From microarray images to Biological knowledge

Junior Barrera

BIOINFO-USP

DCC/IME-USP

Team

Hugo A. Armelin

Junior Barrera

Helena Brentaini

Marcel Brun

Y. Chen

Edward R. Dougherty

Roberto M. Cesar Jr.

Daniel Dantas

Gustavo Esteves

Marco D. Gubitoso

Nina S. T. Hirata

Roberto Hirata Jr

Luiz F. Reis

Paulo S. Silva

Sandro de Souza

Walter Trepode

....

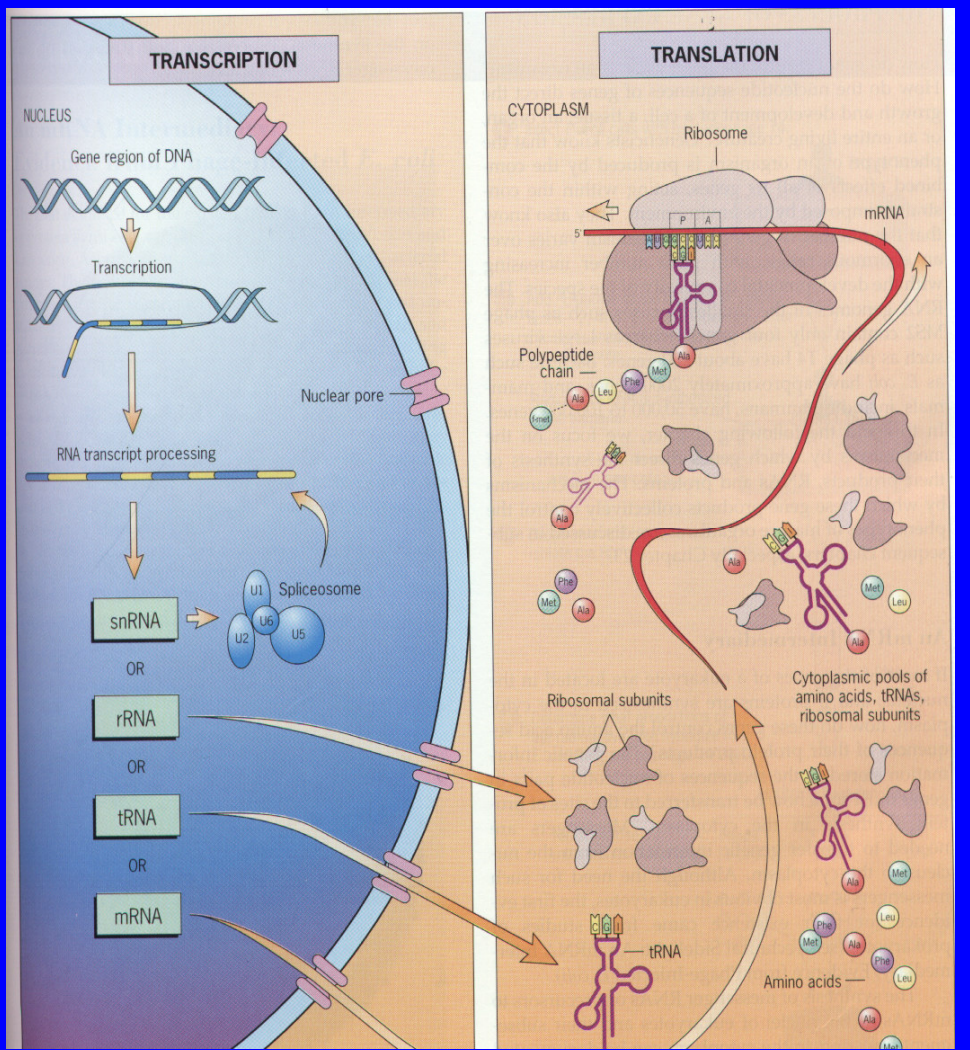
Layout

- Introduction
- Chip design
- Image analysis
- Normalization
- Genes Signature
- Clustering
- Genetic networks
- An environment for knowledge discovery

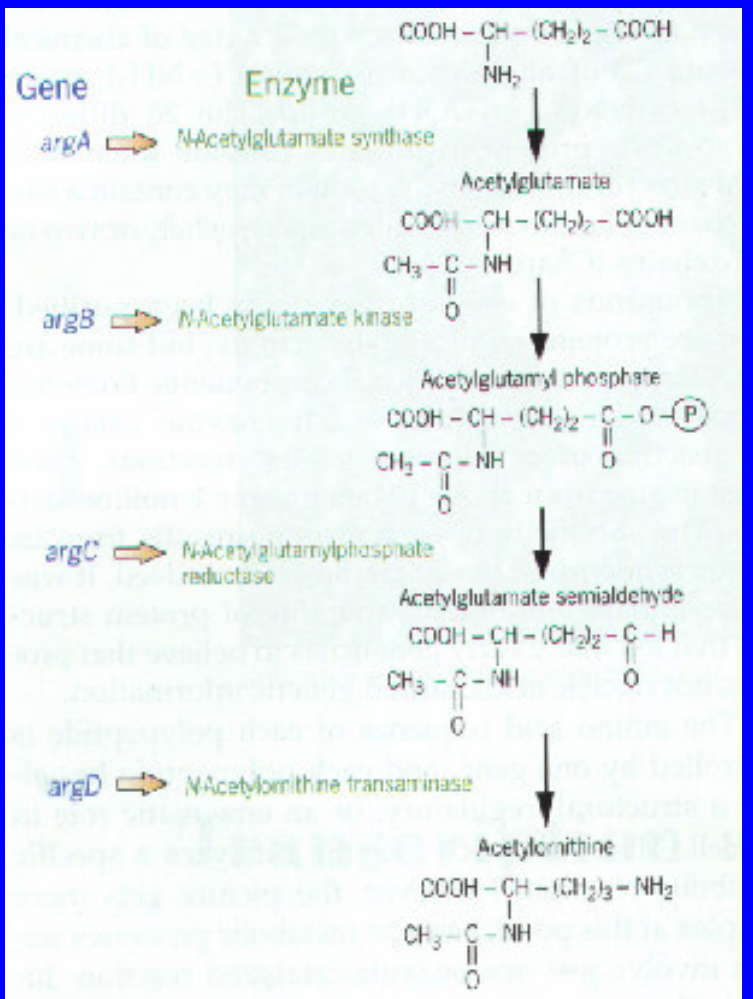
Introduction

Knowledge evolution in genetics

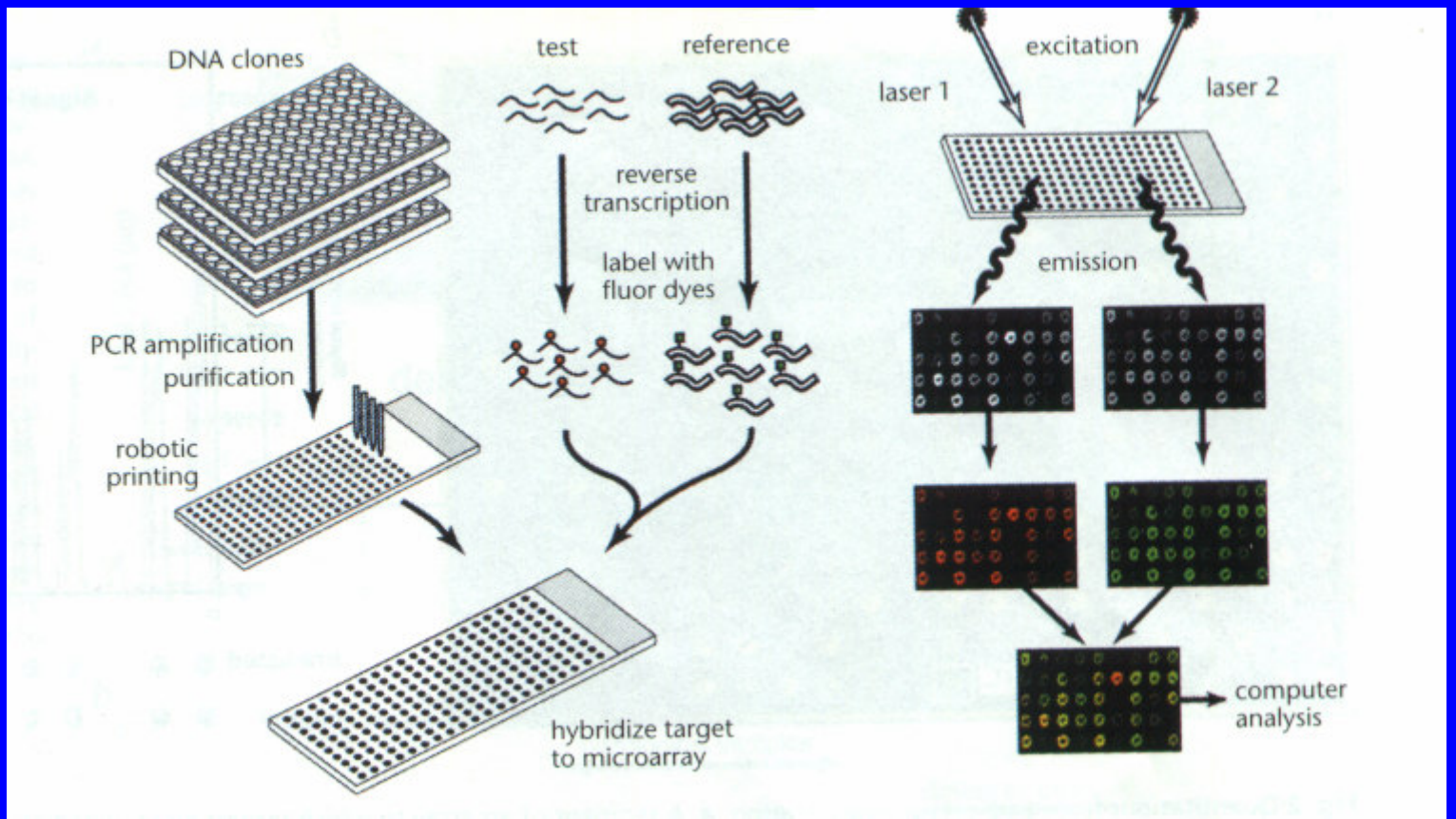
- Gene expression



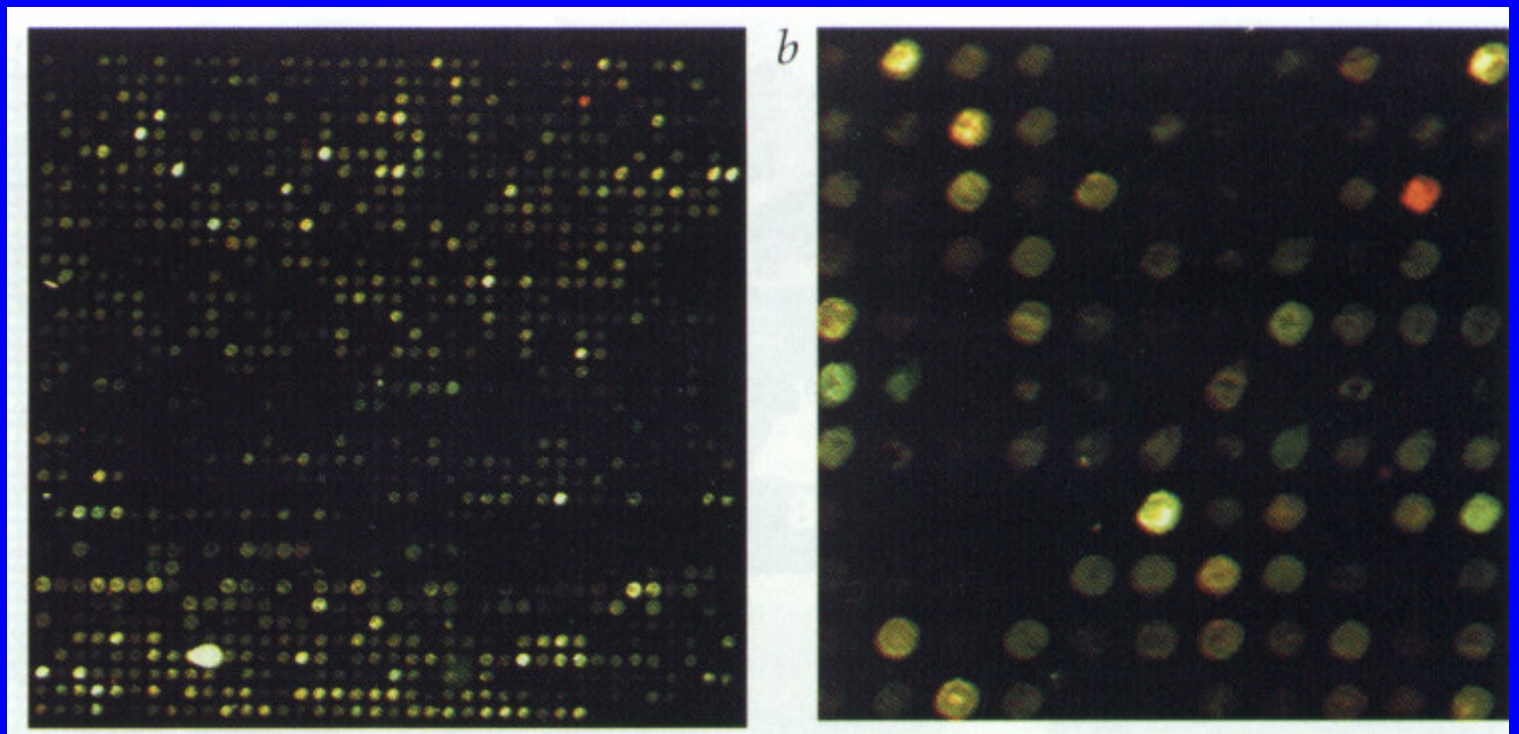
Knowledge evolution in genetics



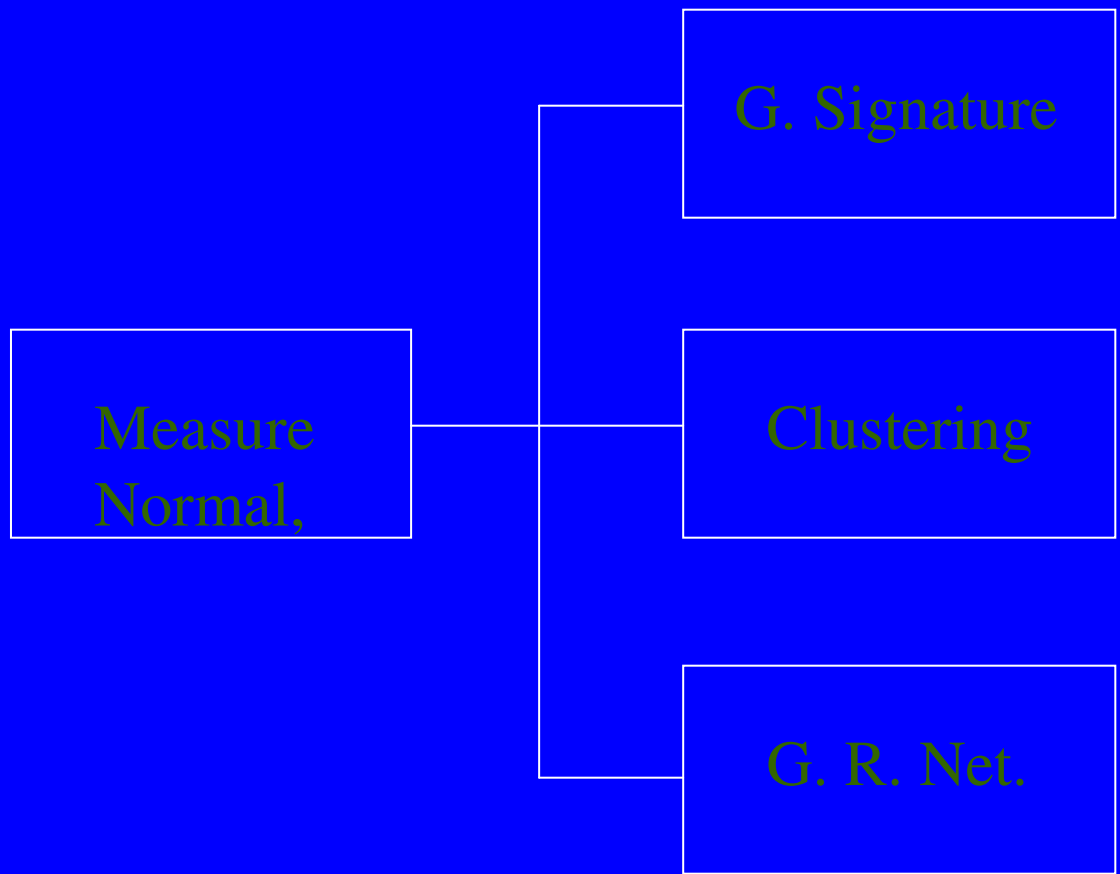
Data acquisition



Data acquisition



Analysis Phases



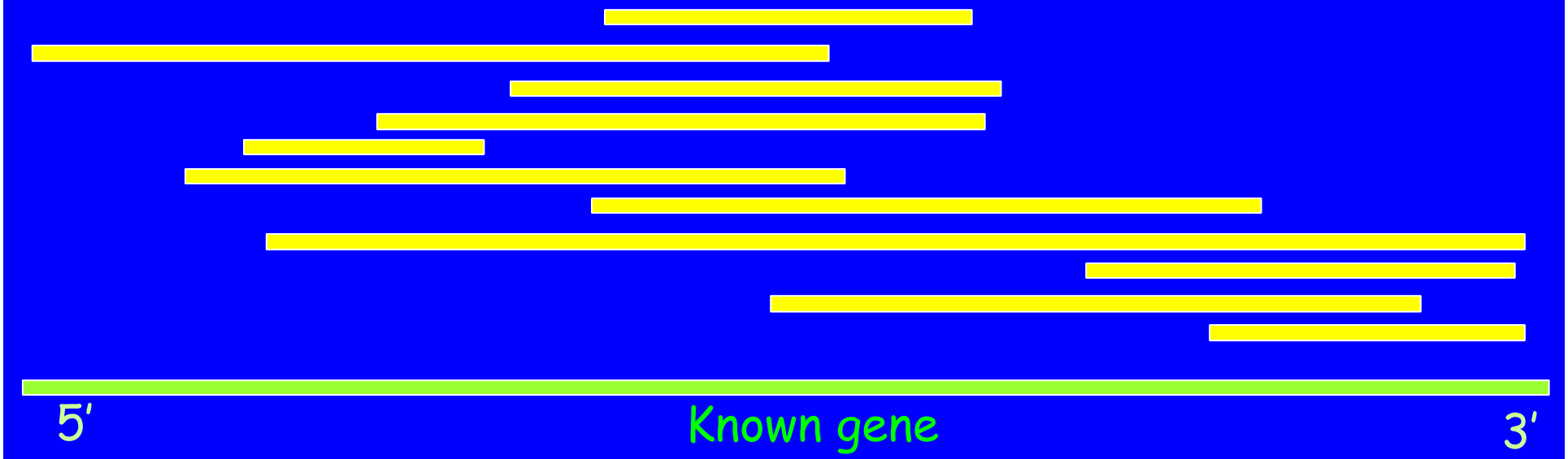
Biological questions

- What genes are enough to separate a cancerous tissue from a normal one?
- What genes can separate to types of cancer?
- What genes differentiate two types of chickens or trees?
- What genes regulate a pathway?

