIT LEUVEN

- **Initialization options:**
  - Multiple random initializations
  - Initialize biclusters with (nearly) complete dataset
  - Initialize all biclusters simultaneously
  - Init/converge one bicluster at a time, then add next (still allowing first bicluster to change)

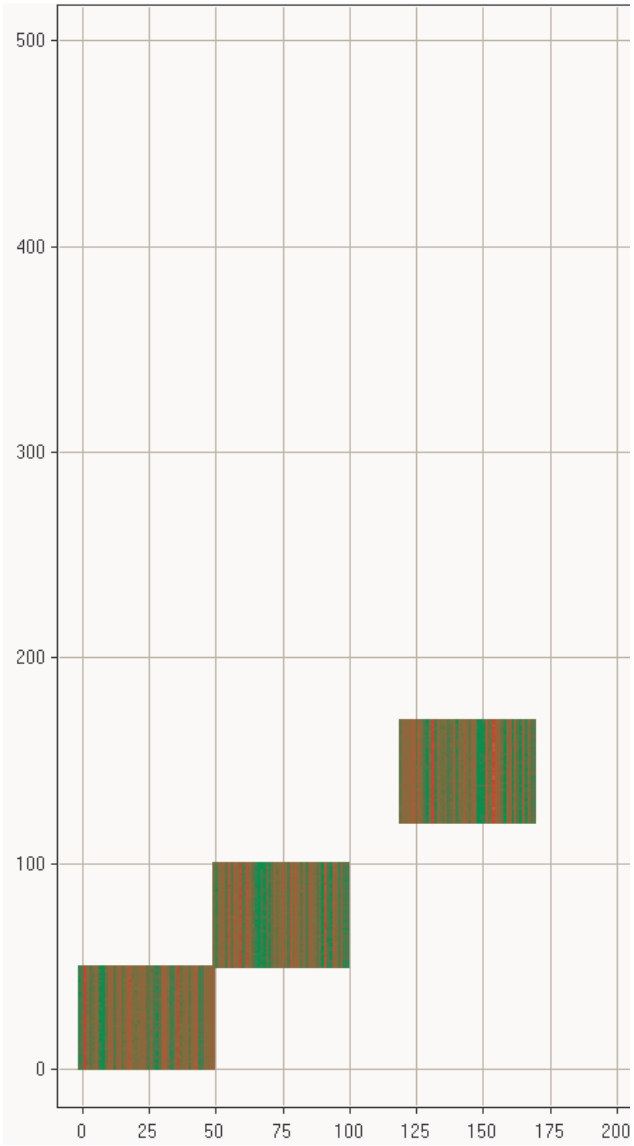
- **Best results:**
  - One initialization: initialize biclusters with (nearly) complete dataset
  - Iteratively add one bicluster and run E.-M.

