

# **News from the ECCB 2002**



# Talks



- Discovery of Differentially Expressed Genes using Fuzzy Technology
- An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis
- Correlating gene promoters and expression in gene disruption experiments

# **Discovery of Differentially Expressed Genes using Fuzzy Technology**



Guthke, Reinhard; Scherf, Uwe

Hans Knöll Institute for Natural Products  
Research, Jena, GERMANY; Gene Logic  
Inc., Gaithersburg, MD, USA

# Statistical methods for differential gene discovery

---

- T-test, SAM, Golub, ...
- All these have underlying assumptions: e.g. normality
- New: fuzzy logic approach of overlap

# Dataset



- Breast tissue
- predict malignant neoplasm
- normalized and logarithmized
- Affymetrix's GeneChip® HumanGenome U95 (62.840 human gene fragments)
- $n^+$  = 25 infiltrating ductal carcinomas (DCIS)
- $n^-$  = 25 "normal" breast tissues (data set "A")

# New fuzzy score

- $Z_j = M_j \times n_{j+} \times n_{j-} / n_{+} / n_{-}$
- where  $M_j$  is the trapezoidal fuzzy membership function quantifying whether the summarized and normalized distance of overlapping is "small"
- distance between a value  $x_{ij}$  obtained under one of both conditions and the minimum or maximum of data set obtained under the opposite condition is defined positive. Thus, the function  $P(x) = (|x|+x)/2$  cuts off negative values.

# Result



- Concordances of the top-100 genes
- Selection criterion Data set A Data set B
- t-test vs. Golub's criterion 82% 84%
- t-test vs. Fuzzy criterion 43% 23%
- Golub's vs. Fuzzy criterion 35% 26%

# Discussion

---

- concordance  $< 50\%$
- Reason: normally assumption violated for 36 % of all genes in the data set "A" and 15 % of all genes in the data set "B". (Lilliefors modification of the Kolmogorov-Smirnov test at 5% level)
- Consequence: applying only statistical or fuzzy methods can lead to false negative results
- Therefore, statistical and fuzzy methods should be applied complementary.

# **An automatic block and spot indexing with k-nearest neighbors graph for microarray image analysis**



■ *Ho-Youl Jung and Hwan-Gue Cho*

■ Department of Computer Science, Pusan National University, San-30, Jangjeon-dong, Keumjeong-gu, Pusan, 609-735, Korea

# Problems

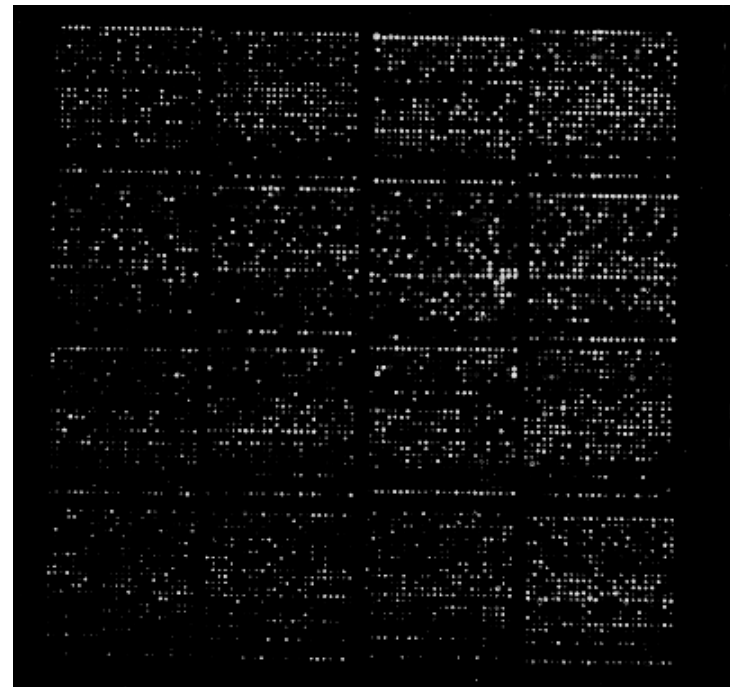


- Block indexing
- Spot indexing
- Intensity computation
  
- Especially hard if many noisy or low expressed spots
- Solution: nearest neighbor graph model

# Image analysis

---

- 4\*4 blocks
- 24\*24 spots each
- Alizadeh
- Microarray



# Image analysis systems



- Assistance software: GUI tools for assisting the user to lay down a template and manually adjust the positions and the sizes of the spots.
- Semi-automated processing system: User specifies the bounding area or guide spots of the array, and system automatically locates each spot. GUI tools for manual corrections of any possibly misidentified spots are provided.
- Fully automated processing system: No human needs interactions. The grid and spots are automatically found and quantitated.