Chip design

Clone Selection

1. Clustering for the same gene

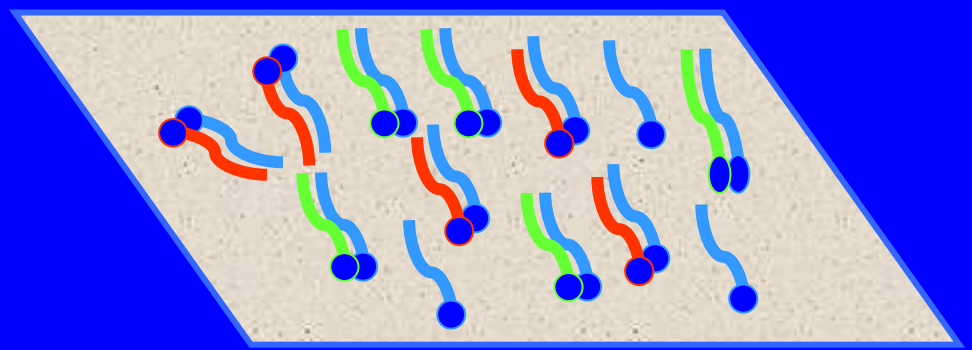


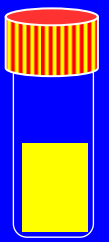
2. Choice of a clone representing the gene

Affect hybridization process

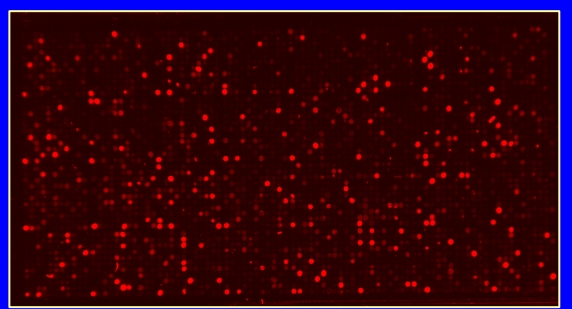
- Clone size - they should have similar size
- Clone position in the gene - they should come from similar regions
- Clone similarity - they should not have large similar regions

Hybridization





Cy5



Cy3

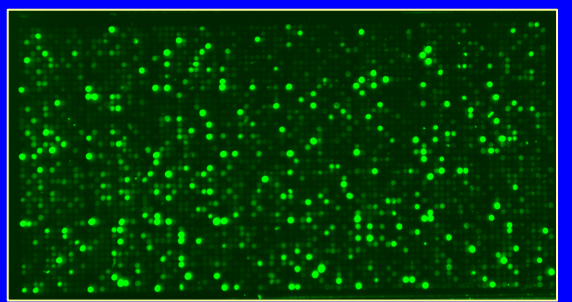
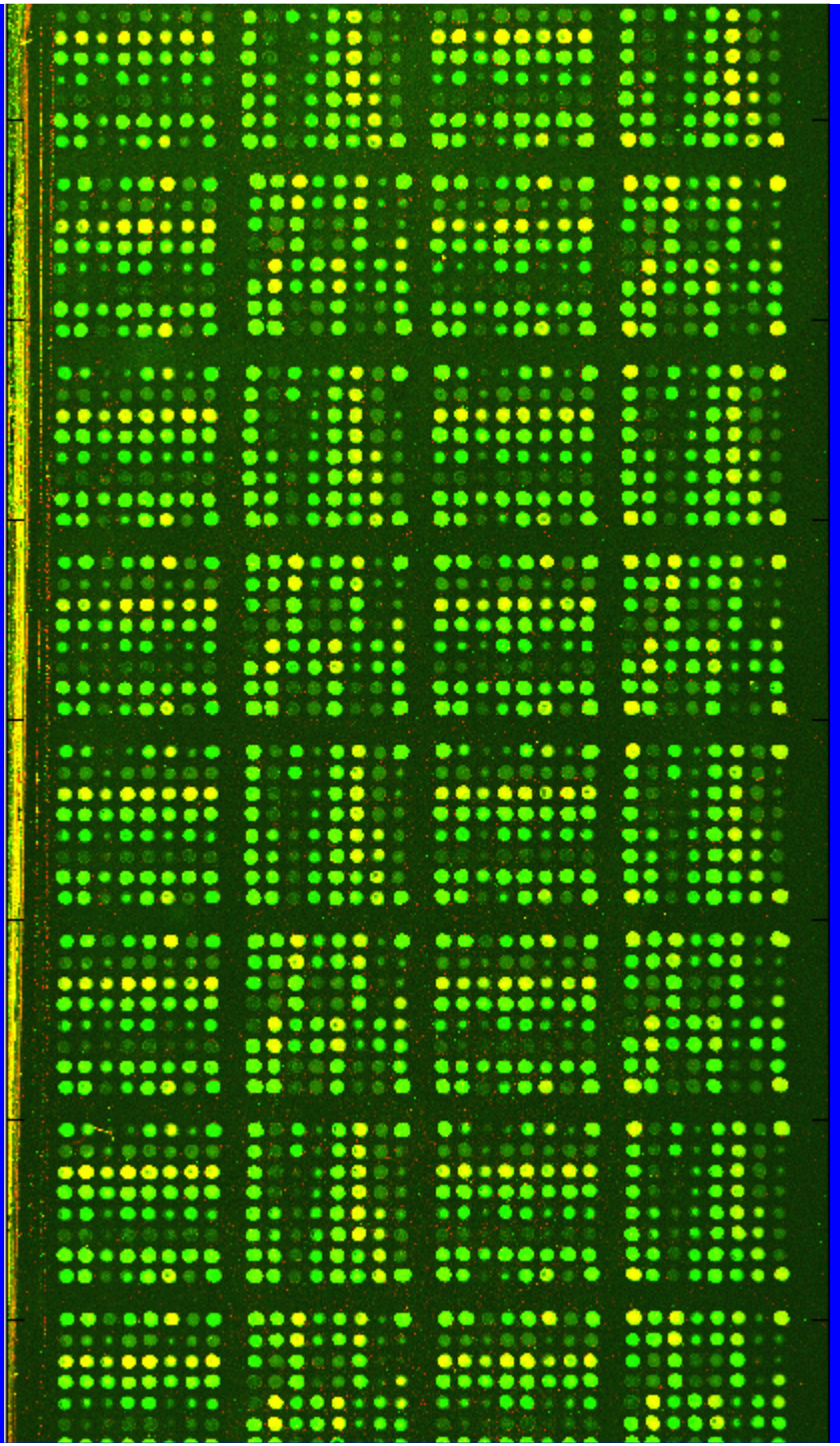
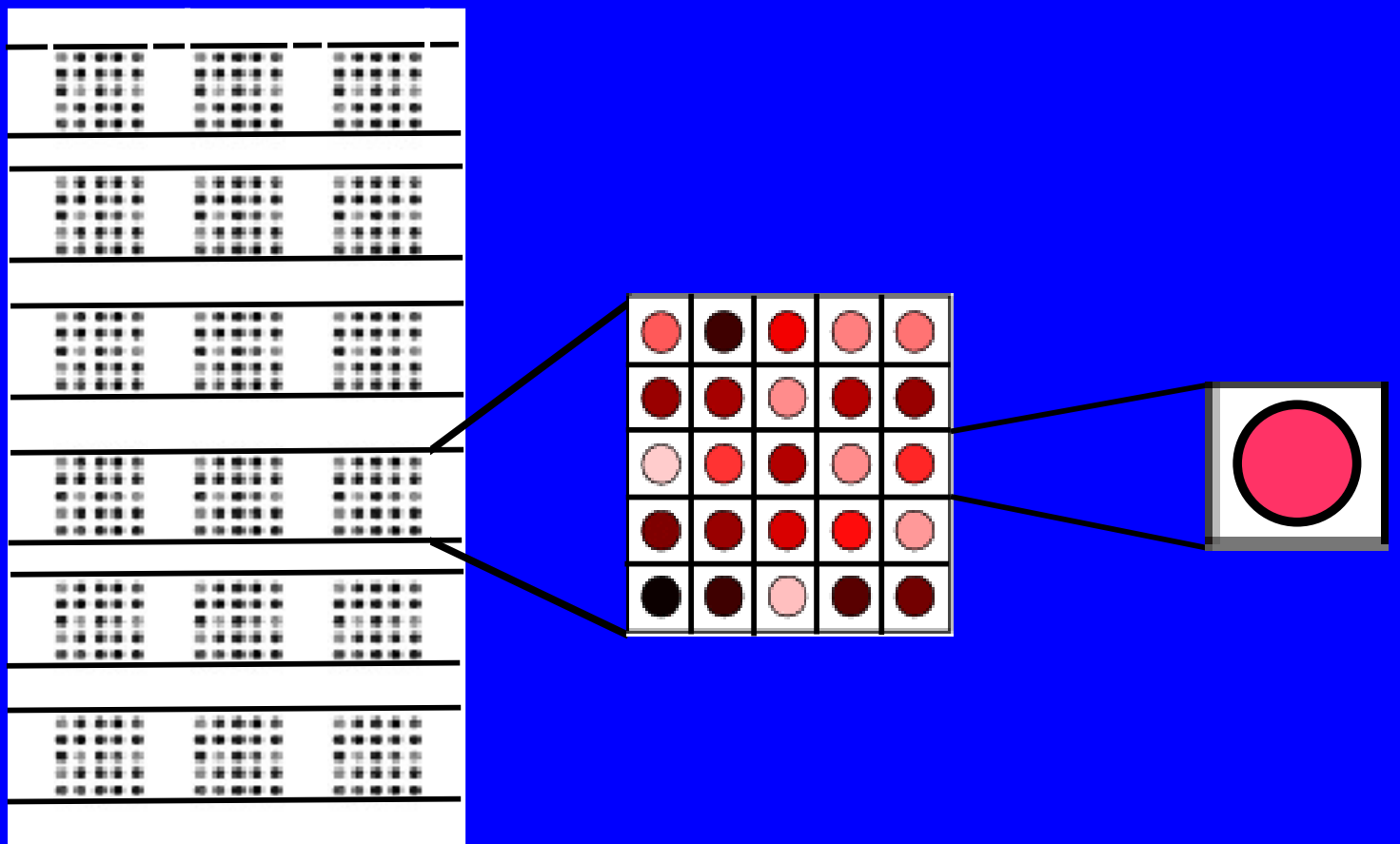


Image Analysis

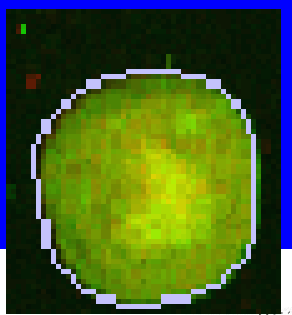
A scanned image of a microarray slide

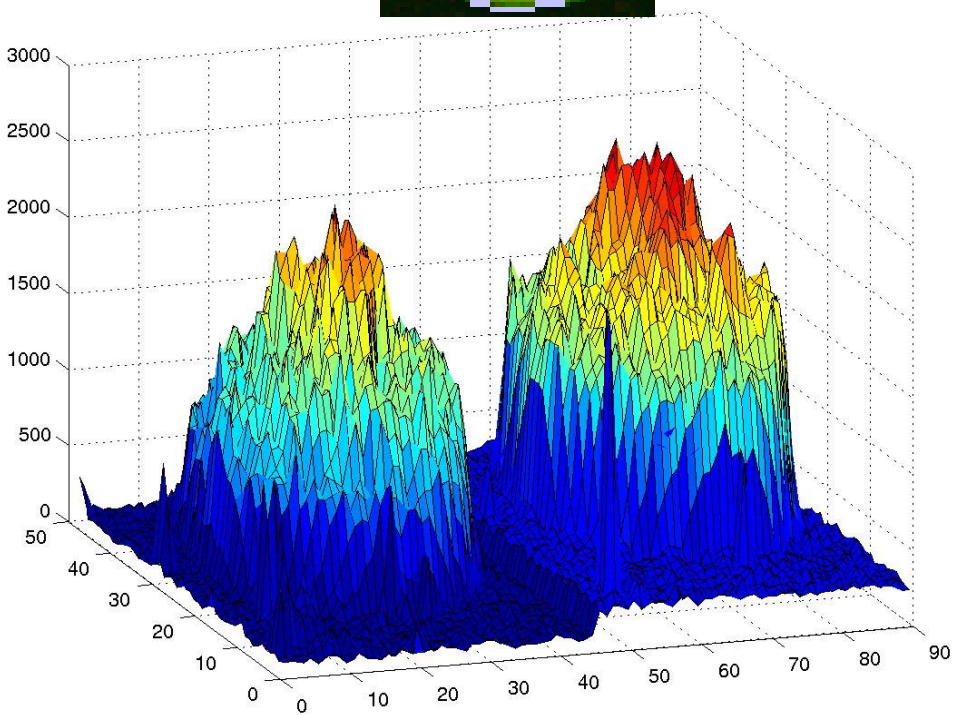


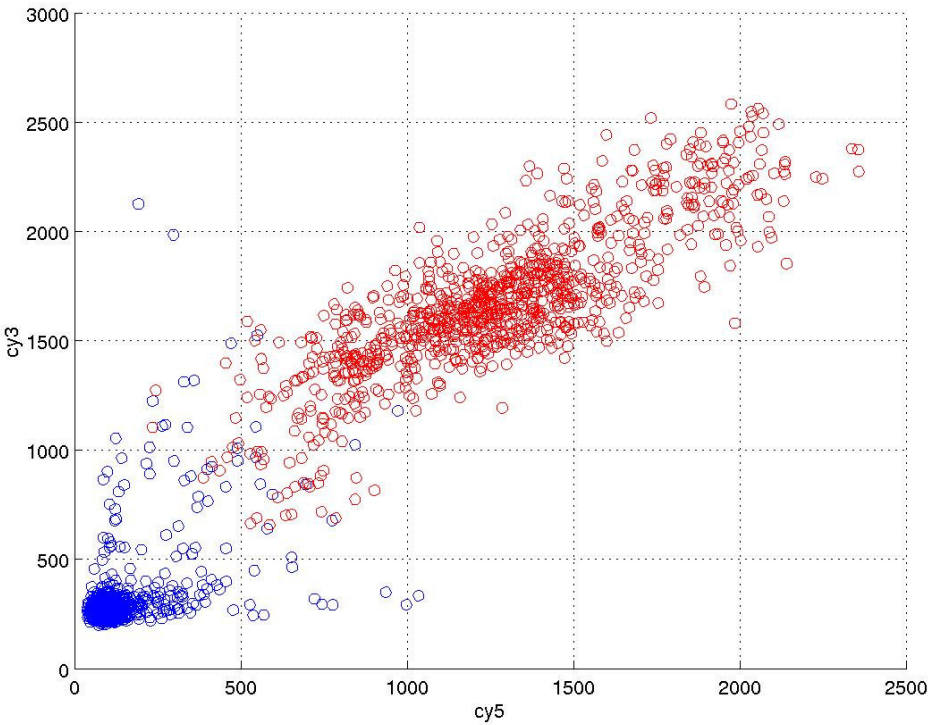
Processo de Segmentação de Imagens de Microarrays



Raw data to the gene expression estimation step







Available solutions

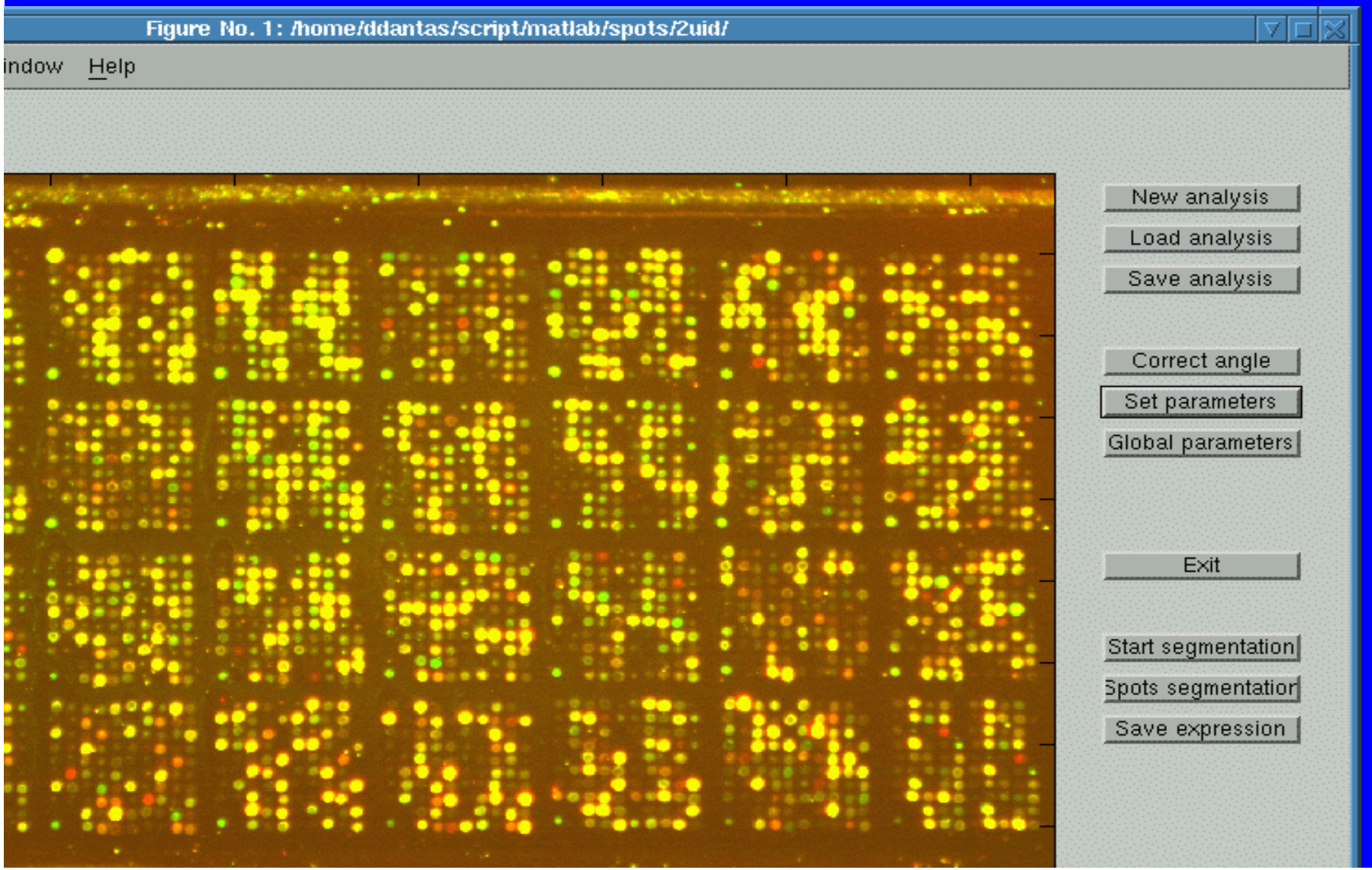
- Scanalyze: usually doesn't find misaligned spots.
- SpotFinder(TIGR): subarrays must be placed manually.
- Arrayvision: very good on locating misaligned spots; many options.
- UCSF Spot: does everything automatically if the image is perfect.
- Quantarray, F-scan, Dapple, Genepix, Imagene etc.