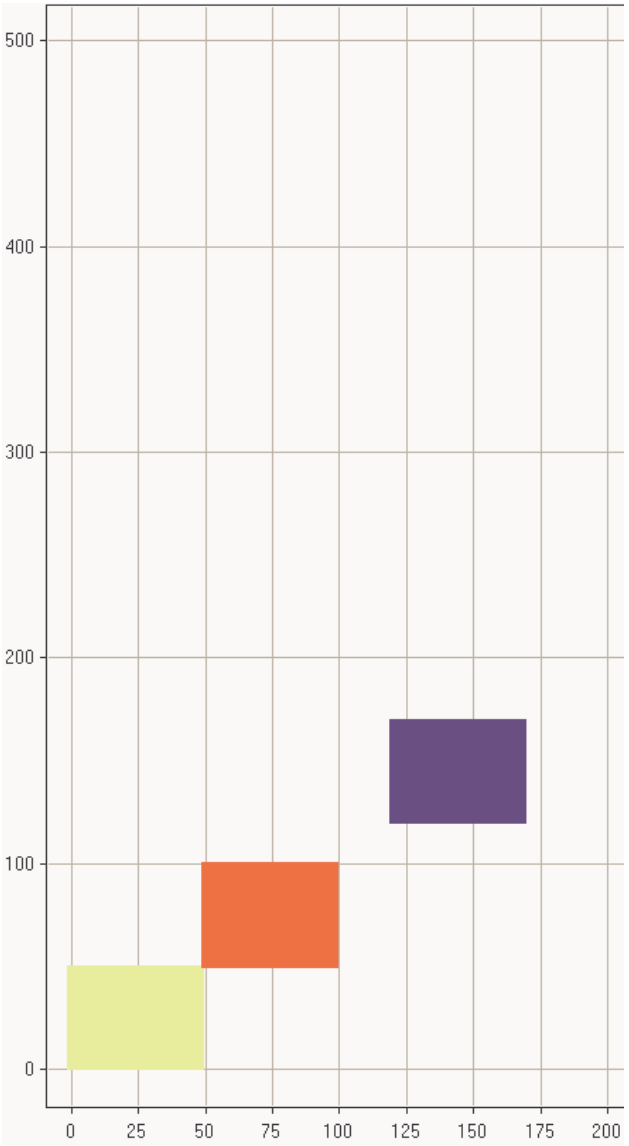
Algorithm: example

KATHOLIEKE UNIVERSITEIT LEUVEN



Algorithm: example





Algorithm properties

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Speed:**
  - 500 genes, 200 conditions, 2 biclusters: 2 min.
  - Scaling:
    - ~ #genes . #conditions . $2^{\#\text{biclusters}}$ (worse case)
    - ~ #genes . #conditions . $(\#\text{biclusters})^{p}$ (in practice), p=1..3

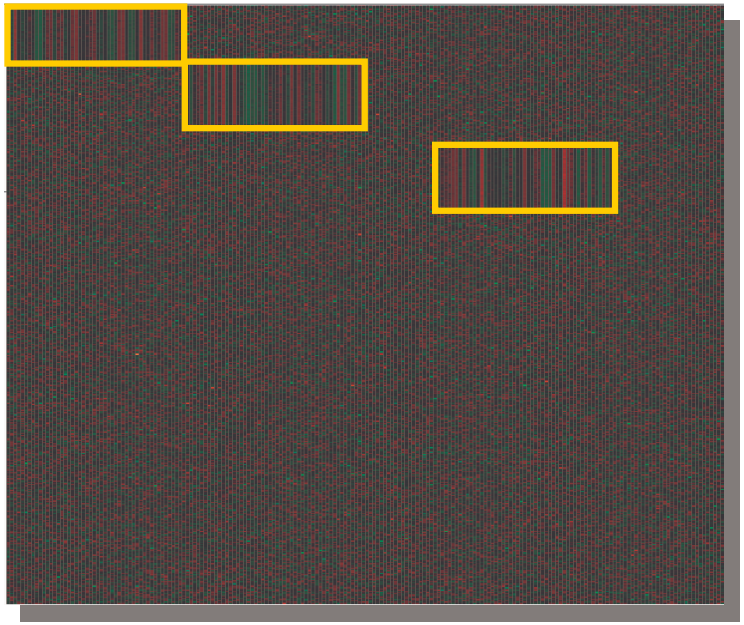
Overview

KATHOLIEKE UNIVERSITEIT LEUVEN

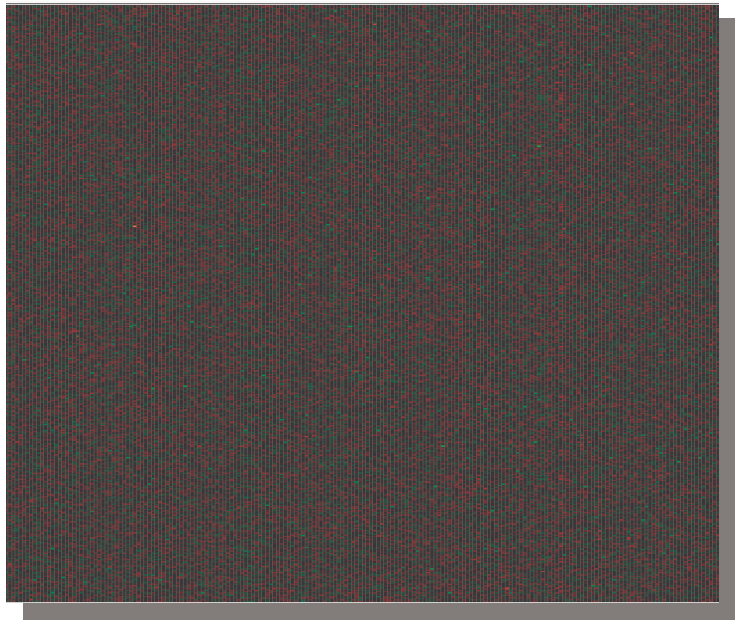
- Biclustering and biology
- Probabilistic Relational Models
- *Pro*Bic biclustering model
- Algorithm
- **Results**
  - Noise sensitivity
  - Bicluster shape
  - Overlap
- Conclusion