# Comparison

**Table 1.** A comparison of microarray image analysis systems

	Block indexing	Spot indexing	Skewed image
ScanAlyze	manual	manual	do not consider
GenePix	manual	manual	do not consider
AutoGene	manual	automatic	do not consider
Steinfath <i>et al.</i>	manual	automatic	consider <sup>a</sup>
Jain <i>et al.</i>	automatic <sup>b</sup>	automatic <sup>b</sup>	consider <sup>c</sup>
Our System	automatic	automatic	consider

<sup>a</sup>If the microarray image has a high positional error then the computation may not be correct.

<sup>b</sup>In this work the histogram method is used, which has a major defect in processing skewed images.

<sup>c</sup>This system corrects the skewed image by asking the user for the angle of skewedness.

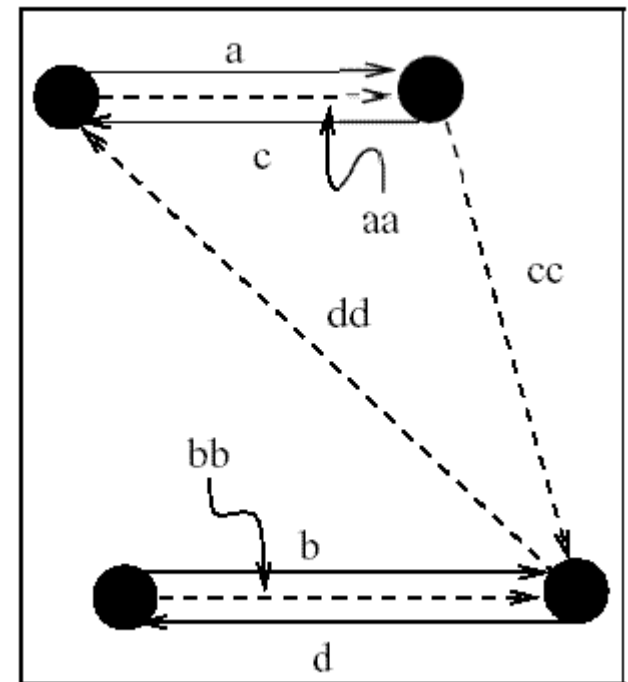
# Real data problems

---

- (i) spot position variation,
- (ii) spot shape and size irregularity,
- (iii) sample contamination,
- (iv) global problems that affect multiple spots (Sचना, 2000).

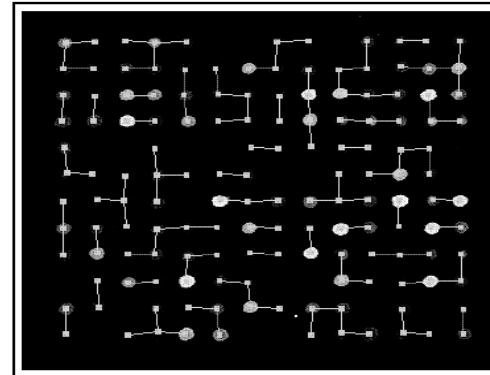
# Block Indexing

- Use image segmentation technique (8-connected-components)
- Use MKNN (modified k nearest neighbours) graph model

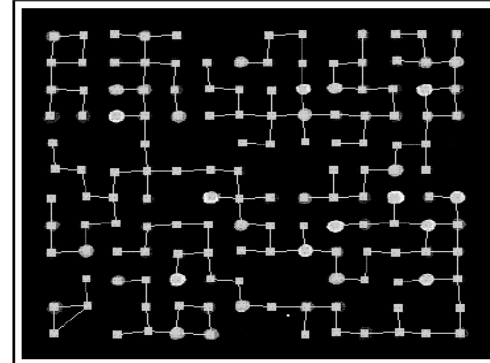


# KNN vs MKNN

As the modified KNN avoids reciprocal pairs it has bigger connected components as KNN.



(a) KNN graph of a sample image : the number of components = 48



(b) Modified KNN : the number of components = 6

Fig. 3. A comparison of KNN and modified KNN graphs ( $k = 1$ ).