- All of them require user interaction to some level.

Our aim...

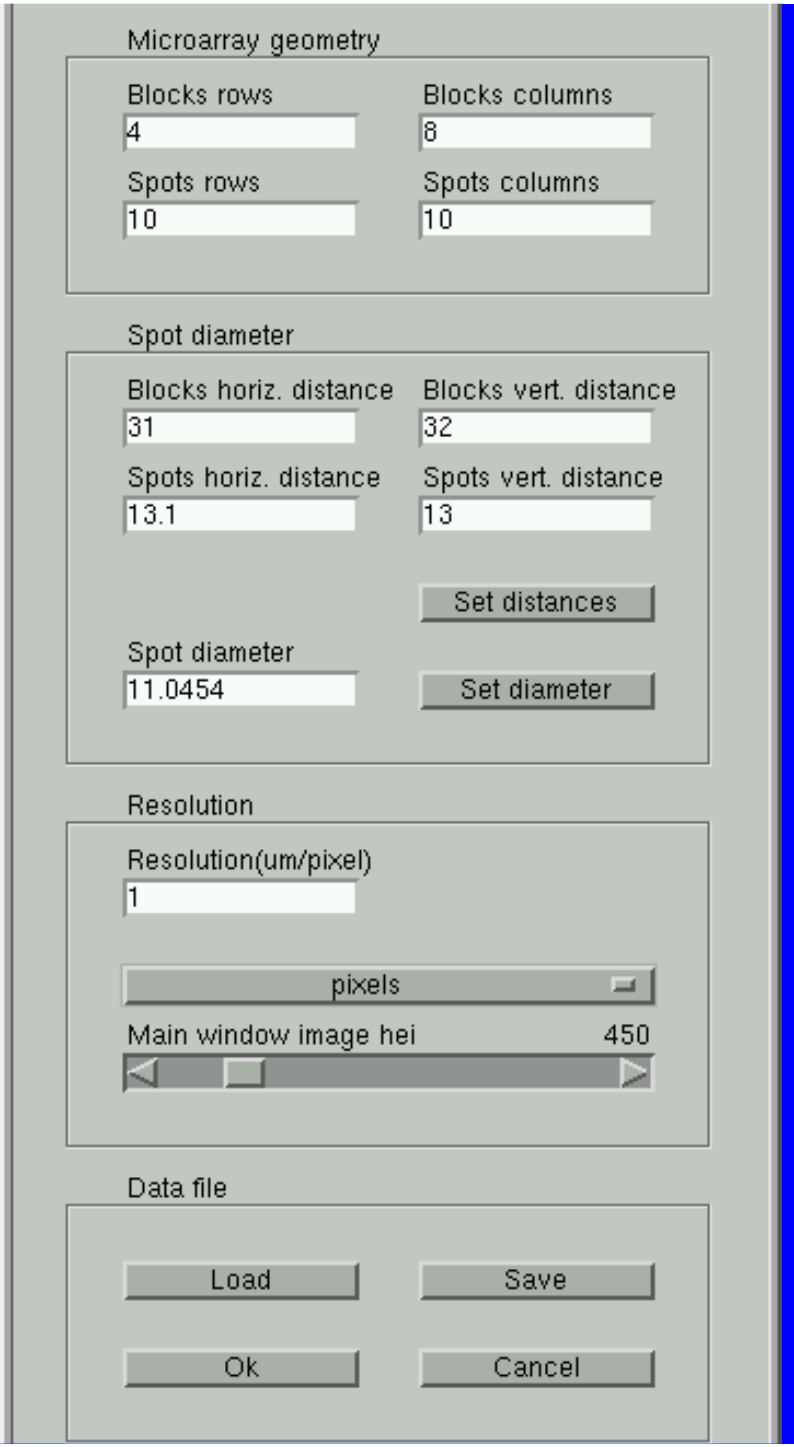
- Is to reduce the user interaction, doing the job automatically and measuring correctly the relative mRNA concentrations.
- This will make the process cheaper and faster.
- User interaction makes the segmentation subjective. Eliminating that, the results may be more reproducible.

Our software



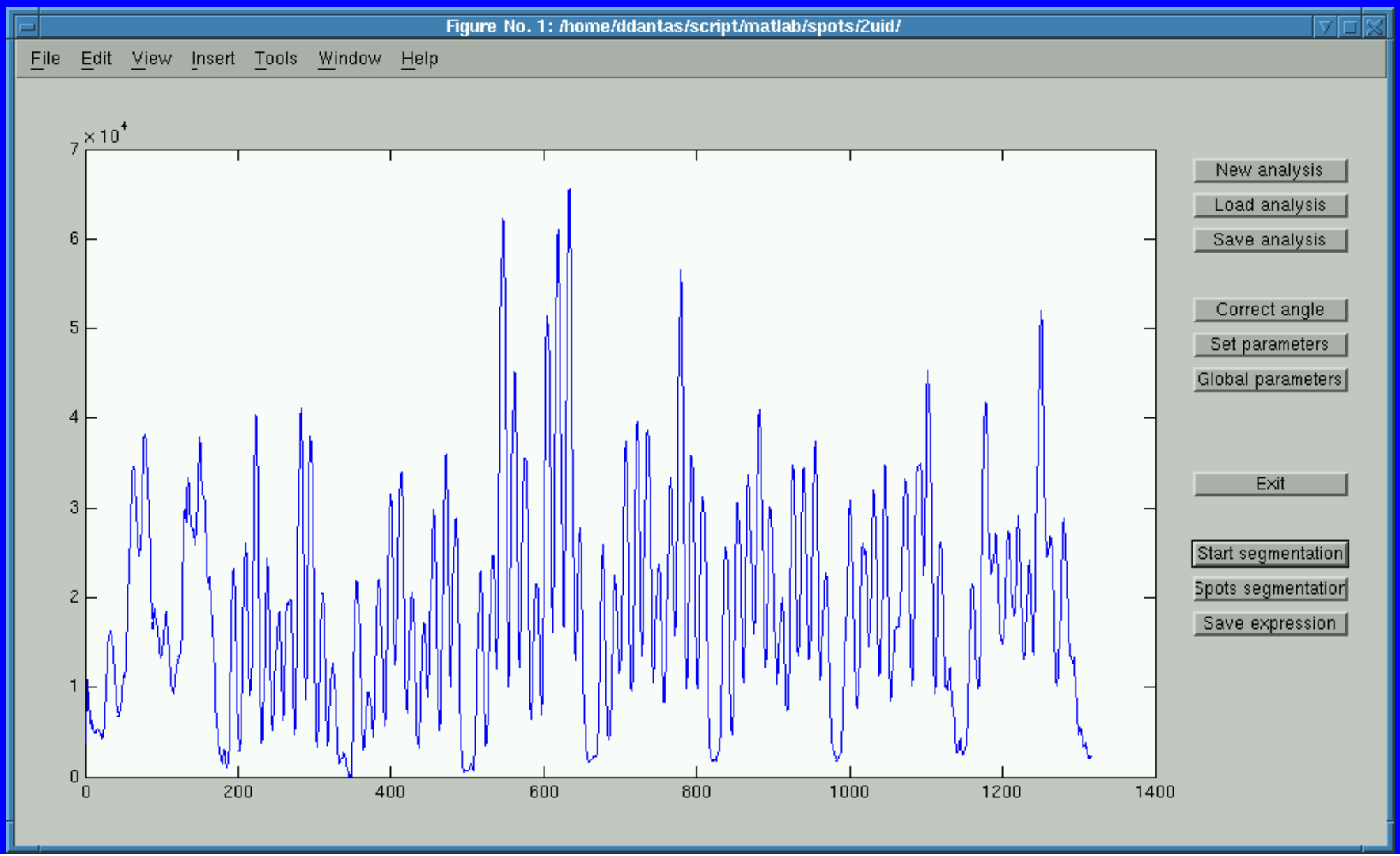
Parameter setting

- In this window the user sets parameters for a whole family of arrays
- He can save in a file for reusing them