Results: noise sensitivity

- **Setup:**
  - Simulated dataset: 500 genes x 200 conditions
  - Background distribution: Normal(0,1)
  - Bicluster distributions: Normal(rnd(N(0,1)), σ), varying sigma
  - Shapes: three 50x50 biclusters



ordered

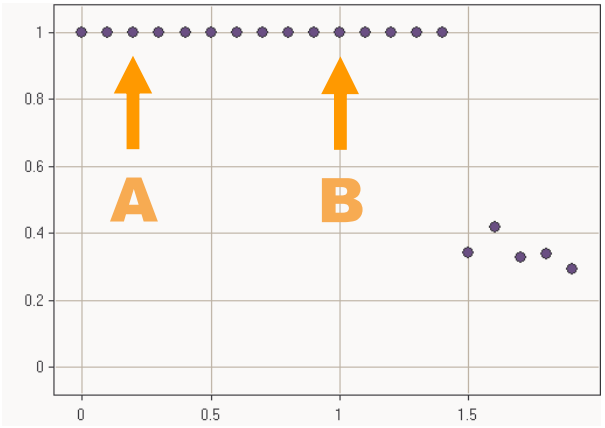


randomized

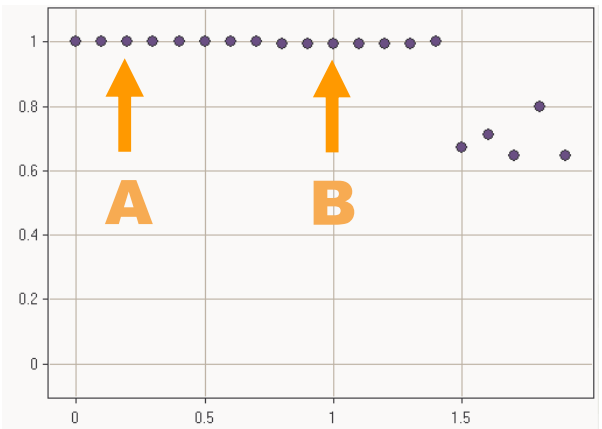
Results: noise sensitivity

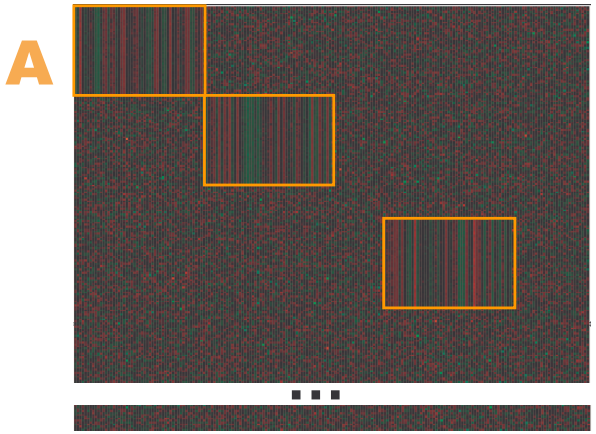
KATHOLIEKE UNIVERSITEIT LEUVEN

Precision (genes)

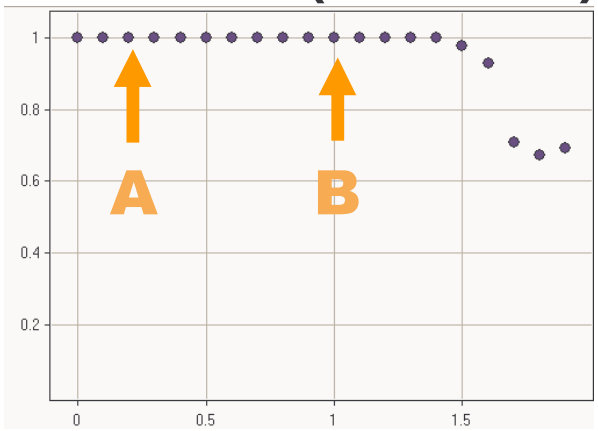


Recall (genes)