# Block identification

Intersection of MBR (min bounding rectangles)

Merge as long as possible

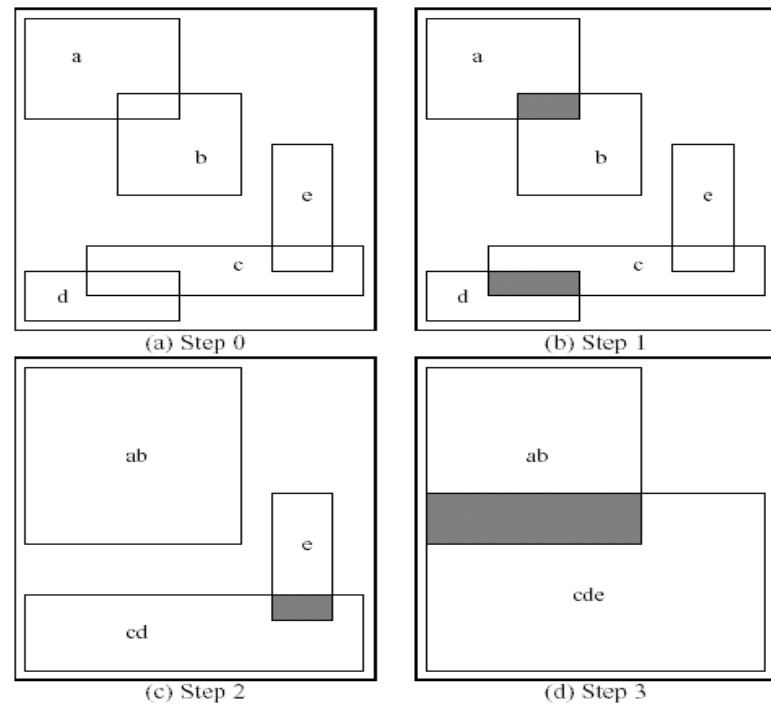
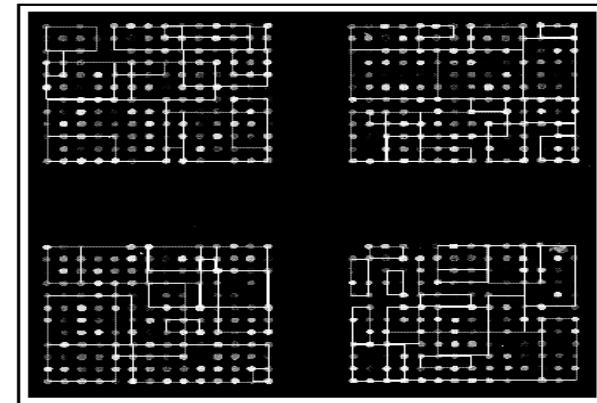


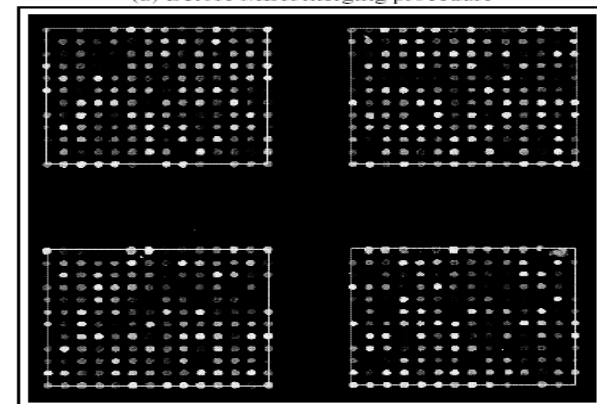
Fig. 4. Identification of a block index using modified KNN graph ( $k = 1$ ).

# Result block identification

Successful block identification if more than 60% of the spots expressed and distance between spots is 1/2 of distance between blocks