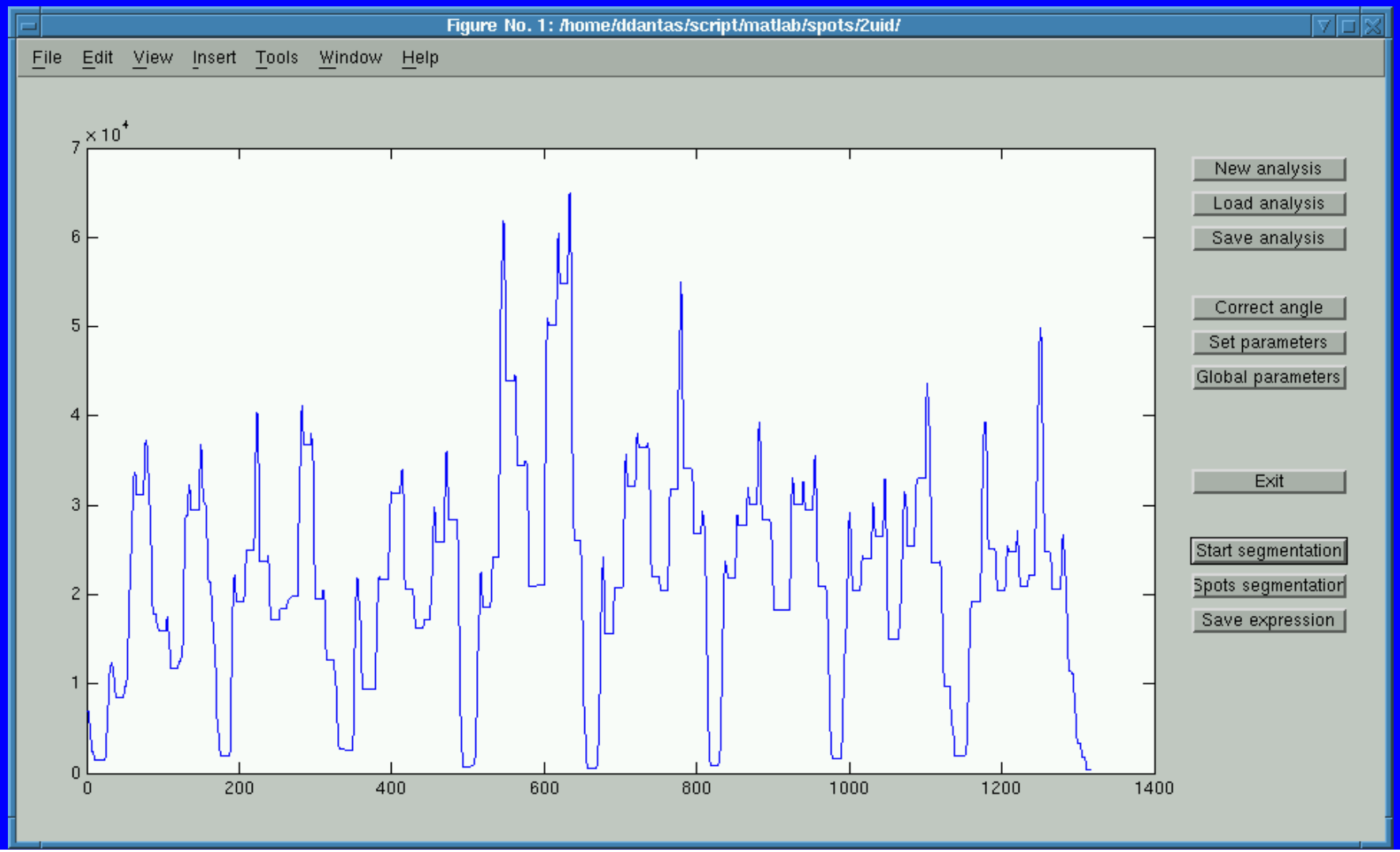
A vertical image profile…

is the sum of the spots values of each image line



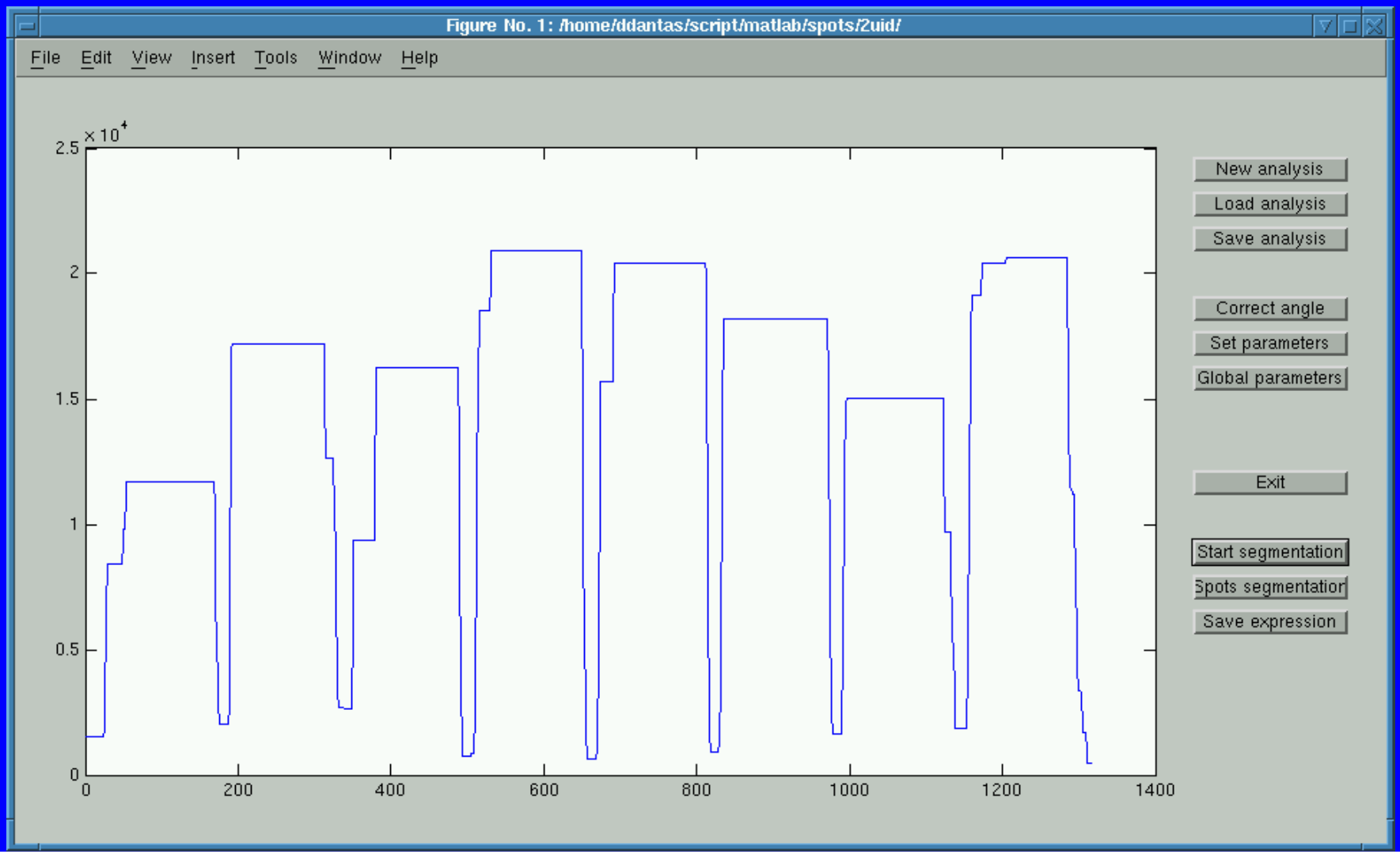
The subarray gridding…

Is done by filtering the horizontal and vertical profiles



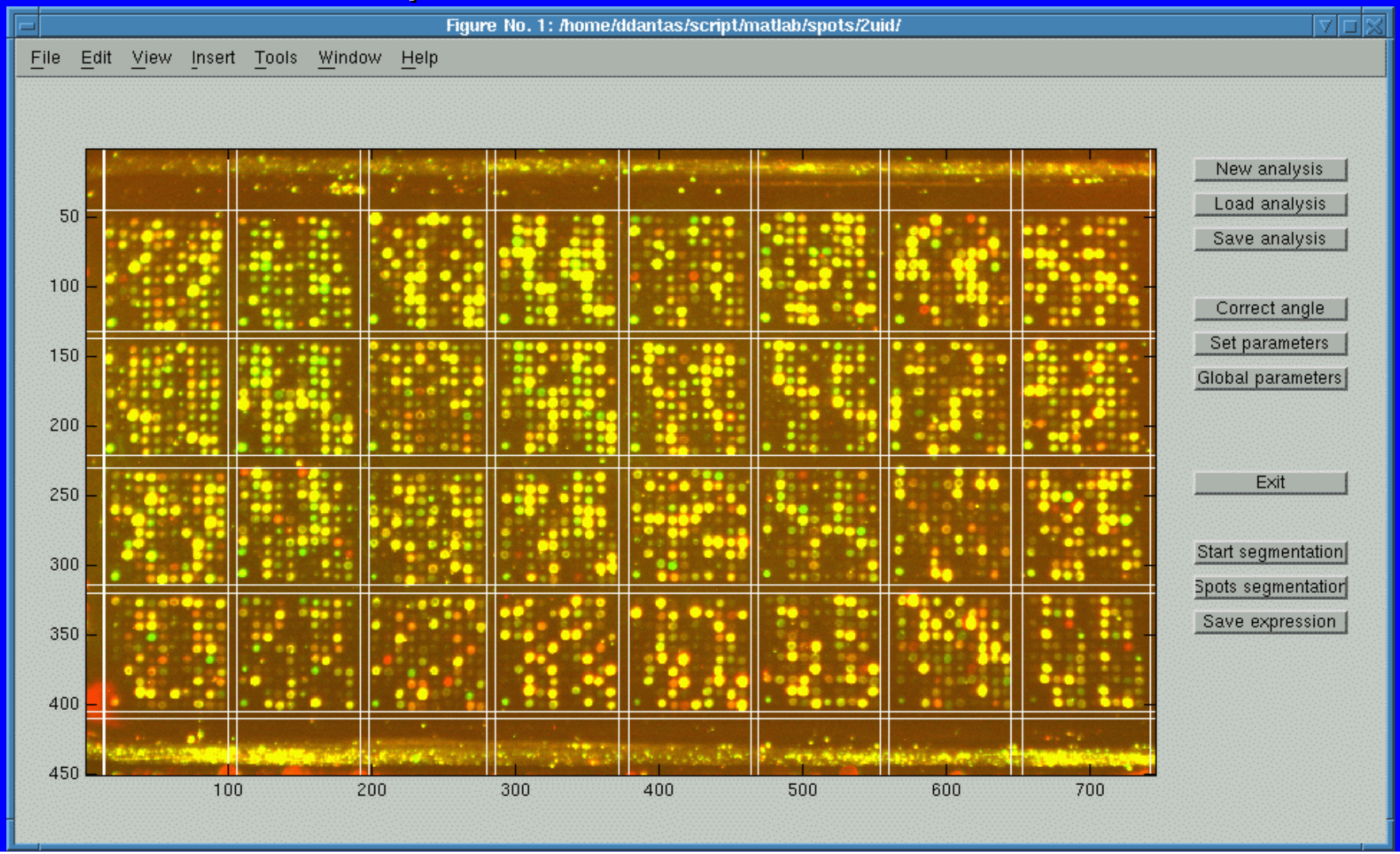
And finally…

taking the local minima of the filtered profile



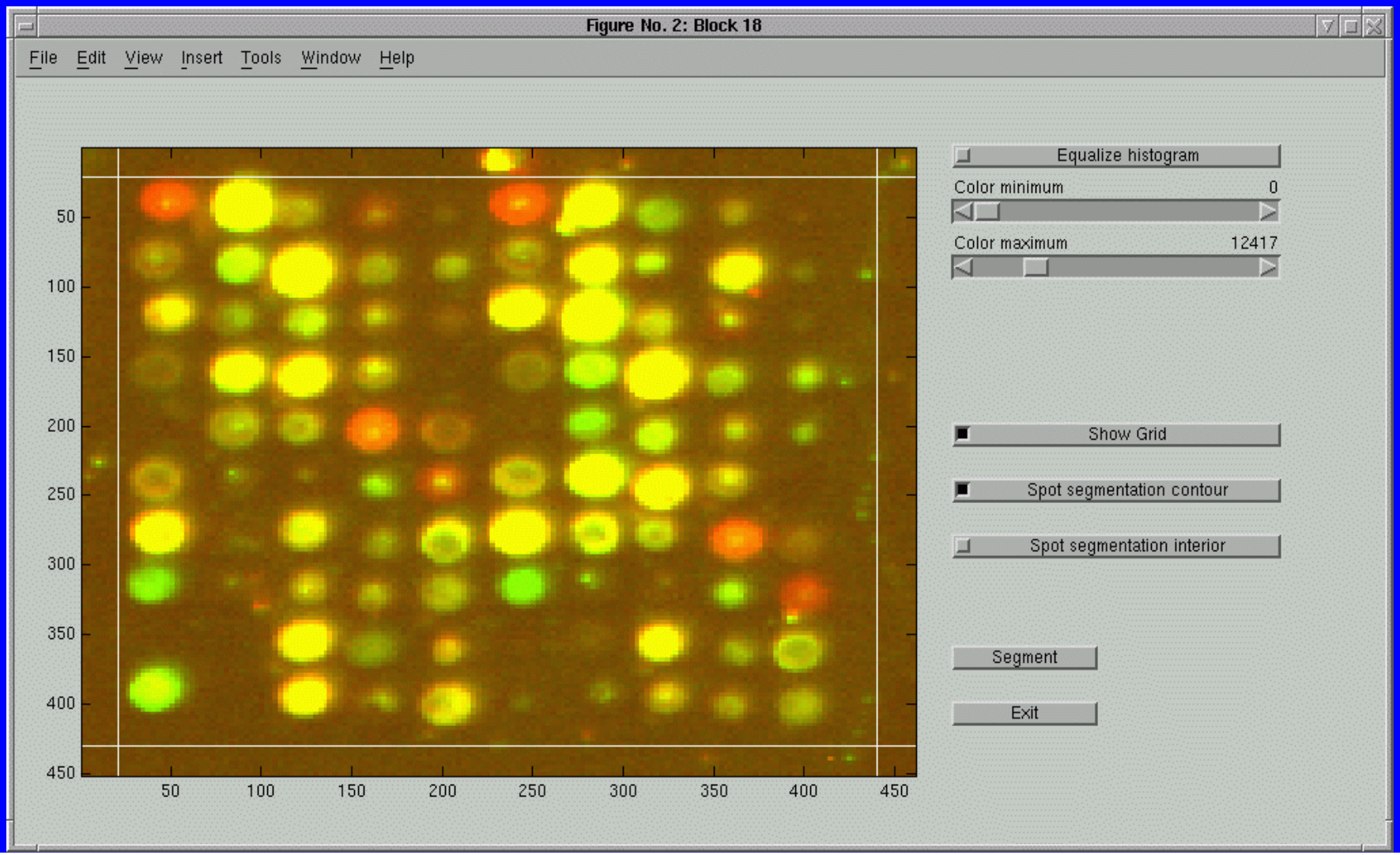
the same is done with…

the horizontal profile. Here the result



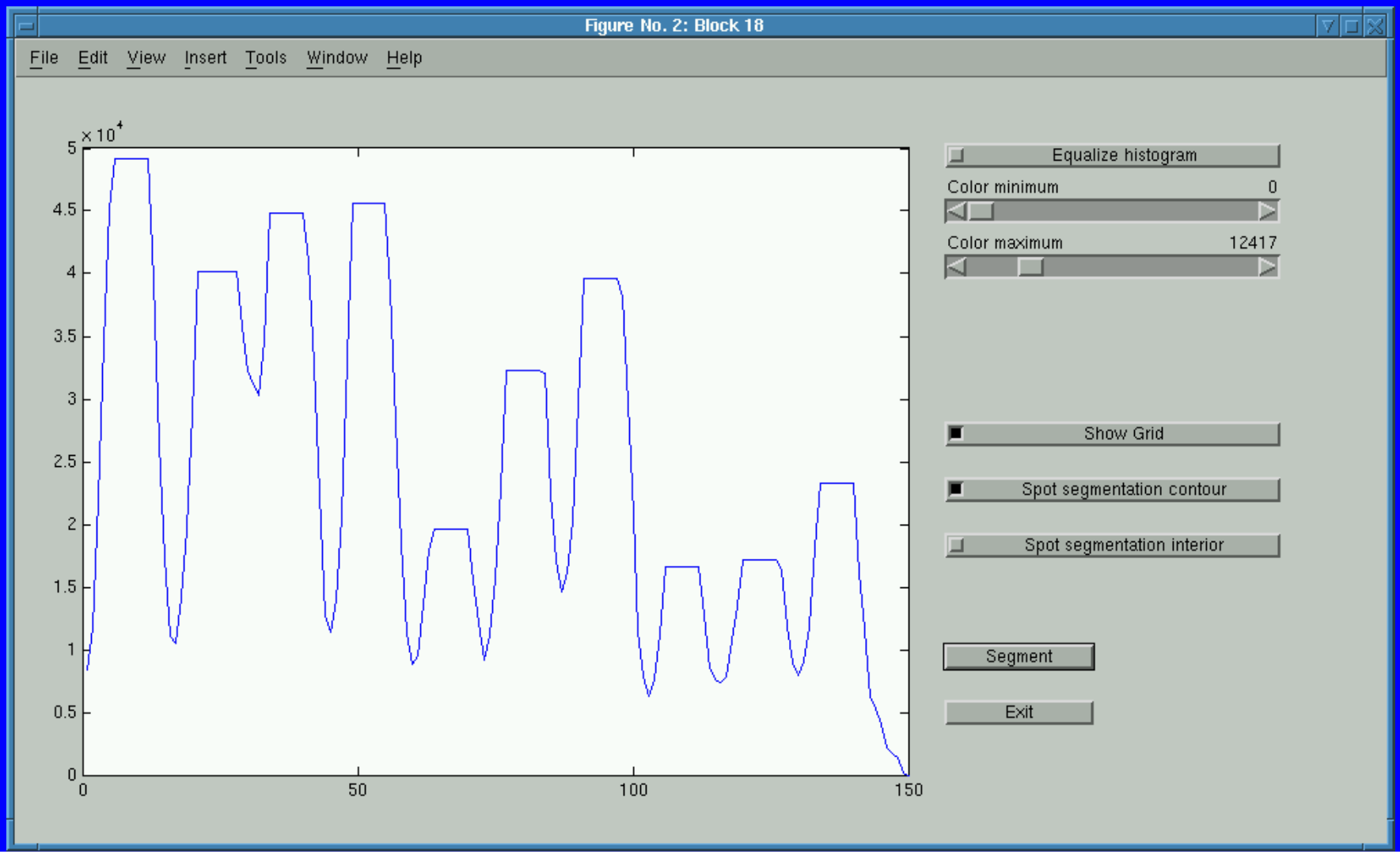
Spots gridding…

is done separately for each subarray



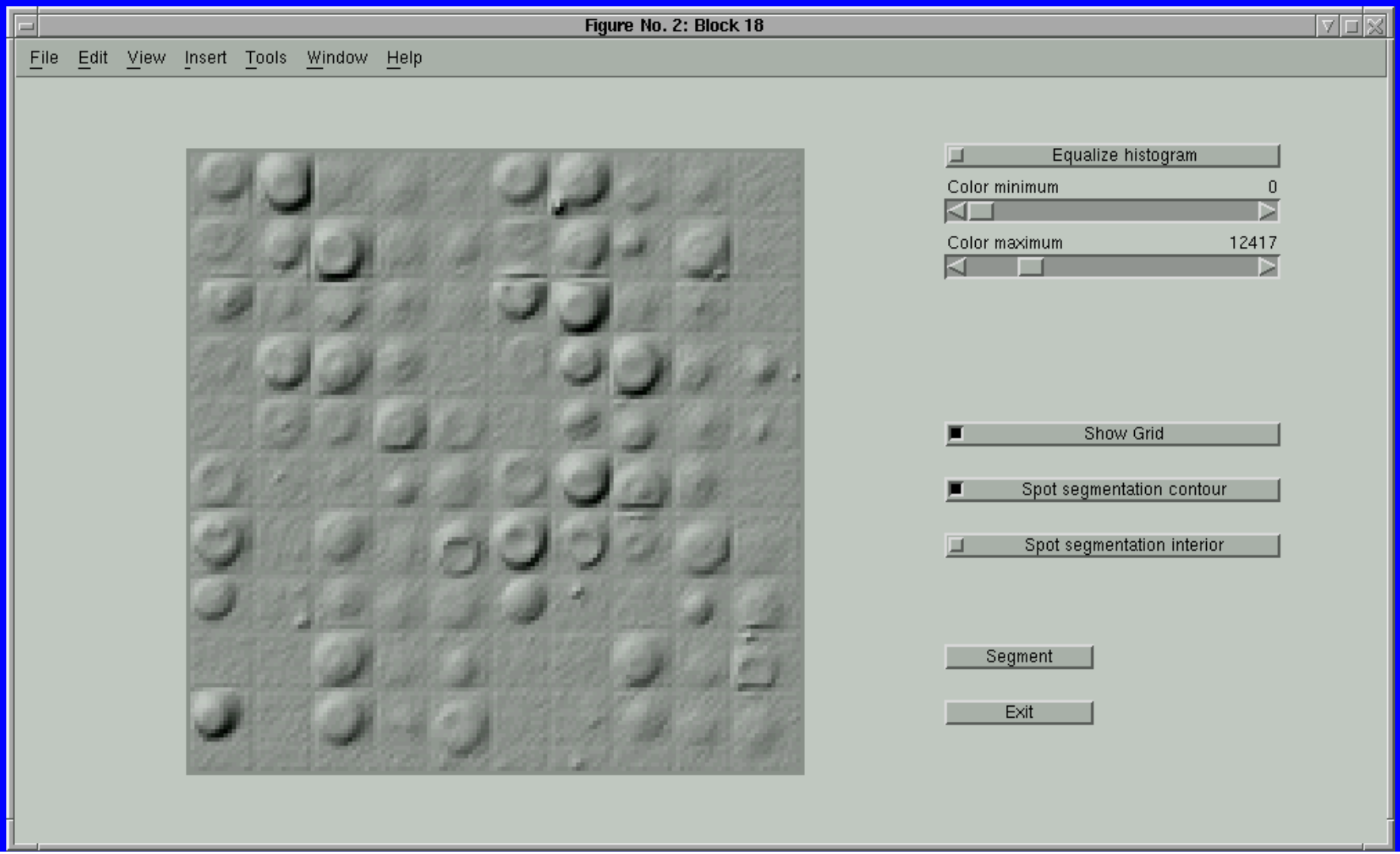
The profile filtering is simpler…

having just one step, and also uses local minima