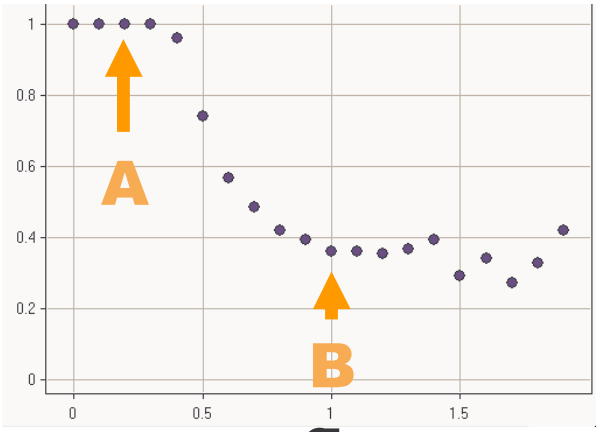


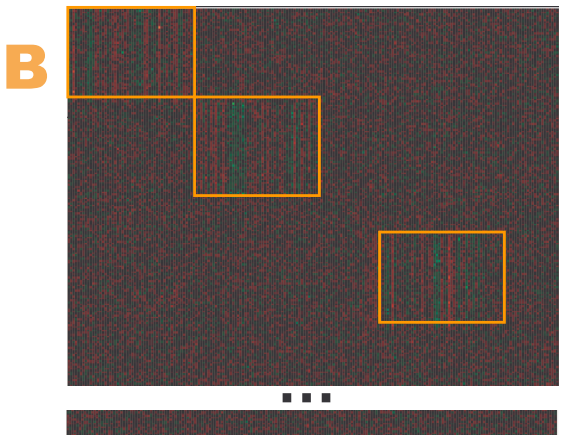


Precision (conditions)



Recall (conditions)





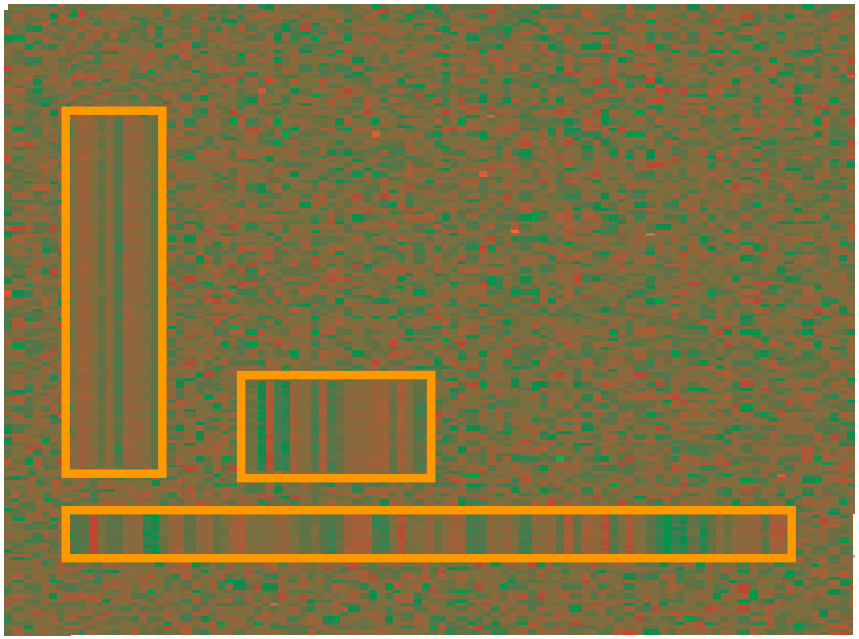
Precision = TP / (TP+FP)

Recall = TP / (TP+FN)