(a) Before MBR merging procedure

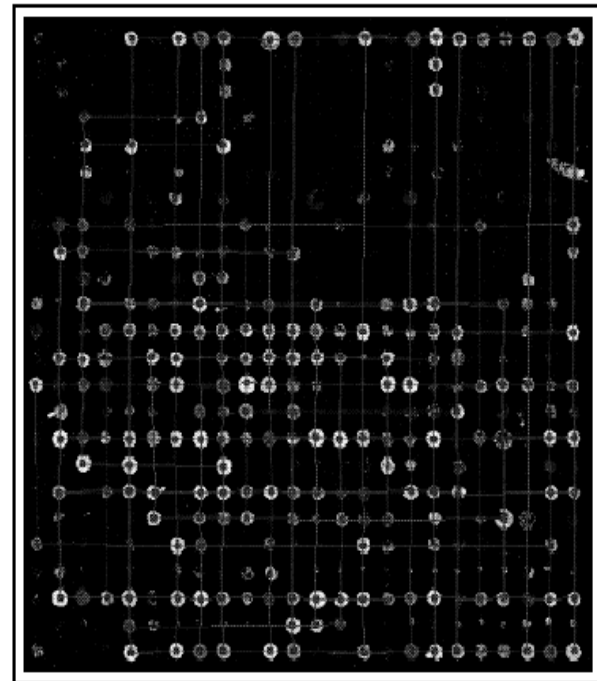


(b) After MBR merging procedure

# Spot indexing

$\epsilon$ -graph model of  
microarray image

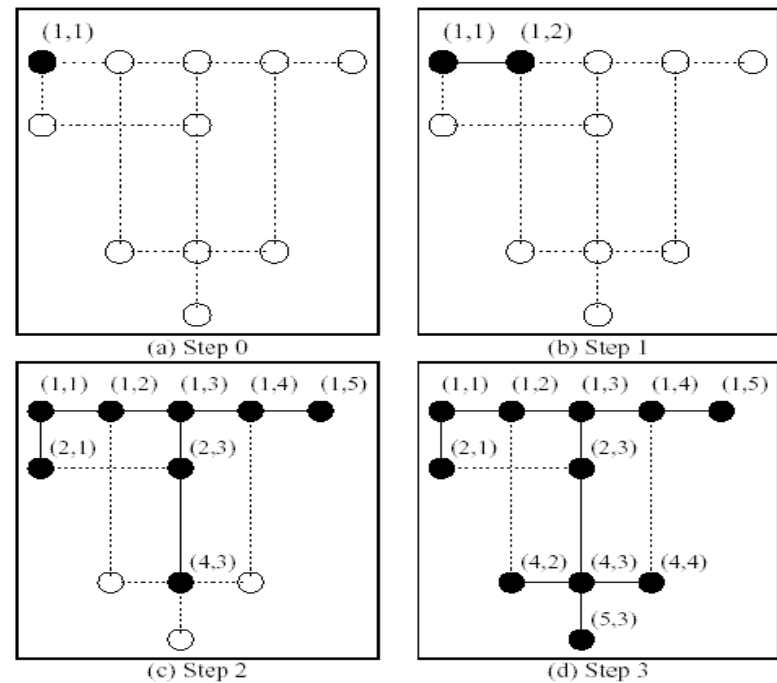
just consider spots vertically  
or horizontally above the  
current spot with offset of  
maximally  $\epsilon$



**Fig. 7.** An example of  $G_\epsilon = (V, E)$ :  $\epsilon = 2$  pixels. The size of the image is  $434 \times 420$  pixels.

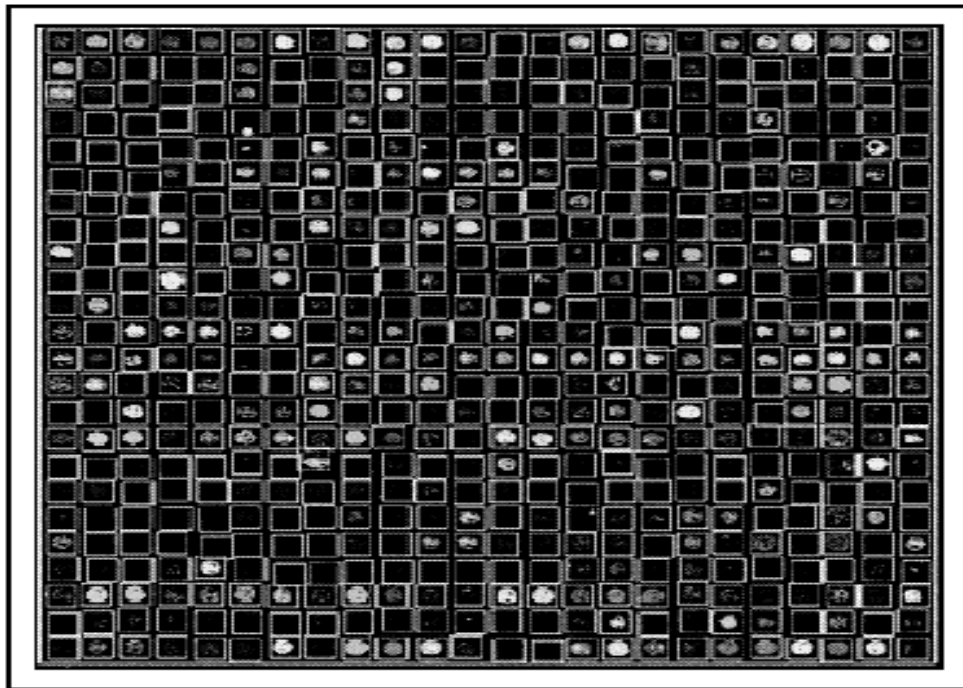
# Spot Numbering

Start with strong guide spots, calc mean distance between spots, mean spot radius



**Fig. 8.** An example of spot indexing procedure: (a) a guide spot is located at (1,1); (b) at the second step the vertex (1,1) links to the vertex (1,2); (c) at the third step vertex the vertex (2,3) links to the vertex (4,3); (d) all vertices are visited.