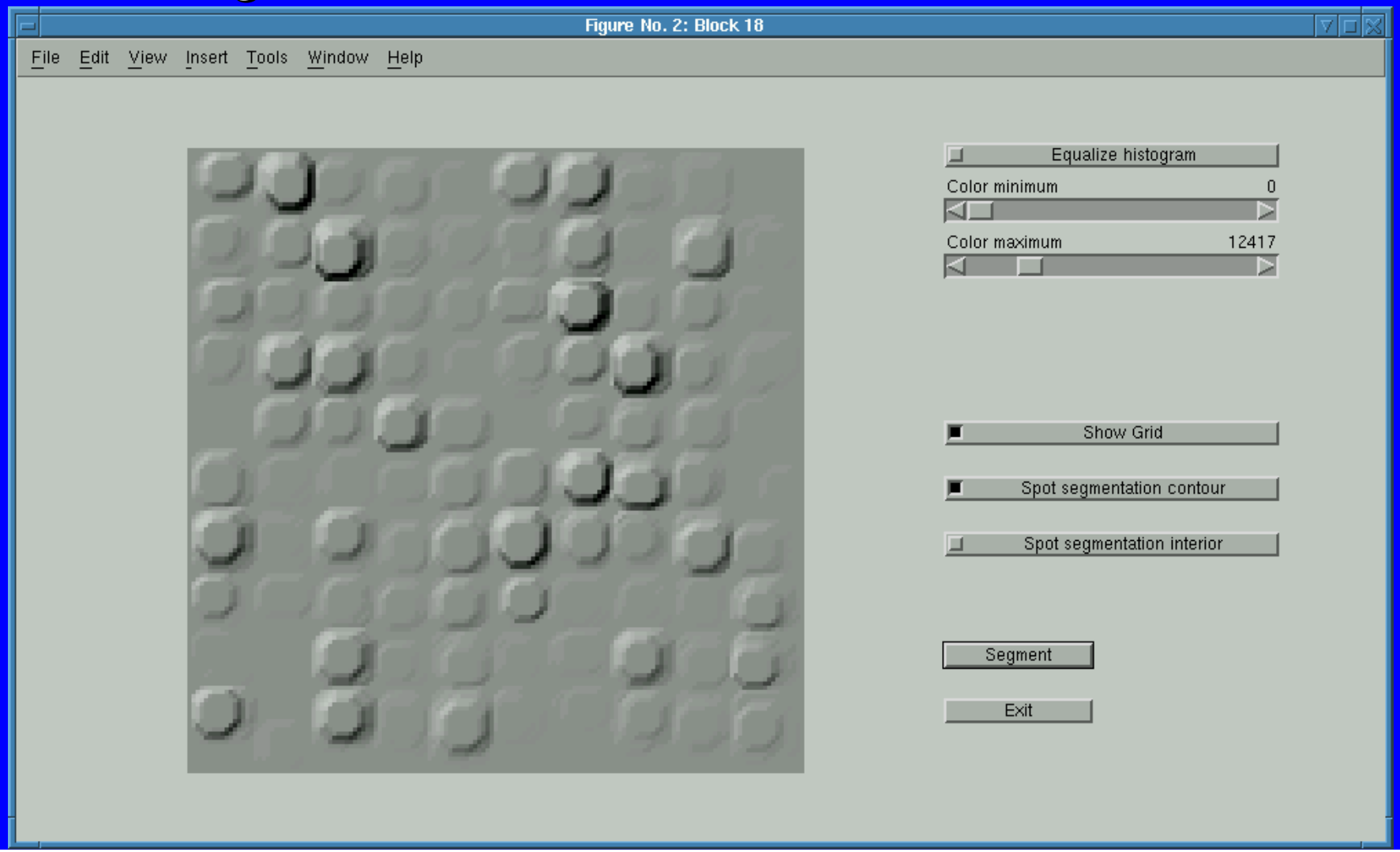
The spots detection step…

is basically the application of the Watershed operator



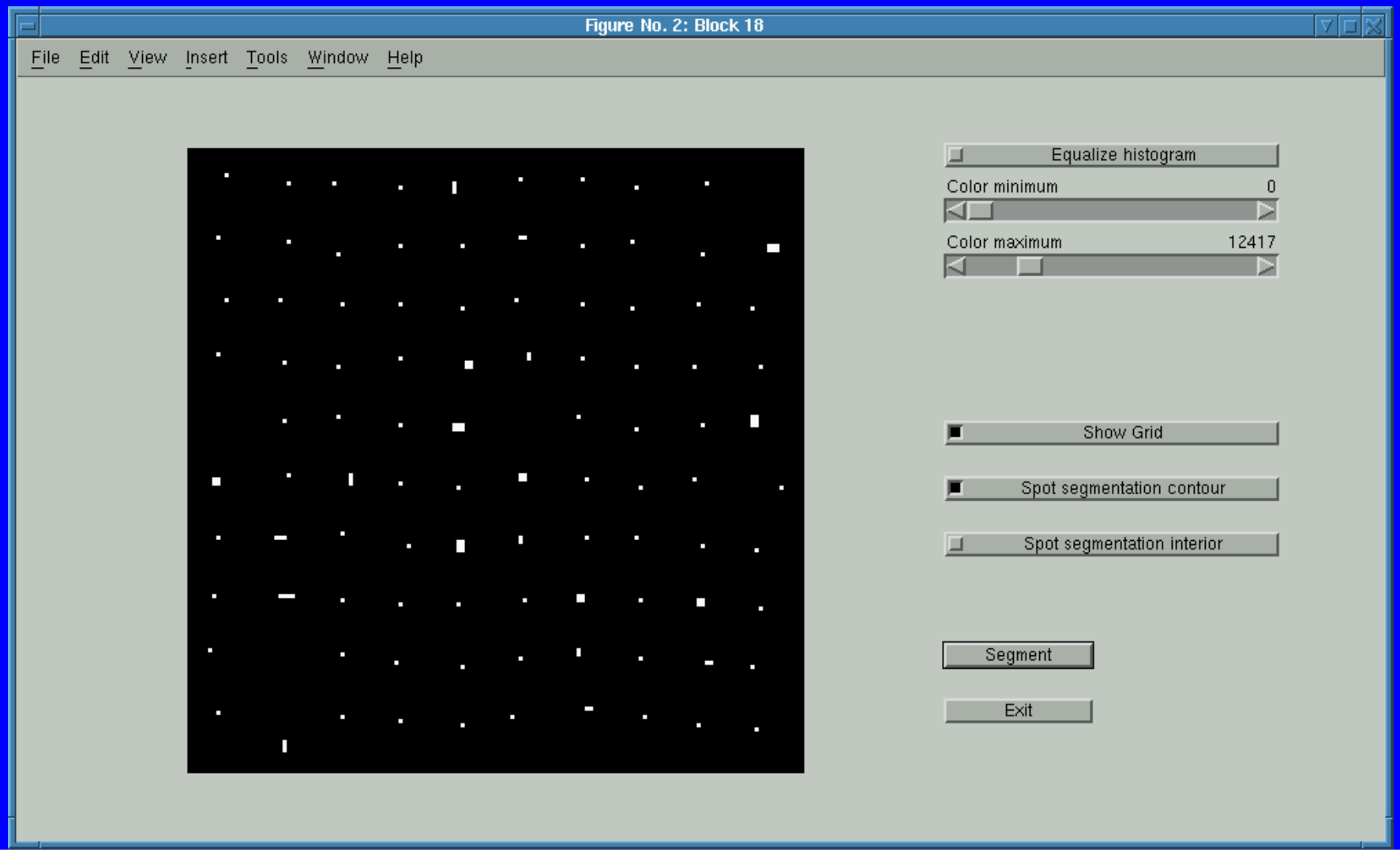
To avoid oversegmentation…

the image must be filtered



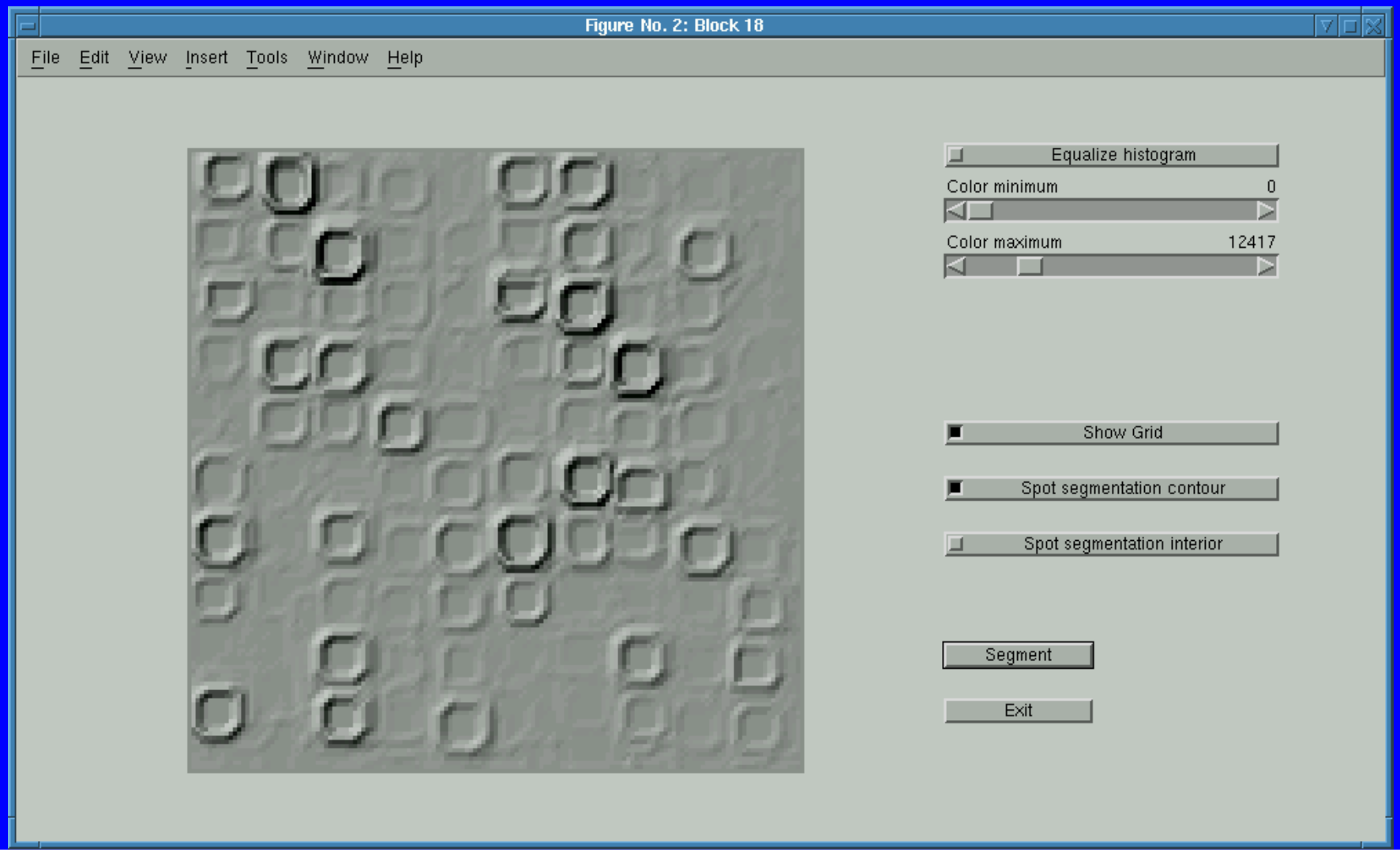
The filtered image also gives…

markers that will be used in Watershed



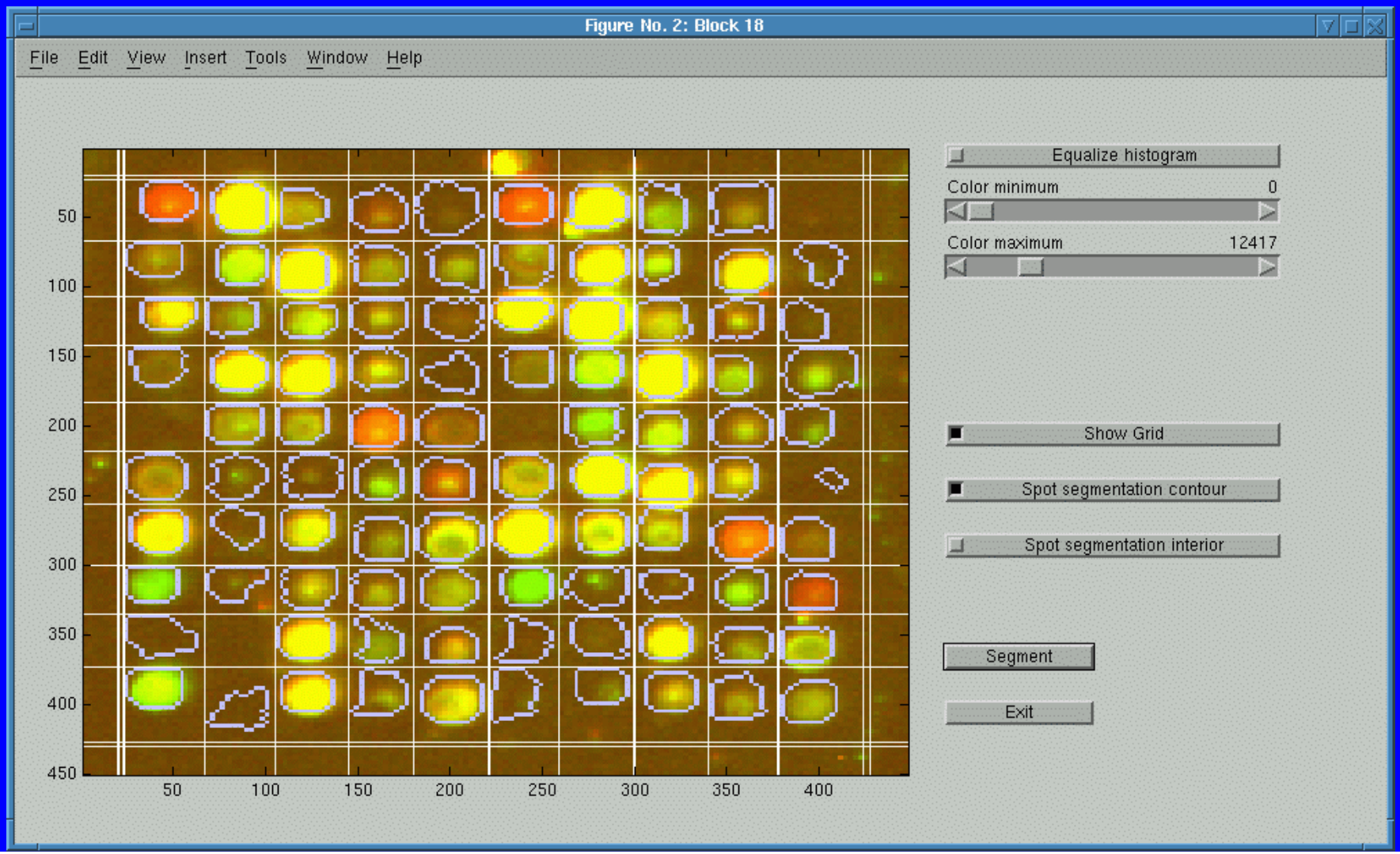
We give as input to the Watershed…

the markers, grid and the filtered image gradient

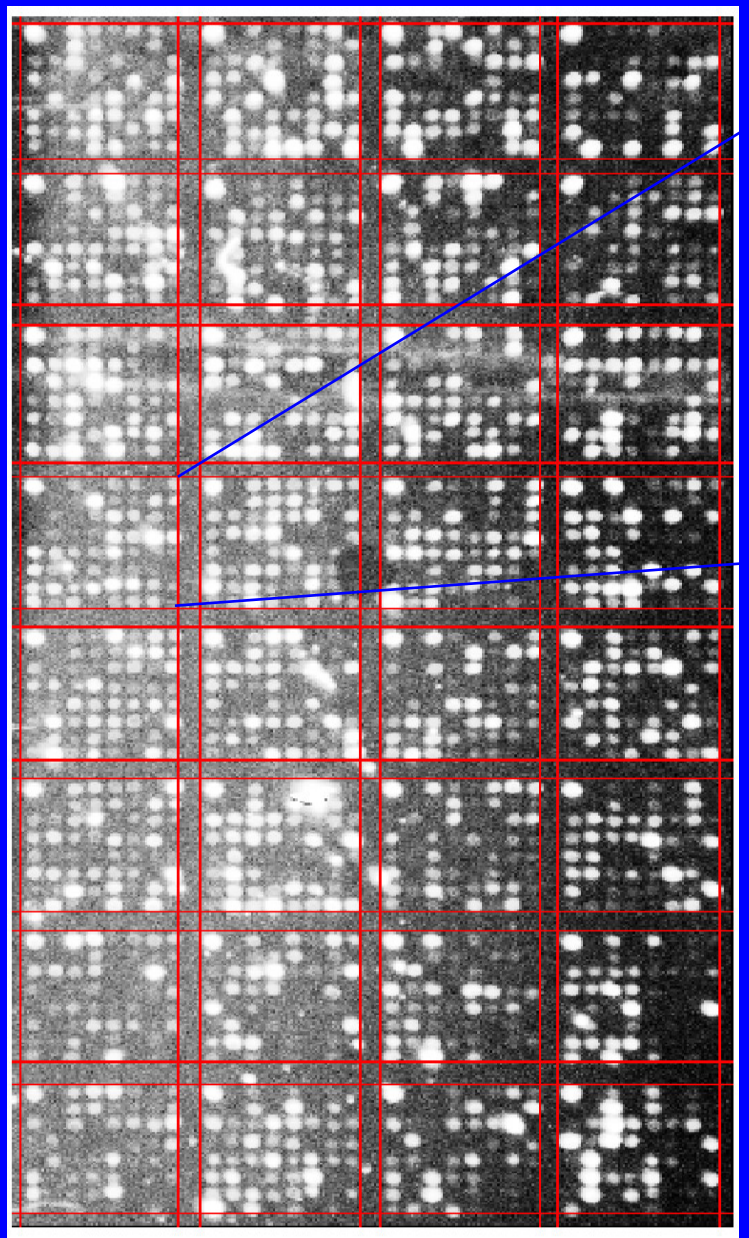


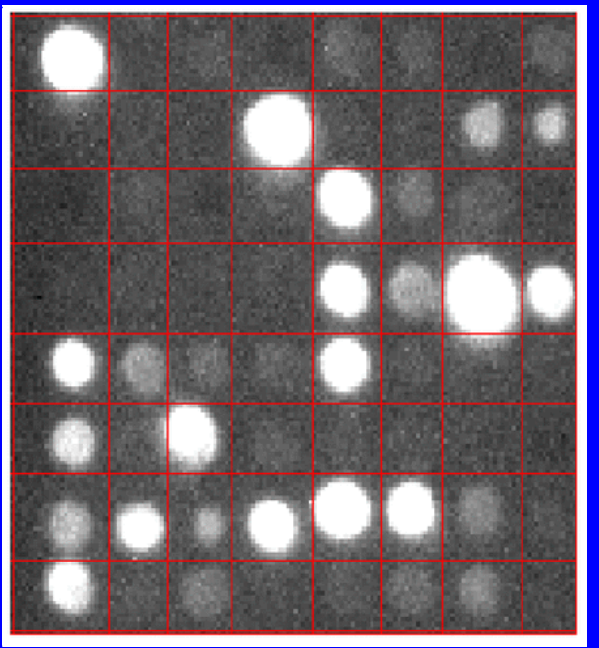
Here the resulting…

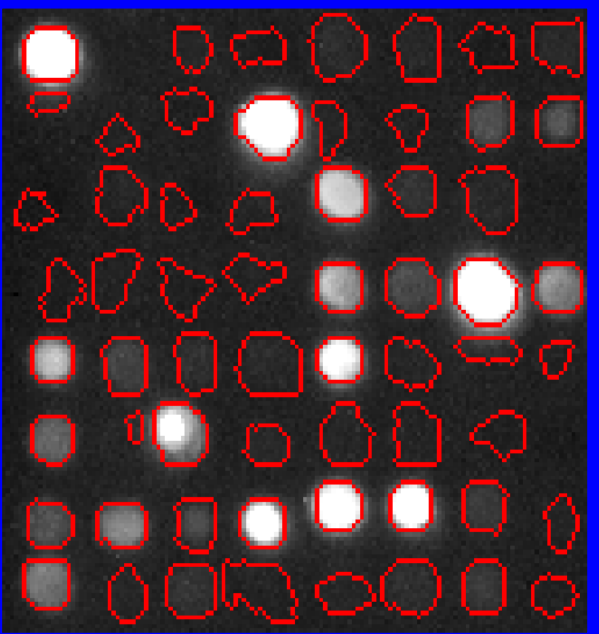
grid in white and spots cortours in light blue