Results: bicluster shape independence

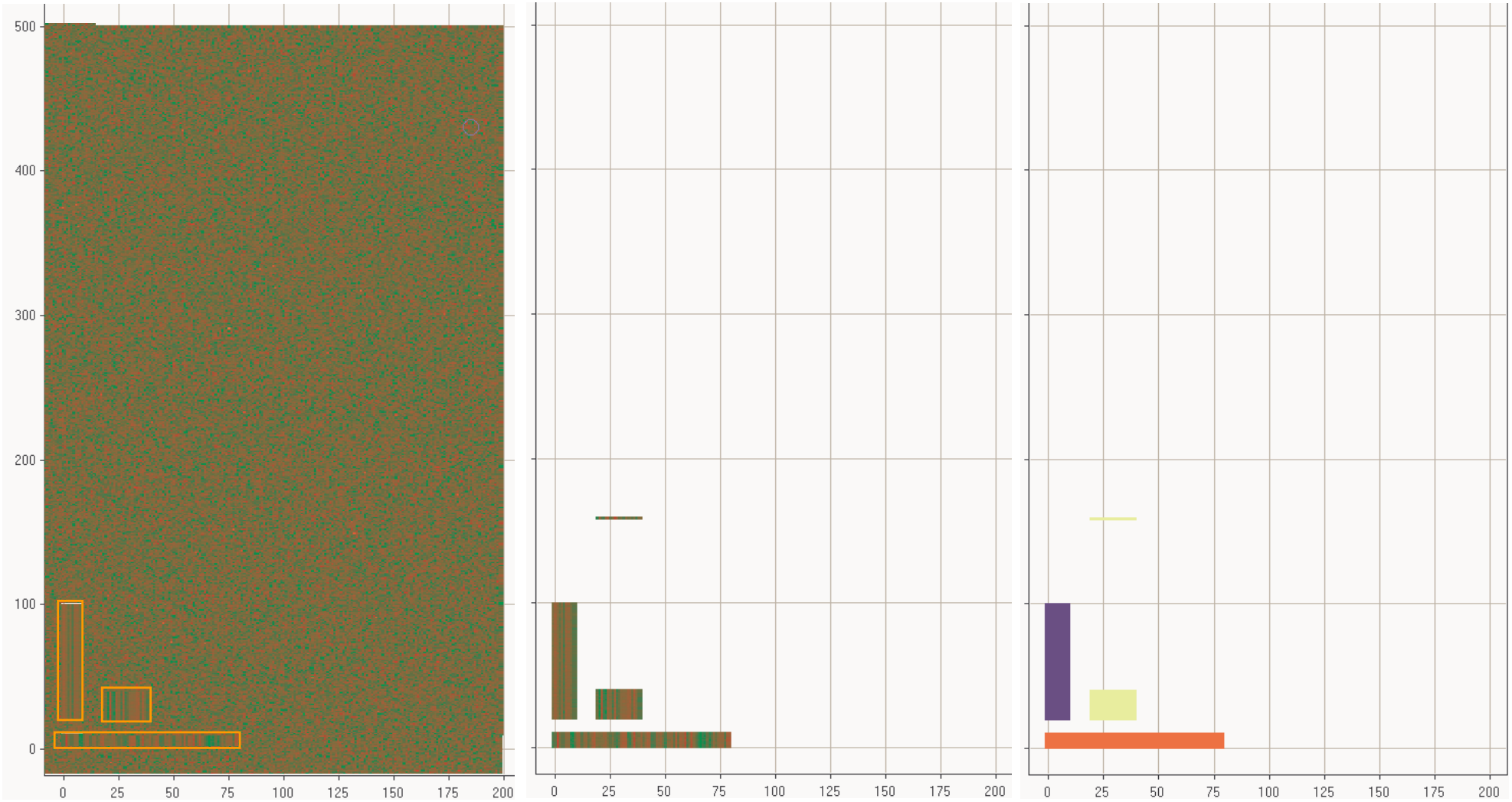
KATHOLIEKE UNIVERSITEIT LEUVEN

- **Setup:**
  - Dataset: 500 genes x 200 conditions
  - Background distribution: N(0,1)
  - Bicluster distributions: N(rnd(N(0,1)), 0.2)
  - Shapes: 80x10, 10x80, 20x20