# Result spot indexing

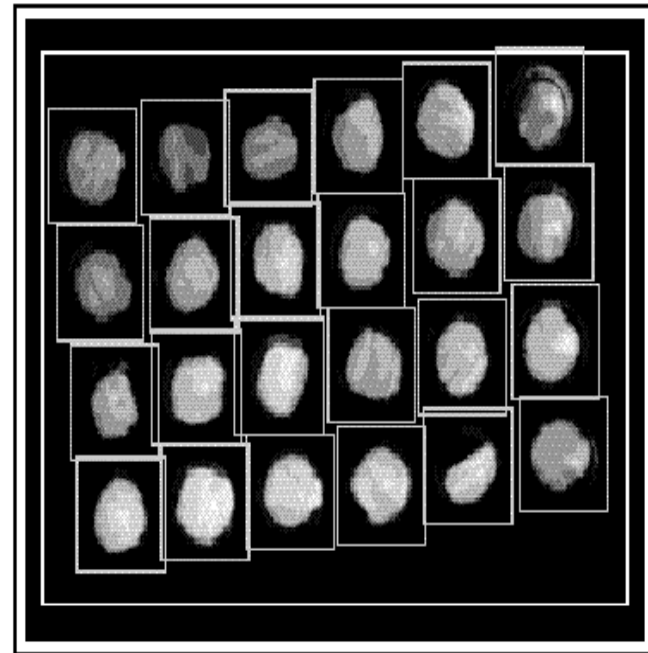


**Fig. 9.** Spot addressing based on the graph traversal algorithm: Each quadrangle represents each region of spots.

# Skewed Images

Steinfath et al used block edges

Here: use of all edges on  $\epsilon$ -graph



**Fig. 10.** Microarray image is skewed: This image was obtained from the photosynthesis of a rice leaf.

# Test data

---

- Artificial data (ART10, ART20, . . . , ART100): their expression rates are from 10% to 100%.  $br$  and  $bc$  is 2 and the number of rows ( $sr$ ) and columns ( $sc$ ) in each block is 20.
- Real microarray data about cabbage gene (CBG1, CBG2, CBG3, CBG4, CBG5) :  $br = bc = 4$  and  $sr = 12$ ,  $sc = 14$ , respectively.
- Microarray image from NIH (NIH1, NIH2, NIH3) : Real data generated by Alizadeh *et al.* (2000) where  $br = bc = 4$  and  $sr = sc = 24$ .
- Real microarray data from Stanford Univ. (STF1, STF2, STF3) :  $br = bc = 2$  and  $sr = sc = 44$ .

# Results

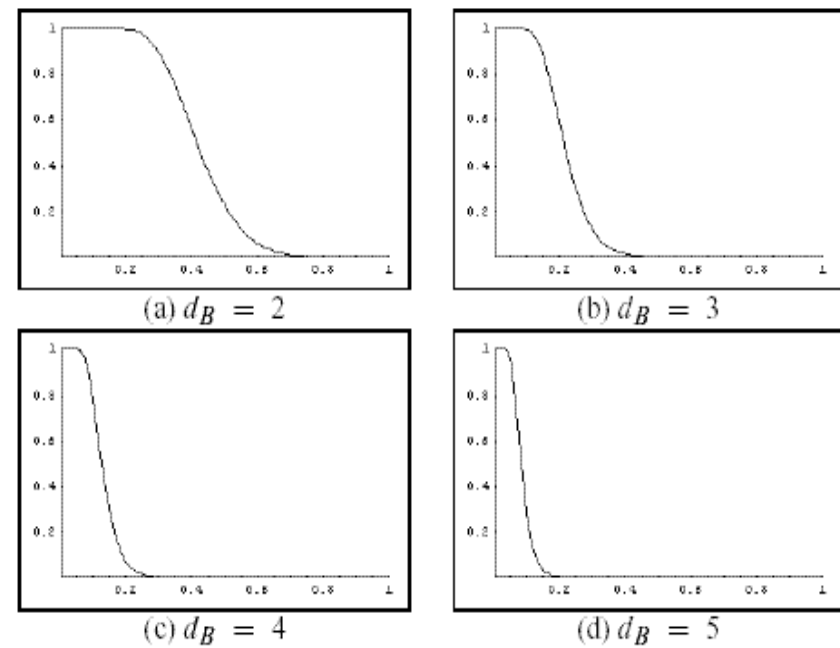


Fig. 13. The probability curve of  $d_B = 2, 3, 4$ , and 5. X-axis is the expression rate  $e_p$  and Y-axis is probability  $\hat{P}_I$ .