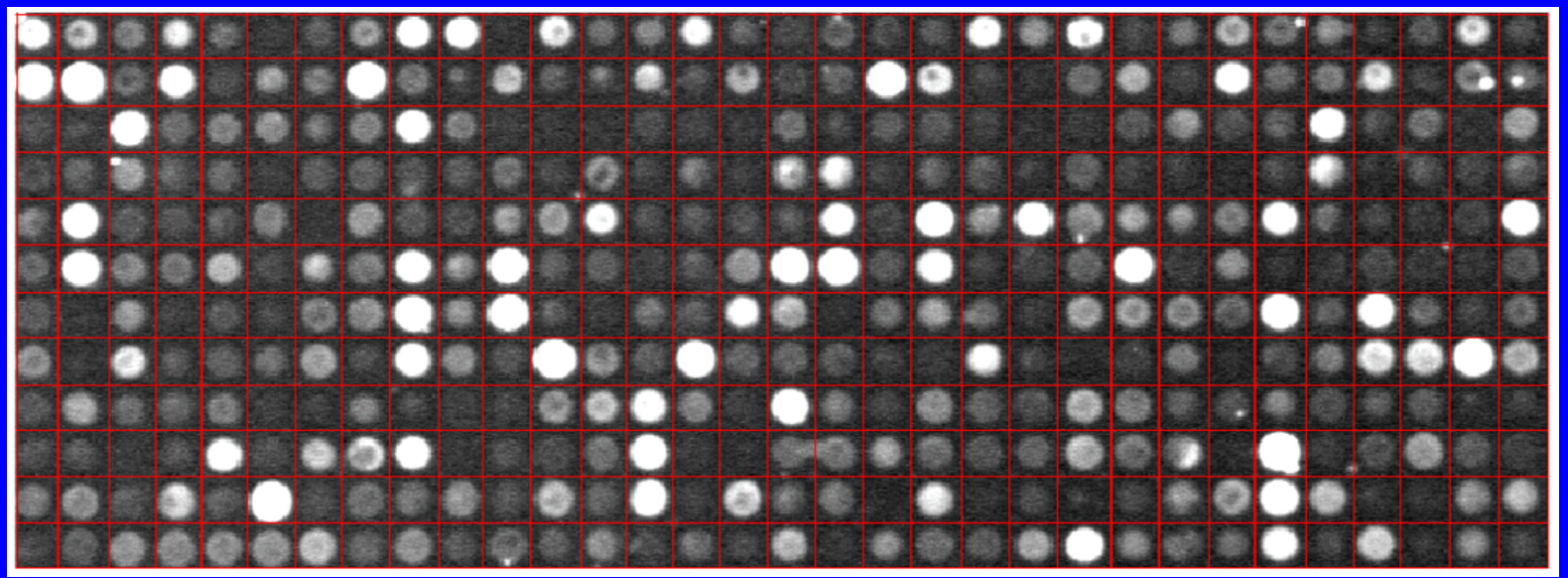
Segmentation example

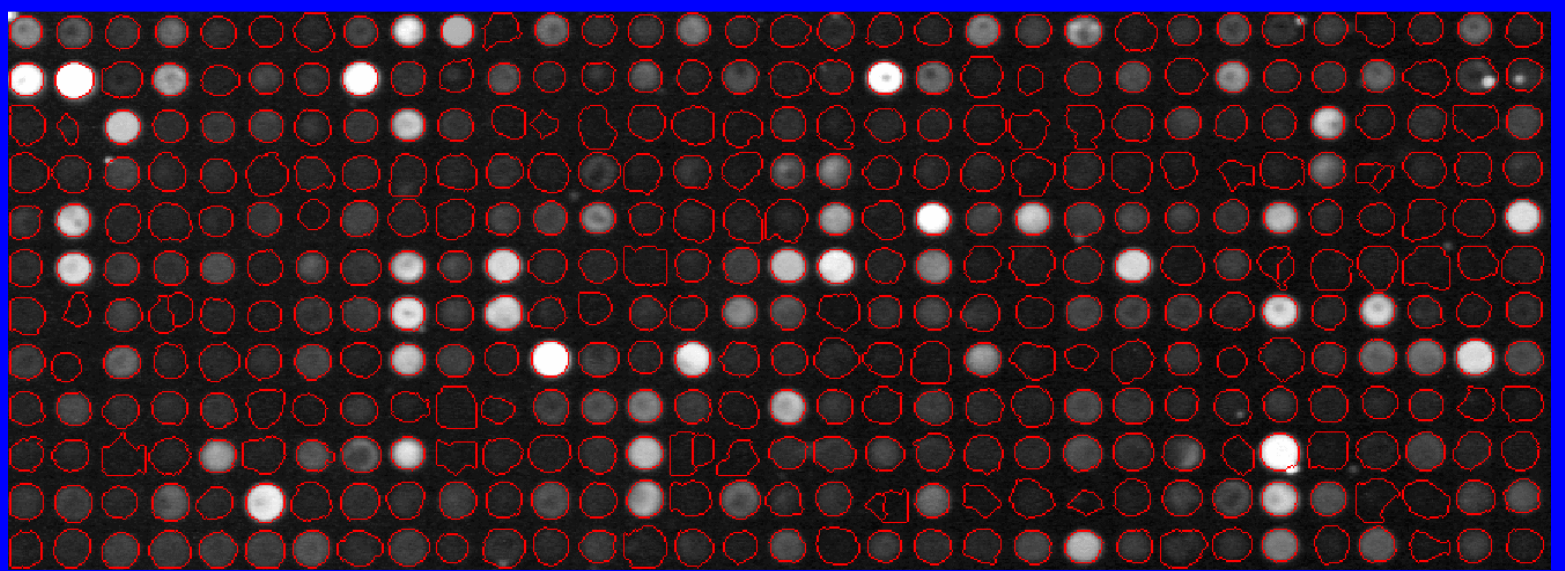






Segmentation example

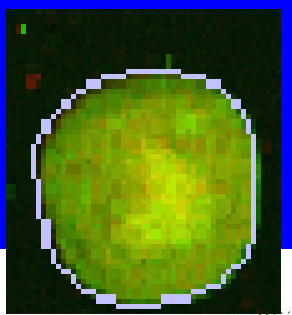


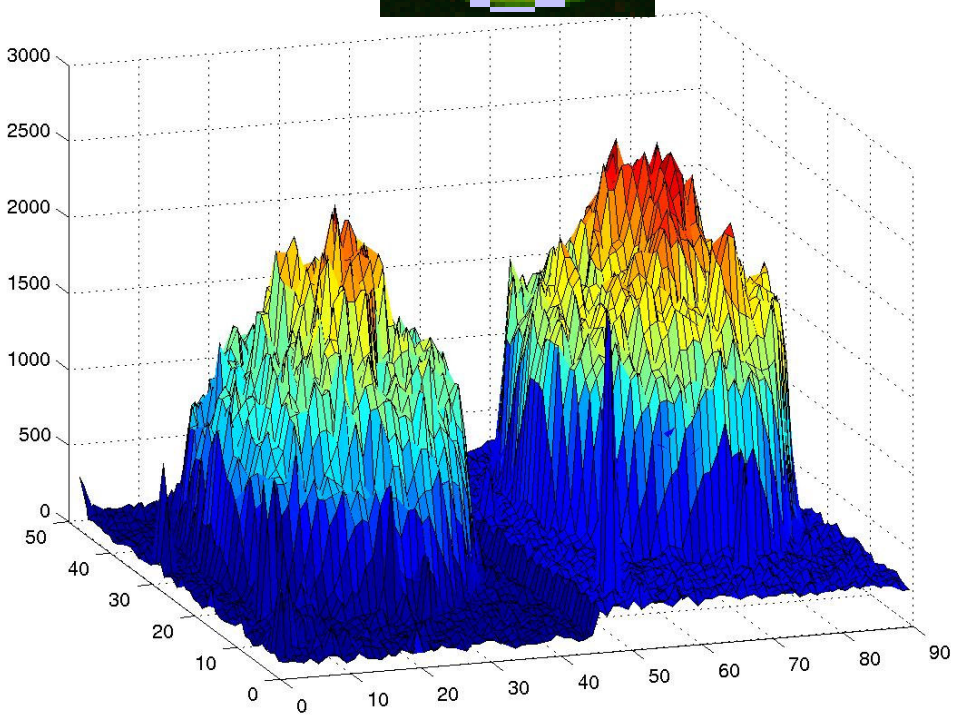


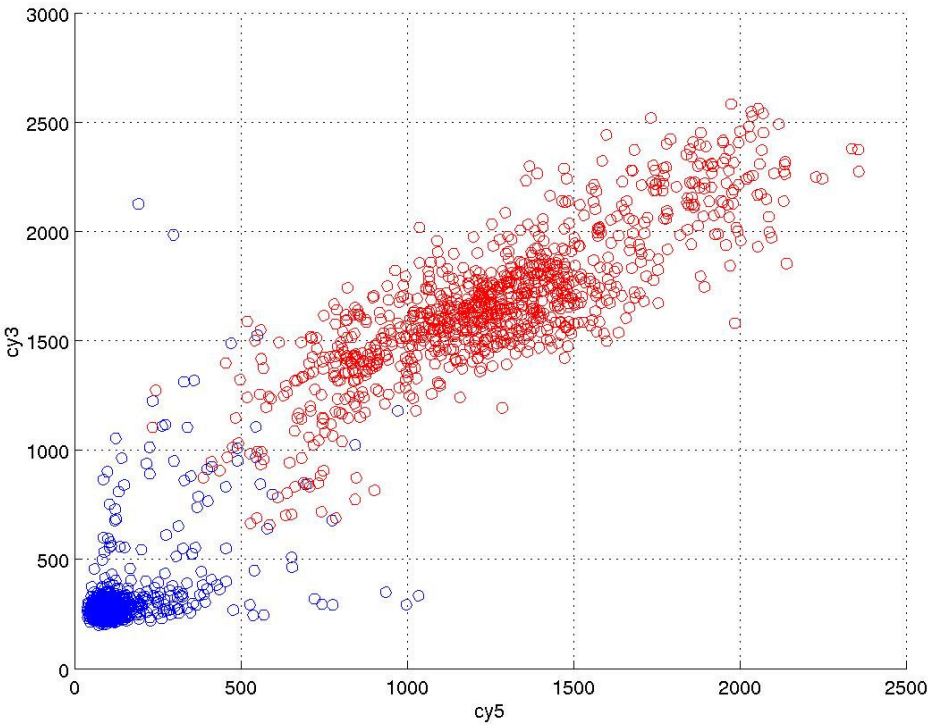
Raw data to the gene expression estimation step

- The raw data of a spot consists on:
  - the pixels values of both channels inside its rectangular region of interest
  - which pixels belong to foreground or backround
- Foreground is the region with spotted cDNA
- Background is the region without it.

Raw data to the gene expression estimation step





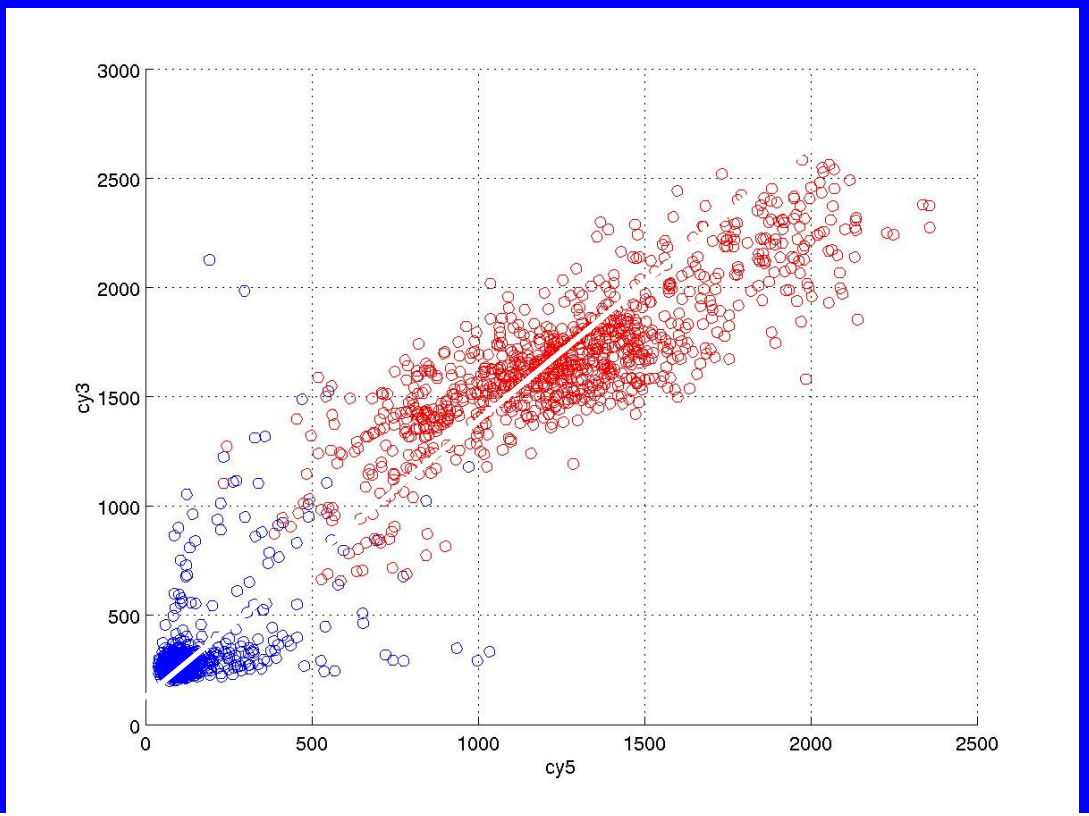


Gene expression estimation

- Is to find a value that represents the relative quantity of mRNA in the two samples.

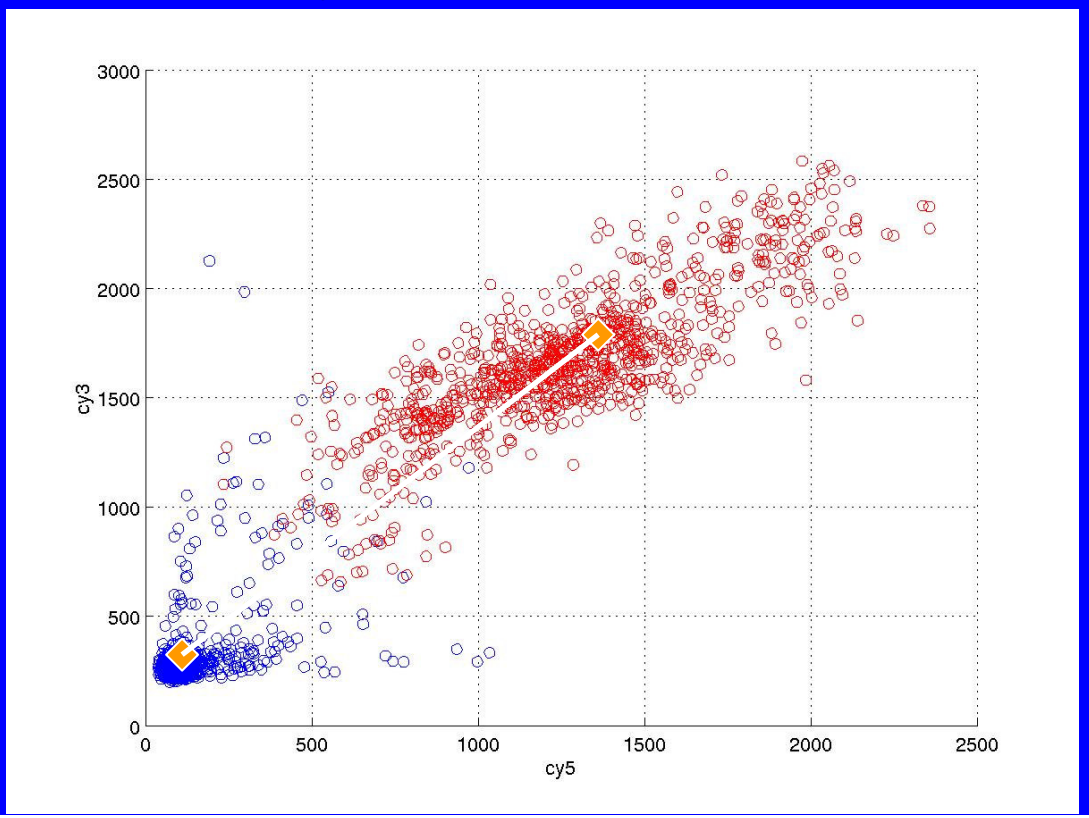
Some techniques to estimate gene expression

- Linear regression or least-squares fit of the values of pixels in the two channels.



Some techniques to estimate gene expression

- (ch1i-ch1b) / (ch2i-ch2b) where chXi is the estimated foreground intensity and chXb is the estimated background intensity of channel X.



Some techniques to estimate gene expression

- To estimate chXi and chXb we can do:
  - mean or median of all pixels in the foreground and background.
  - mean or median of some percentiles in the foreground and background (fixed region method)
  - mean or median of higher percentiles of all the pixels in the rectangle to estimate chXi and of lower percentiles to estimate the chXb. Foreground and background information is ignored (histogram method)

Gene expression generation

```
HEADER      SPOT  GRID  TOP   LEFT  BOT   RIGHT ROW   COL   CH1I  CH1B  CH1AB CH2I  CH2B  CH2AB SPIX  BGPIX EDGE  RAT2  MRAT  REGR
      CORR  LFRAT CH1GTB1     CH2GTB1     CH1GTB1     CH2GTB2     CH1EDGEA    CH2EDGEA    FLAG  CH1KSD      CH1KSP      CH2KSD
      CH2KSP
REMARK
REMARK
REMARK
REMARK
REMARK      DATE  21-Mar-2002
REMARK      TIME  16:51:57
SPOT  1     1     88    24    102   40    1     1     7341  6704  6818  4669  4016  4312  105   150   0     0     0.9624      0.636
      1.0000      0.0000      0.8952      0.9143        0   0.01905     -8.68E+03   -3.30E+03   0         0   0.00E+00        0   0.00E+00
SPOT  2     1     88    40    102   53    1     2     7075  6704  6708  4419  3920  3989  89    121   0     0     0.9751      0.5965
      1.0000      0.0000                                                                                    E+00        0   0.00E+00
SPOT  3     1     88                                                                                        0.6031      0.5906
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  4     1     88                                                                                        1.422 0.7494
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  5     1     88                                                                                        1.018 0.8077
      1.0000      0.0000                                                                                   0.00E+00
SPOT  6     1     88                                                                                        0.9687      0.6442
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  7     1     88                                                                                        1.125 0.6169
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  8     1     88                                                                                        1.044 0.6949
      1.0000      0.0000                                                                                   0.00E+00
SPOT  9     1     88                                                                                        0.7028      0.7353
      1.0000      0.0000                                                                                       0   0.00E+00
SPOT  10    1     88                                                                                        0.4444      0.4603
      1.0000      0.0000                                                                                   E+00        0   0.00E+00
SPOT  11    1     102                                                                                       1.417 0.3783
      1.0000      0.0000                                                                                   0.00E+00
SPOT  12    1     102   40    118   53    2     2     10875 7552  7803  7889  5232  5464  75    163   0     0     0.7956      0.8394
      1.0000      0.0000        1     1   0.2533      0.5467      -3.89E+04   -9.60E+04   0         0   0.00E+00        0   0.00E+00
SPOT  13    1     102   53    118   68    2     3     9546  7136  7245  6922  4688  4759  98    174   0     0     0.9084      0.8939
      1.0000      0.0000        1   0.9898      0.1327      0.4286      1.48E+04    -1.99E+04   0         0   0.00E+00        0   0.00E+00
SPOT  14    1     102   68    118   83    2     4     49743 9808  12869 26888 7288  8971  100   172   0     0     0.5192      0.6069
      1.0000      0.0000        1     1     1   0.99  -2.61E+05   -3.82E+05   0         0   0.00E+00        0   0.00E+00
SPOT  15    1     102   83    118   97    2     5     11367 8160  9020  7880  6032  6499  92    163   0     0     0.5999      0.6874
      1.0000      0.0000        1   0.9022      0.2609      0.1739      5.17E+04    1.76E+04    0         0   0.00E+00        0   0.00E+00
```

- The program saves the expression data in a tab separated text file
- The file has the same format of the ones generated by *ScanAlyze*

Validation

- We made controlled experiments to test the expression estimation techniques.
- The objective of the experiment was to test how expression was affected by:
  - position in the slide
  - dilution of cDNA
  - length of mRNA fragments
  - being marked with cy3 or cy5

Validation

- We spotted microarrays with 32 blocks, each block with

  6 genes x 5 dilutions x 2 repetitions + 4 landmarks = 64 spots

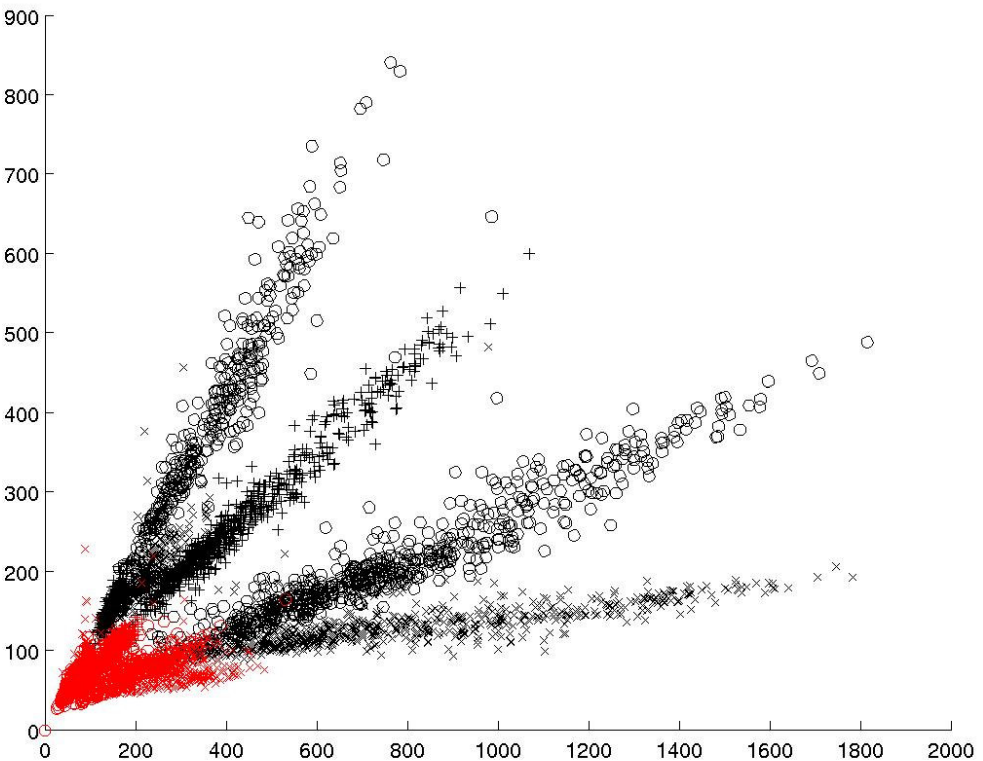
- We made six slides like this and, onto them, we poured six different mRNA soups:

	Dilution					
gene	43A	43B	44A	44B	45A	45B
Irf	1	5	1	2	1	10
Trp	1	5	1	2	1	10
ST0280	1	5	1	2	1	10
IL	5	1	2	1	10	1
Q	5	1	2	1	10	1
Lys	5	1	2	1	10	1

Validation

- Here each point is the value of a spot obtained by the fixed region method. Spots from different dilutions are grouped. The black ones are from the three bigger mRNA fragments, and the red, from the three smaller.

sqrt(($ch1i_A$-$ch1b_A$) x ($ch2i_B$-$ch2b_B$))