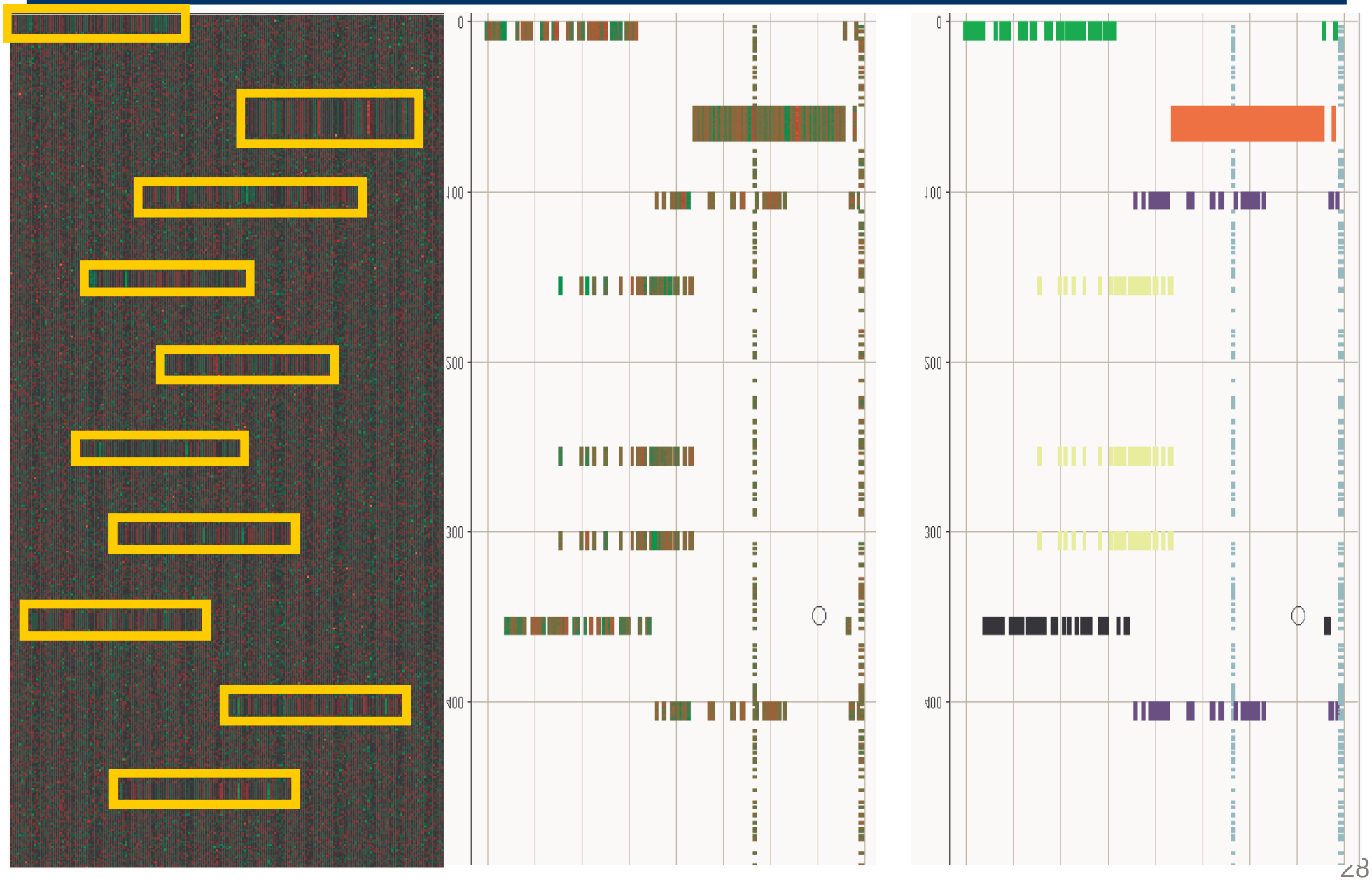
Results: bicluster shape independence

KATHOLIEKE UNIVERSITEIT LEUVEN



Results: 10 biclusters

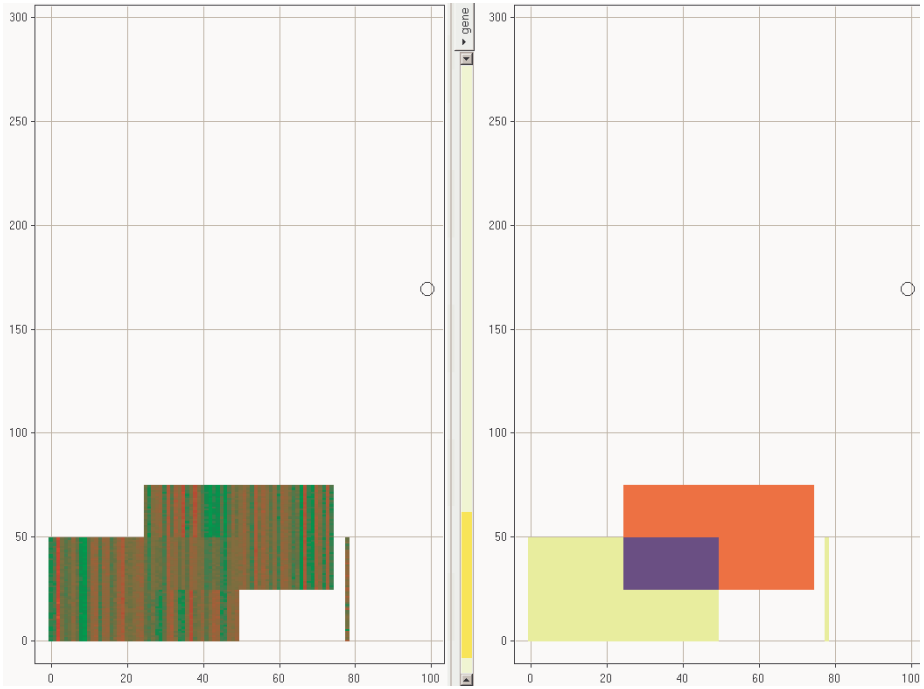
KATHOLIEKE UNIVERSITEIT LEUVEN