# Features



- Using the minimum precondition (e.g. the resolution of blocks in image data and the resolution of spots in a single block), we can index the block and spot successfully.
- Our algorithm can be adapted to not only microarray image data, but also other grid structured image data.
- Our algorithm is robust within a  $10^\circ$  rotation of the grid
- If the expression rate is more than 40%, and the distance between two neighbor blocks,  $dB$ , is more than three times the unit distance,  $d\mu$ , then our algorithm can give a fully automatic image analysis.
- If  $dB = 5$  then our algorithm can index each block automatically even if the expression rate 10%.

# Correlating gene promoters and expression in gene disruption experiments

---

- *Kimmo Palin, Esko Ukkonen, Alvis Brazma and Jaak Vilo*
- *Department of Computer Science, PO Box 26, FIN-00014 University of Helsinki, Finland*
- *European Bioinformatics Institute, Wellcome Trust Genome Campus,*
- *Hinxton, Cambridge, CB10 1SD, UK*

# Gene regulation

---

- So far: Search in clusters of coexpressed genes for signals in sequence that could explain coregulation
- Problems: promotor region in eukaryotes large 30-10000bp (easier in yeast -600bp)

# Problems with predictions

---

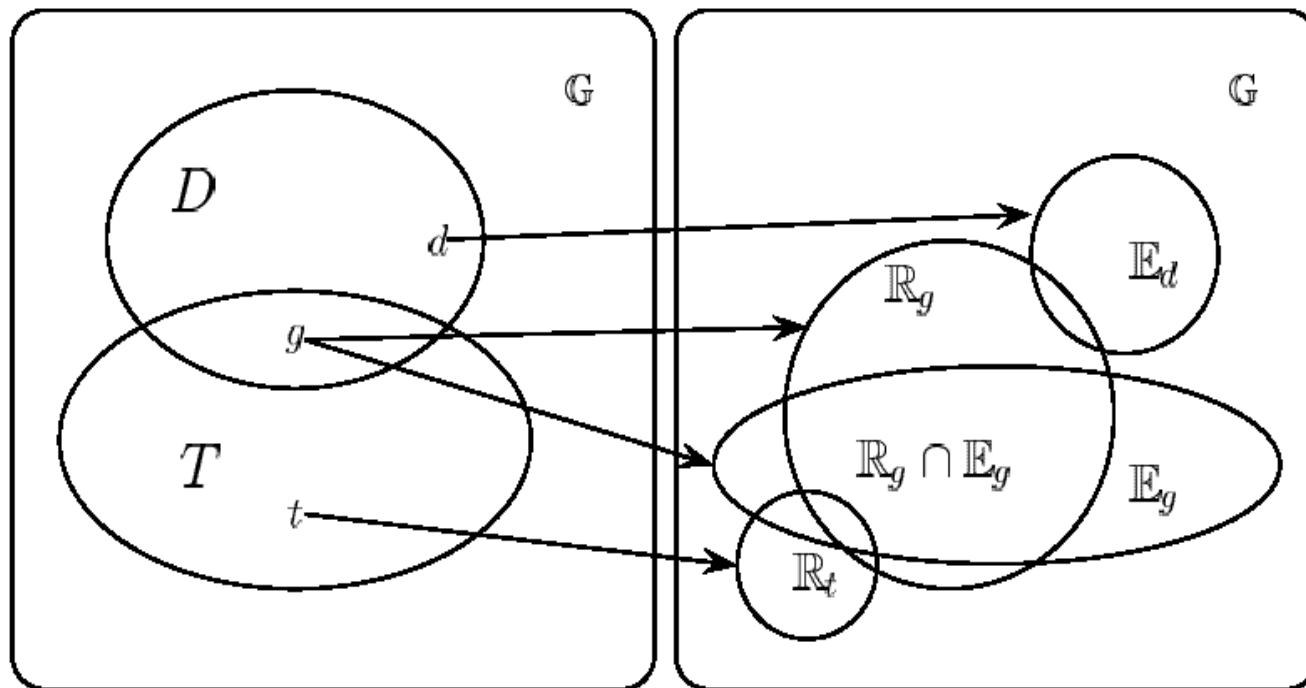
- Chromatin structure plays a role
- protein complexes that bind to site (one protein can be a member of several TF complexes)
- mRNA does only have a 0.6 correlation with protein level (posttranslational modification, ..., Ideker 2001)