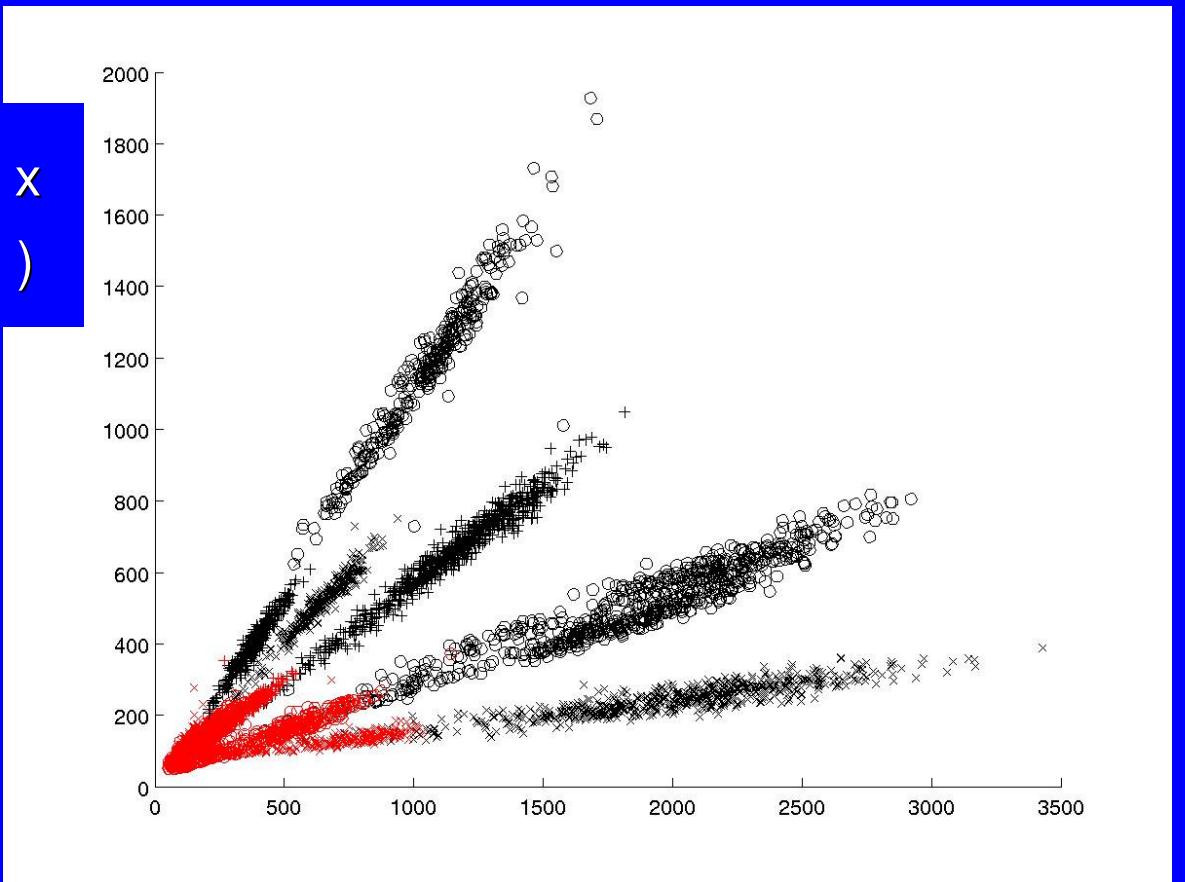


sqrt(($ch2i_A$-$ch2b_A$) x ($ch1i_B$-$ch1b_B$))

Validation

- And here is the best result, obtained with the histogram method.

sqrt(($ch1i_A$-$ch1b_A$) x ($ch2i_B$-$ch2b_B$))



sqrt(($ch2i_A$-$ch2b_A$) x ($ch1i_B$-$ch1b_B$))

Normalization

Normalization

- The expected expression of the gene IRF was 1.0 but the expression found was 1.6
- This is due to the physical properties of the dyes.

Normalization

- When we have a single slide, we must eliminate the constant *k* assuming, when appropriate, that
  - we can normalize all the spots using the expression of a housekeeping gene
  - we can normalize assuming that the mean of normalized expression rates is one

$$x = k \frac{(\mathrm{ch1i} - \mathrm{ch1b})}{(\mathrm{ch2i} - \mathrm{ch2b})}$$

Normalization by swap

- Consists on eliminating the influence of the dyes properties by using two slides, and swapping the dye used to label the mRNA sample.
- Use it if you find the single slide normalization hypotheses too strong.

Normalization by swap

- Better results can be achieved by doing swap experiments.

$$x = k\frac{(\mathrm{ch1i_A} - \mathrm{ch1b_A})}{(\mathrm{ch2i_A} - \mathrm{ch2b_A})} = \frac{(\mathrm{ch2i_B} - \mathrm{ch2b_B})}{(\mathrm{ch1i_B} - \mathrm{ch1b_B})} \cdot \frac{1}{k}$$

Normalization by swap

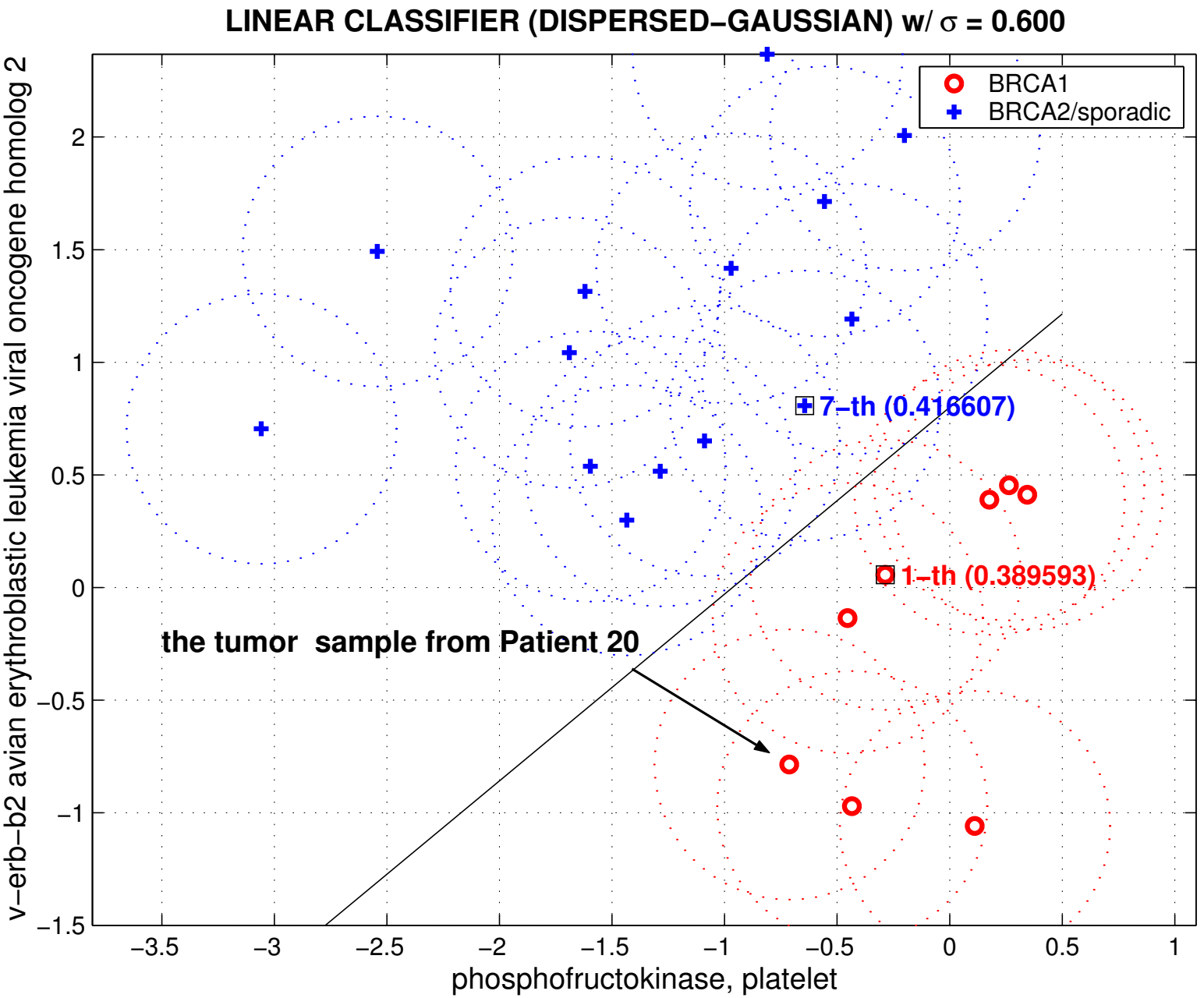
- Better results can be achieved by doing swap experiments.

$$x = \sqrt{\frac{(\mathrm{ch1i_A} - \mathrm{ch1b_A})}{(\mathrm{ch2i_A} - \mathrm{ch2b_A})} \cdot \frac{(\mathrm{ch2i_B} - \mathrm{ch2b_B})}{(\mathrm{ch1i_B} - \mathrm{ch1b_B})}}$$

Genes Signature

Cancer tissue characterization

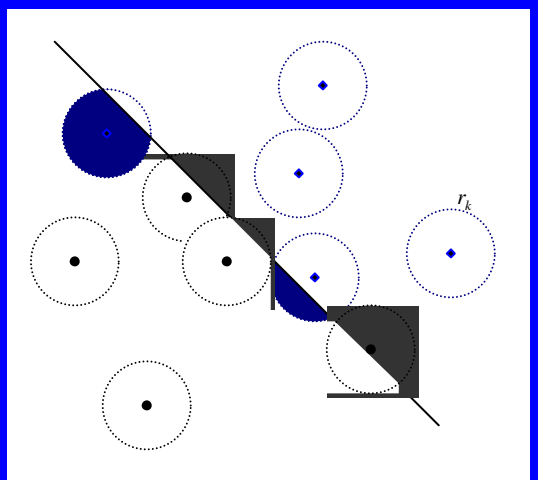
- Problem: from a small set (60) of microarrays, find a minimum number of genes that are enough to separate cancer A and B.

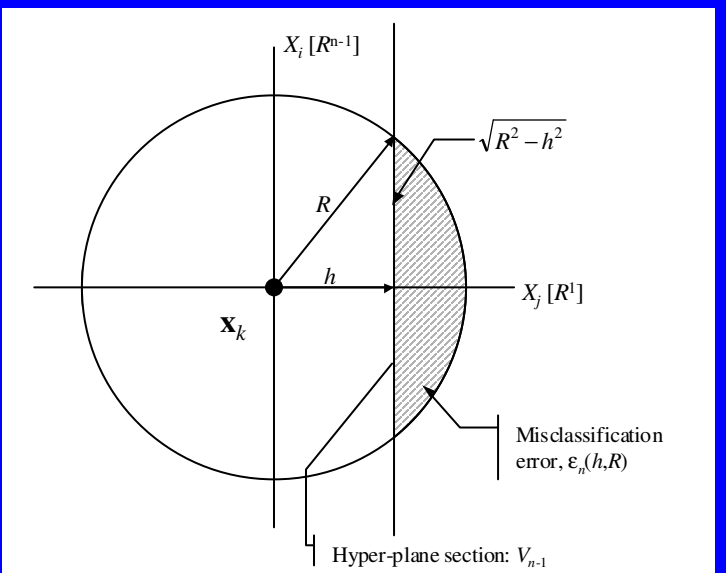

LINEAR CLASSIFIER (DISPERSED−GAUSSIAN) w/ σ = 0.600
BRCA1
BRCA2/sporadic
7−th (0.416607)
1−th (0.389593)
the tumor sample from Patient 20
v−erb−b2 avian erythroblastic leukemia viral oncogene homolog 2
phosphofructokinase, platelet
2
1.5
1
0.5
0
−0.5
−1
−1.5
−3.5
−3
−2.5
−2
−1.5
−1
−0.5
0
0.5
1

Approach

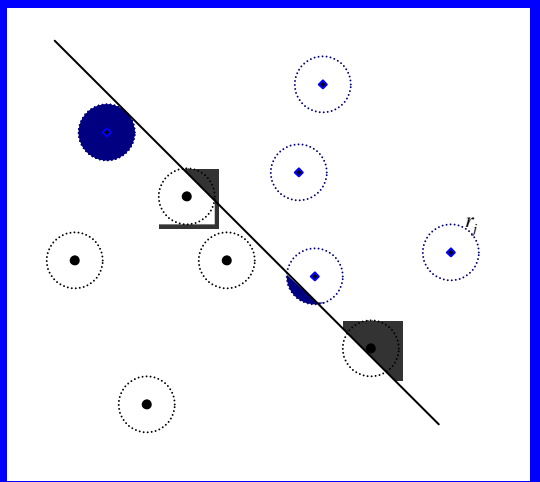
- randomize data
- compute classifier using genes subsets
- measure error for different dispersions
- choose the subset that balance small error and high dispersion.
- A supercomputer is required.

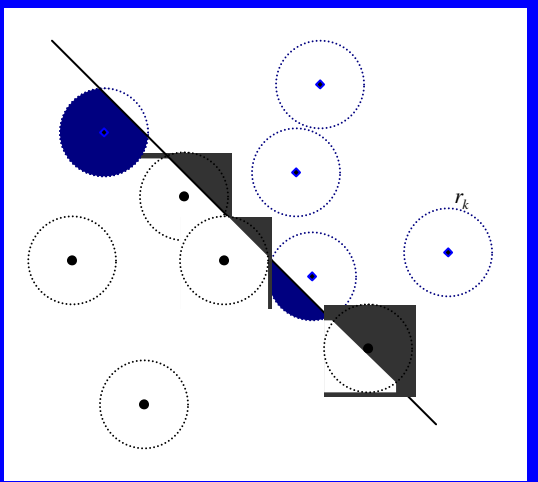
- Linear classifier
- Dispersion centered in the sample
- Flat round dispersion model
- Error computed analytically (faster)

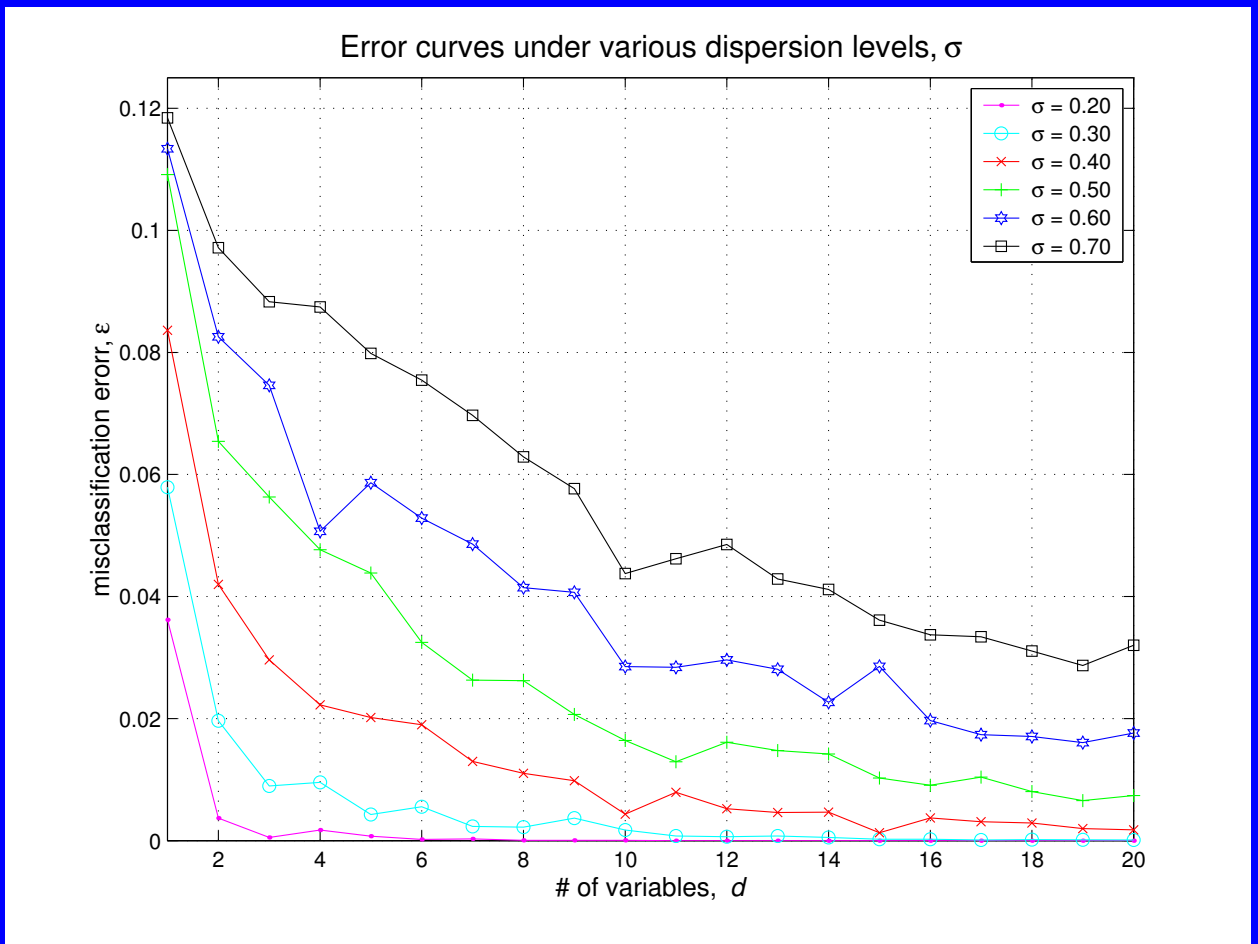




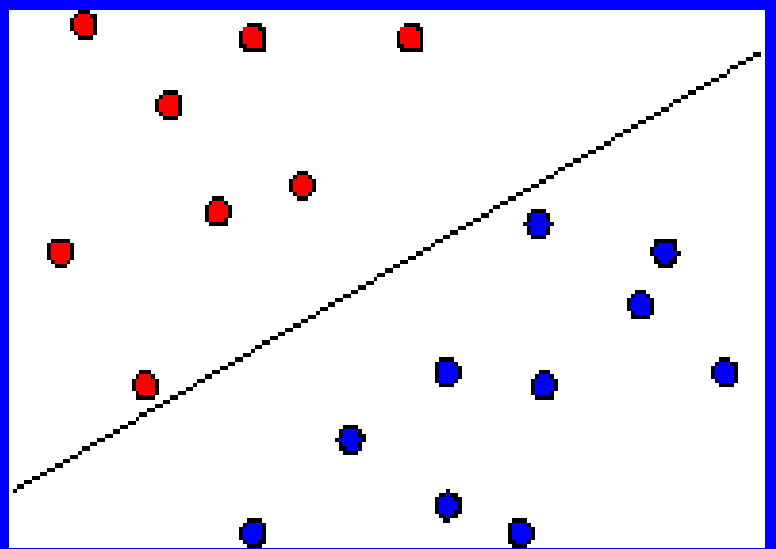
- Robustness analysis

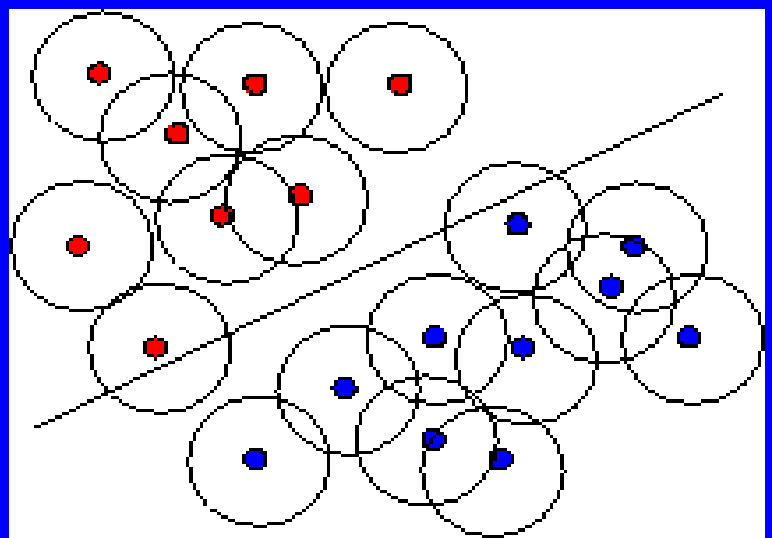


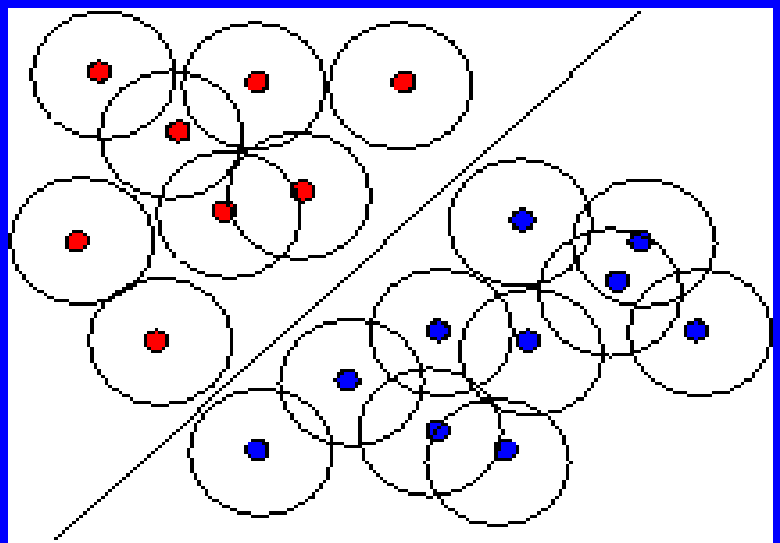



Error curves under various dispersion levels, σ
misclassification erorr, ε
# of variables, d
σ = 0.20
σ = 0.30
σ = 0.40
σ = 0.50
σ = 0.60
σ = 0.70
0
0.02
0.04
0.06
0.08
0.1
0.12
2
4
6
8
10
12
14
16
18
20