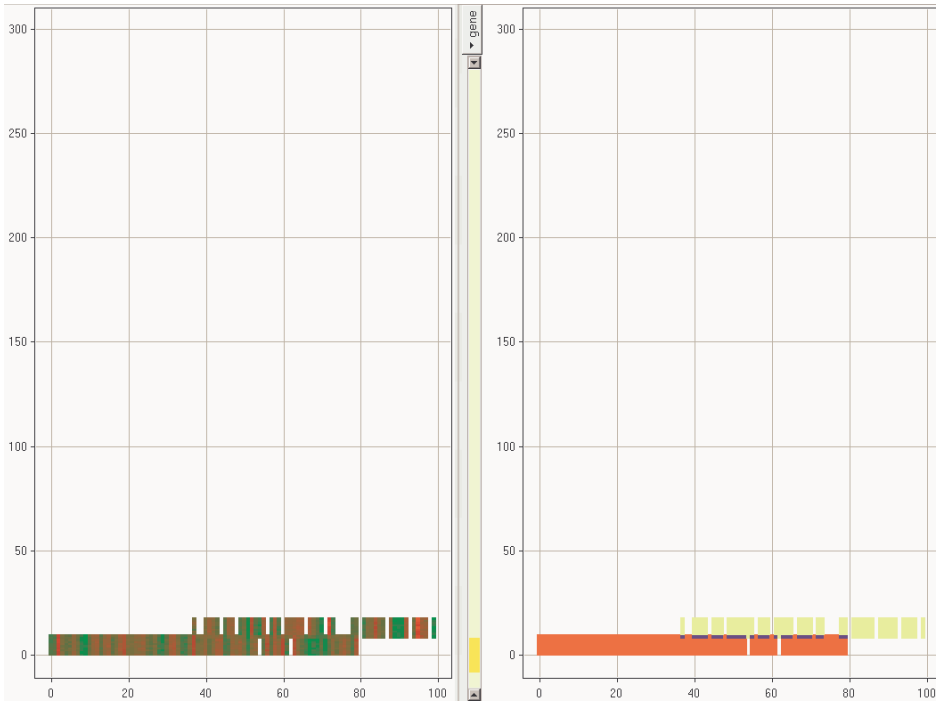
Overlap examples

KATHOLIEKE UNIVERSITEIT LEUVEN

- **Two biclusters**
  (50 genes, 50 conditions)
- **Overlap:**
  25 genes, 25 conditions