# General idea

---

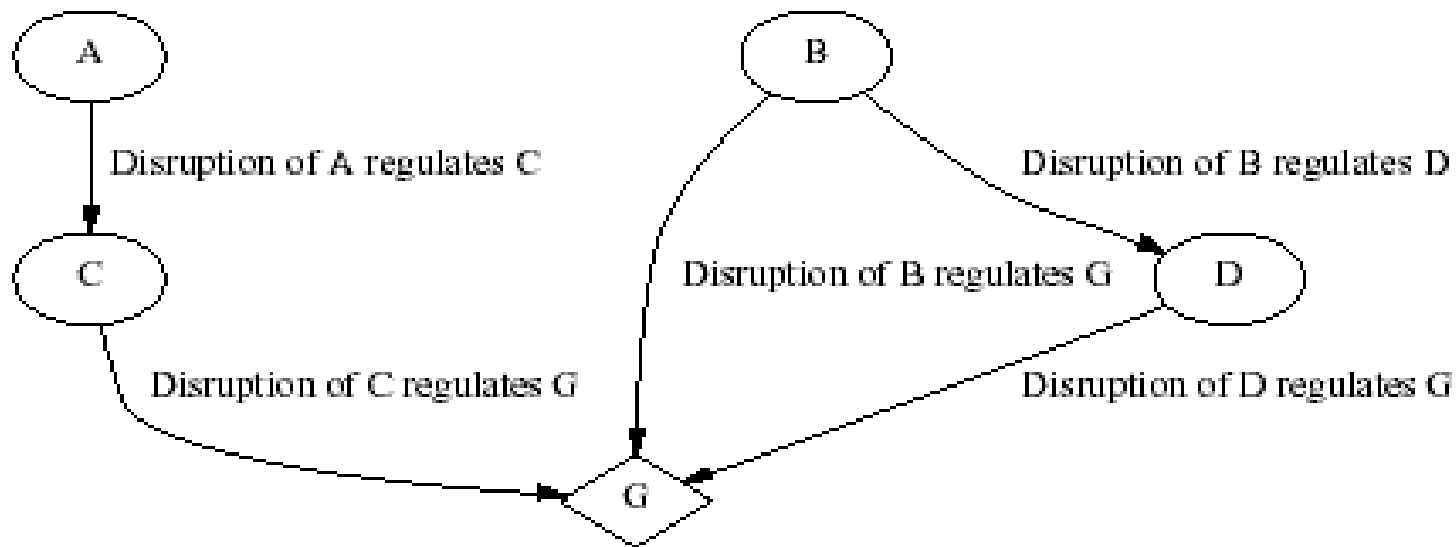
- Combine promotor analysis and expression data
- For each disrupted gene find set of genes that are changed significantly (0.05, ...)
- Find genes that have known binding site in upstream region
- Look at intersection of these sets (use hypergeometric distribution, Holms correction)

# Expression and Binding sets



**Fig. 1.** Disrupted gene  $d$  maps to its effectual set  $\mathbb{E}_d$  and transcription factor  $t$  maps to its regulation set  $\mathbb{R}_t$ . Gene  $g$  that belongs to  $D \cap T$  has both sets  $\mathbb{R}_g$  and  $\mathbb{E}_g$ .

# Disruption network



**Fig. 2.** A disruption network. Disruption of source–gene affects the target-gene.

# Datasets

---

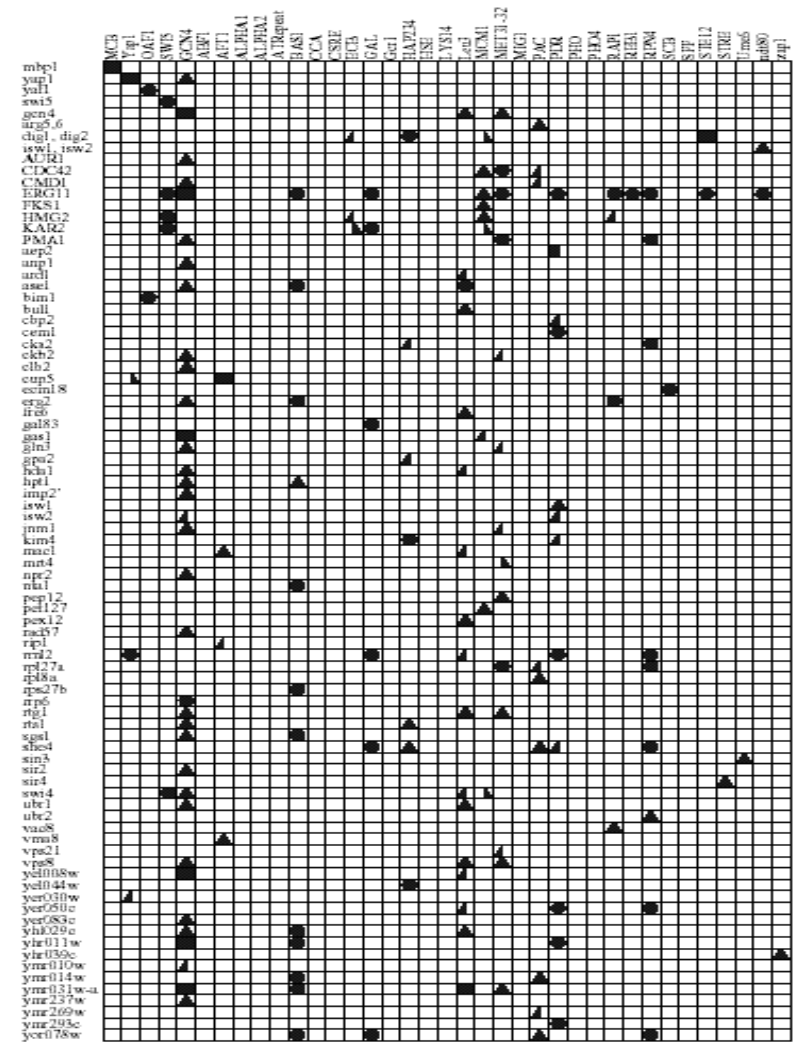
- Expression data:  
263 disruption experiments selected from the compendium of expression profiles Hughes *et al.* (2000). Excluded drug treatments or experiments conducted with yeast under haploid state.
- Regulation set (binding site motifs):  
356 defined by yeast binding sites in Pilpel *et al.* (2001). Of these 37 are previously known and the rest are putative sites generated from MIPS families.

# Results

Most of the analyzed transcription factor binding sites did not show significant correlation with any of the disruptions. Out of the 356 analyzed motifs only 102 had at least one correlating disruption at 0.01 confidence level.

Positively out of the 37 known motifs 20 had at least one correlation.

Out of the 319 generated motifs only 85 had a correlating disruption. In general, for most of these putative motifs the correlations are rare and weak.



**Fig. 3.** Correlations between the disruptions on the left and binding sites on the top. Only left or right part of the triangle or rectangle is shown if the correlation occurs only with weaker or stronger (left and right half of an icon) definition of significantly altered expression. A triangle ▲ for correlation (6 strong, 29 weak, 61 both), a circle ● (54) for the disruption in the expression cascade of the transcription factor. A rectangle ■ is for a correlation explained by the cascade (1 weak, 11 both).

# Results list

**Table 1.** Disruption—Binding Site correlations: Columns for size of the regulation set, name of the binding site motif, size of the effectual set of the best correlating disruption, name of the disruption, size of the intersection of the two sets and description of the result. Size of the effectual set and the intersection is only given for patterns that had one clearly best correlation

$ \mathbb{R}_b $	Site	$ \mathbb{E}_g $	Disruption	$ \mathbb{R}_b \cap \mathbb{E}_g $	Description
184	MCB	8	<i>mbp1</i>	5	Part of a DNA binding complex.
78	YAP1	55	<i>yap1</i>	6	Binding site of <i>yap1</i> factor
116	Ume6	346	<i>sin3</i>	20	Interacting proteins.
210	zap1	3	<i>msc7</i>	3	Relation via hydrogenases.
243	STE12	437	<i>dig11,dig2</i>	36	<i>dig1</i> represses <i>STE12</i> .
153	ndt80	151	<i>isw1,isw2</i>	13	Genetic interaction with <i>isw2</i> .
180	RPN4	33	<i>ubr2</i>	17	Similar cellular role.
257	RAP1	121	<i>vac8</i>	23	Weak link through vacuole
480	BAS1	23	<i>hpt1</i>	11	Adenine response.
149	STRESS	126	<i>sir4</i>	13	Unexplained.
116	HAP234		4 disruptions		Unexplained.
151	GCN4		34 disruptions		Central biosynthesis regulator.
89	Leu3		20 disruptions		In biosynthesis pathway.
58	MET31-32		16 disruptions		In biosynthesis pathway.
188	AFT1		<i>cup5 mac1 vma8</i>		Small molecule transport, iron uptake.
907	rRNA proc.		9 disruptions		Ribosomal activity.
514	PAC		9 disruptions		Ribosomal activity.
356	ECB		5 Weak disruptions		Early Cell-Cycle box
371	PDR		11 Weak disruptions		Unexplained
410	MCM1		10 disruptions		<i>MCM1</i> needs coregulators.