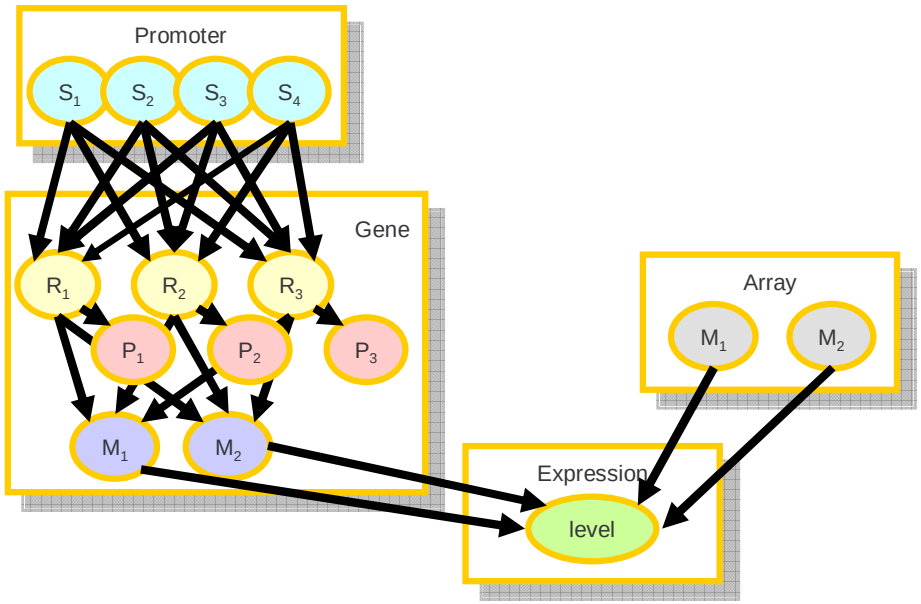


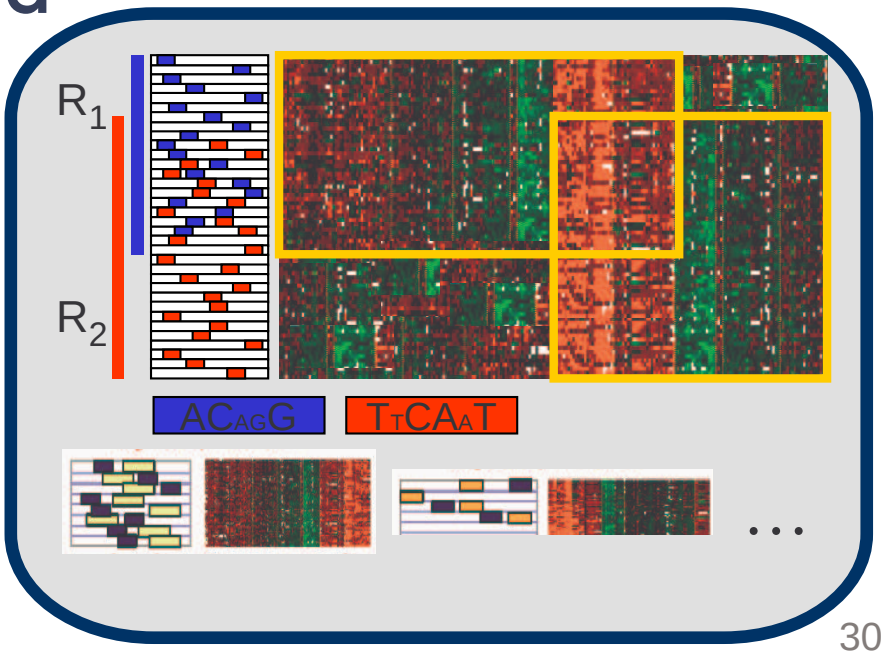
- **Two biclusters**
  (10 genes, 80 conditions)
- **Overlap:**
  2 genes, 40 conditions



Near future

- **Automated definition of algorithm parameter settings**
- **Application biological datasets**
  - Dataset normalization
- **Extend model with different overlap models**
- **Model extension from biclusters to regulatory modules include motif + ChIP-chip data**





Conclusion

KATHOLIEKE UNIVERSITEIT LEUVEN

- Noise robustness
- Naturally deals with missing values
- Independent of bicluster shape
- Simultaneous identification of multiple overlapping biclusters
- Can be used query-driven
- Extensible

Acknowledgements

KULeuven:

- **whole Biol group, ESAT-SCD**
  - **Hui Zhao**
  - Thomas Dhollander
- **whole CMPG group**
  (Centre of Microbial and Plant Genetics)
  - Kristof Engelen
  - Kathleen Marchal

UGent:

- **whole Bioinformatics & Evolutionary Genomics group**
  - Tom Michoel