Linear programming optimization



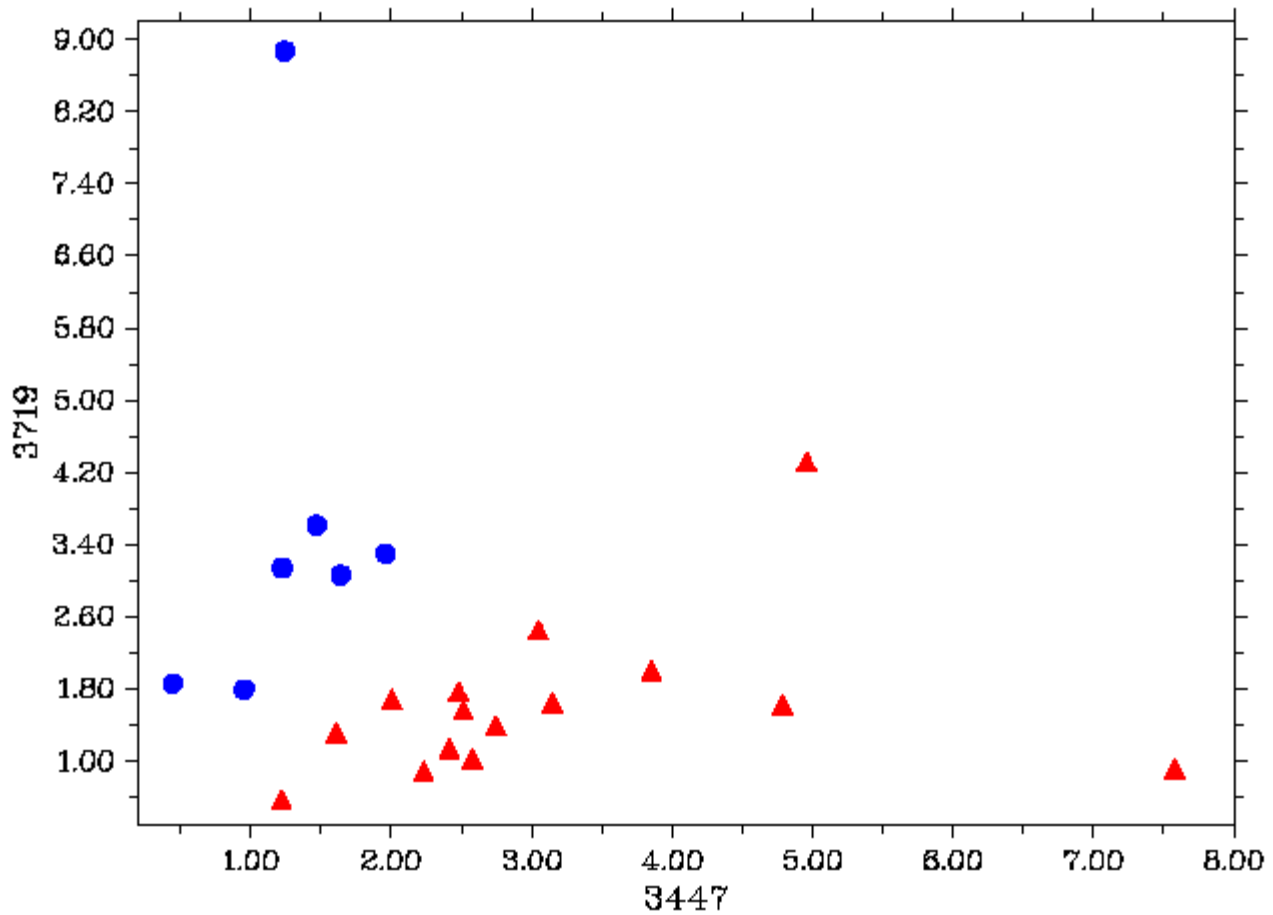




Steps

- The best linear classifier uses about 20-25 genes
- Genes used are eliminated and the best linear classifier is computed, more 20-25 genes are separated
- The procedure is repeated till having about 100 genes
- The full search is applied in the selected subset of genes

Separation

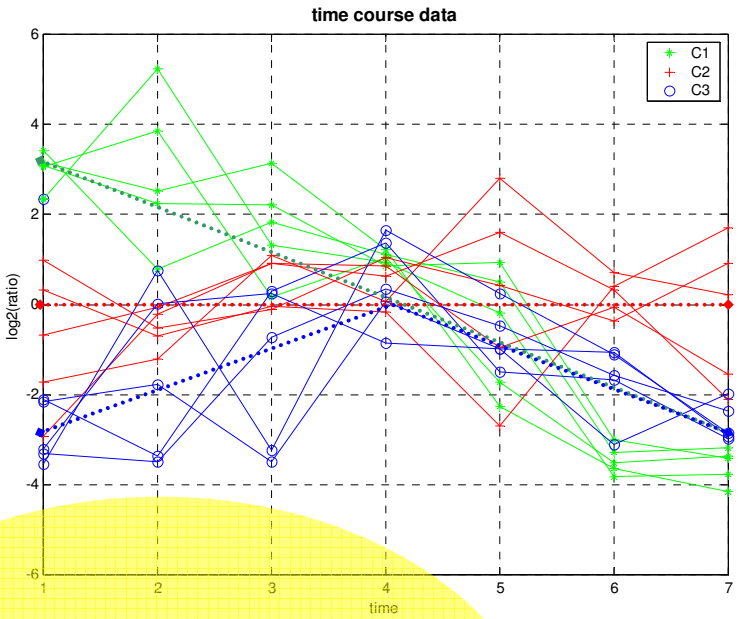

9.00
8.20
7.40
6.60
5.80
5.00
4.20
3.40
2.60
1.80
1.00
3719
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
3447

Validation

- Expression of chosen subsets of genes are measured several times in low cost experiments
- If the experiments reveal compact clusters the subset of genes chosen should be correct.

Clustering

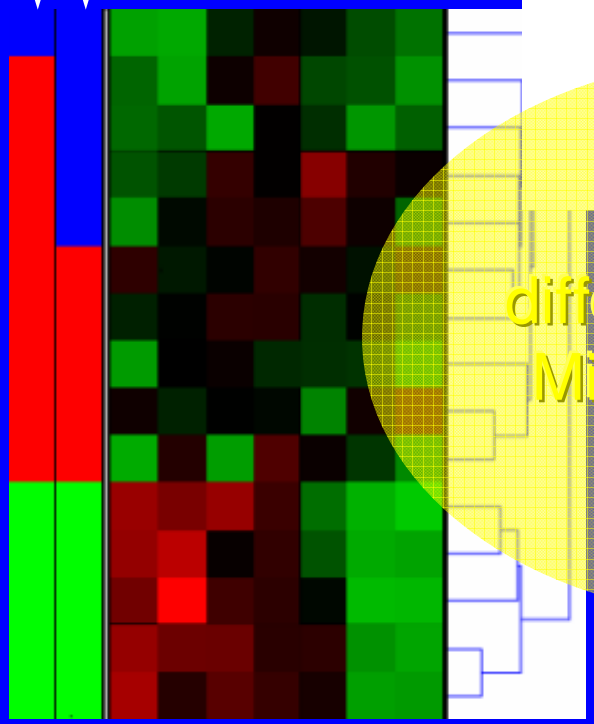
Time course data



Clustered by dendrogram

Original clusters

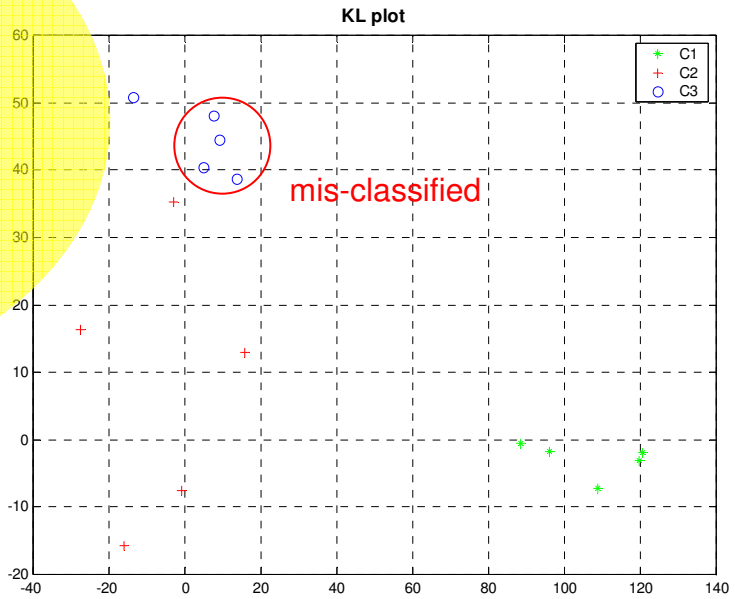
Dendrogram



different views of Microarray data

KL plot

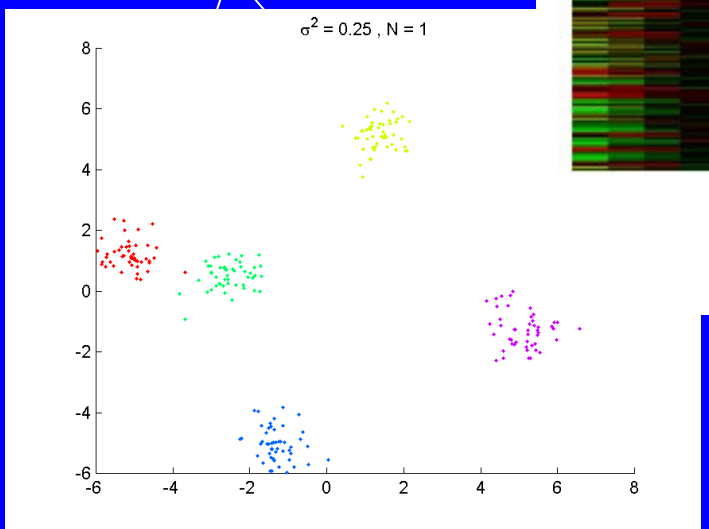
multidimensional space

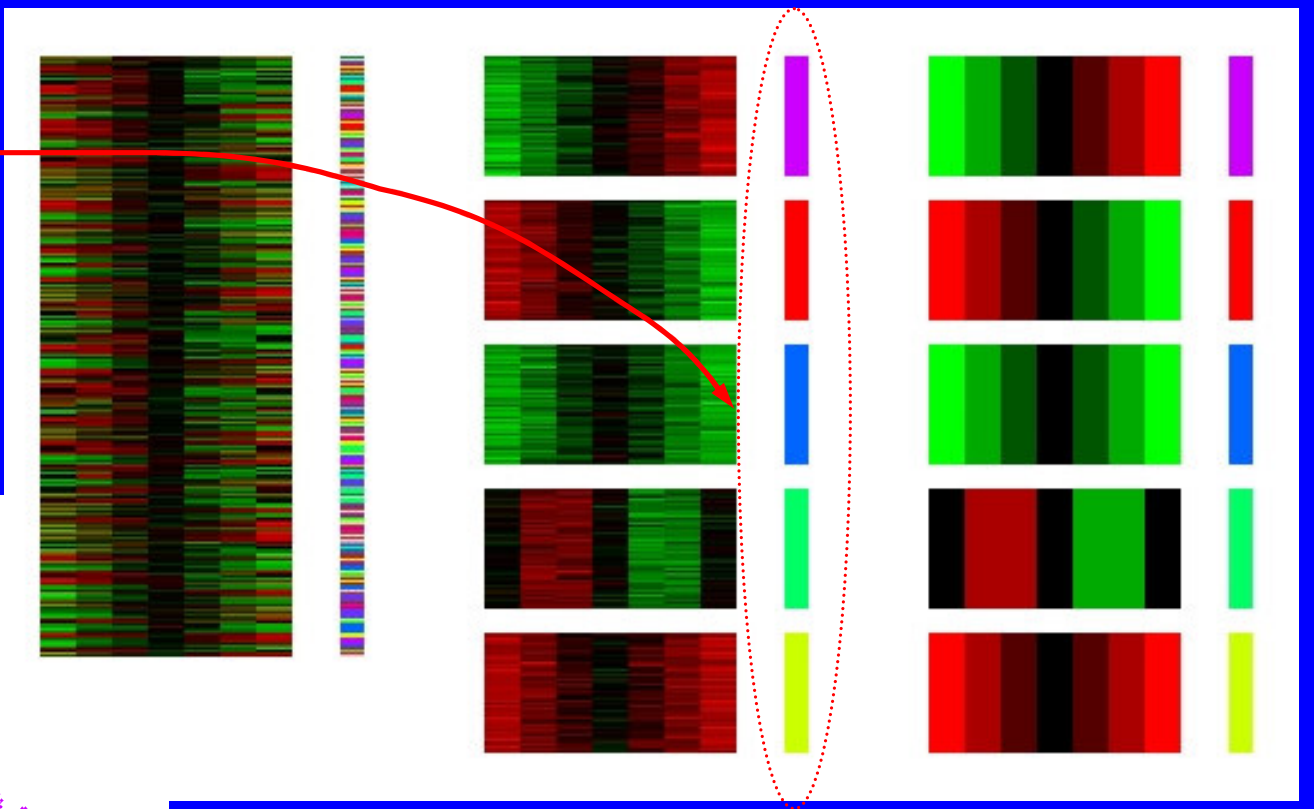


Example

No error!

Tighter clusters due to small variance

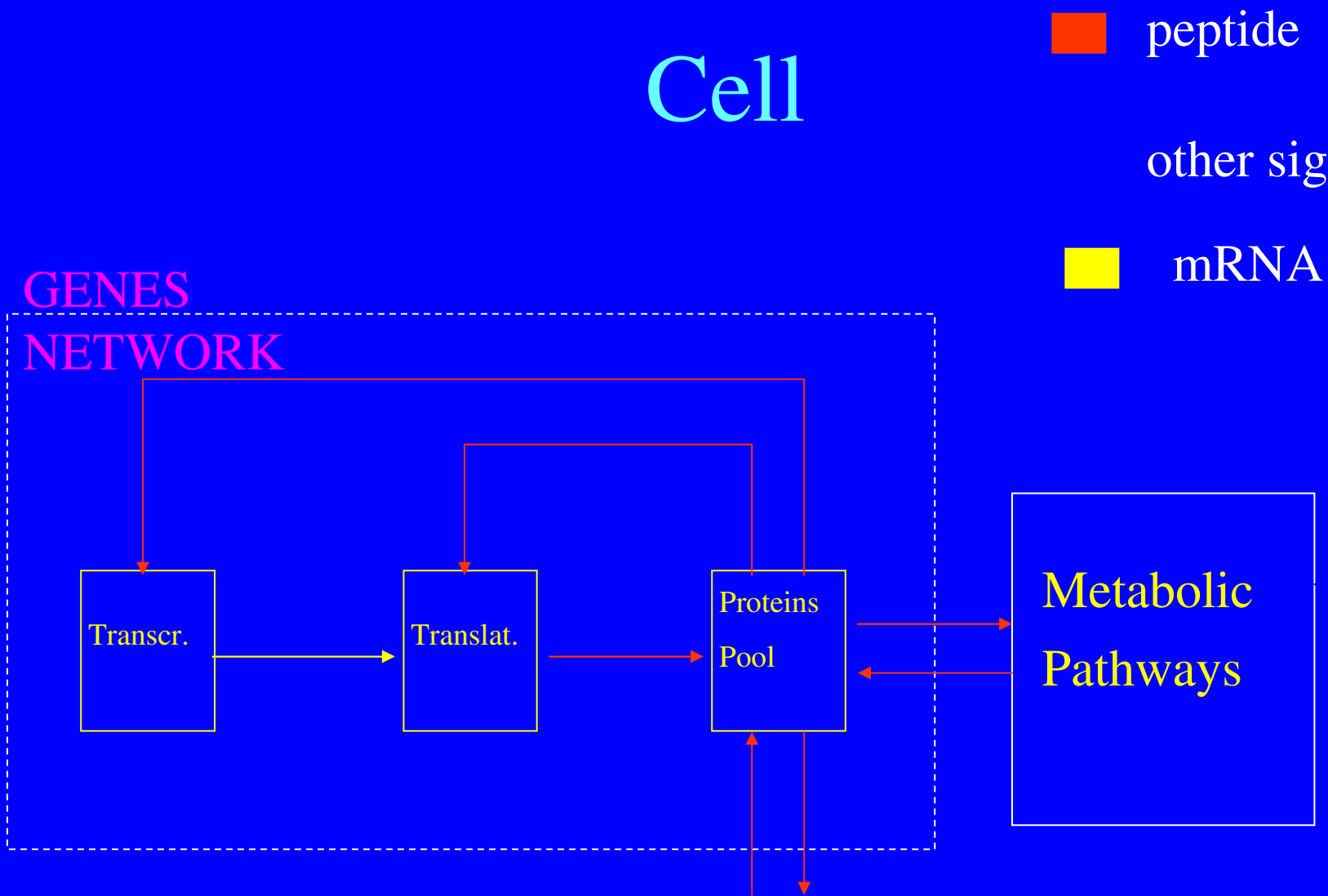




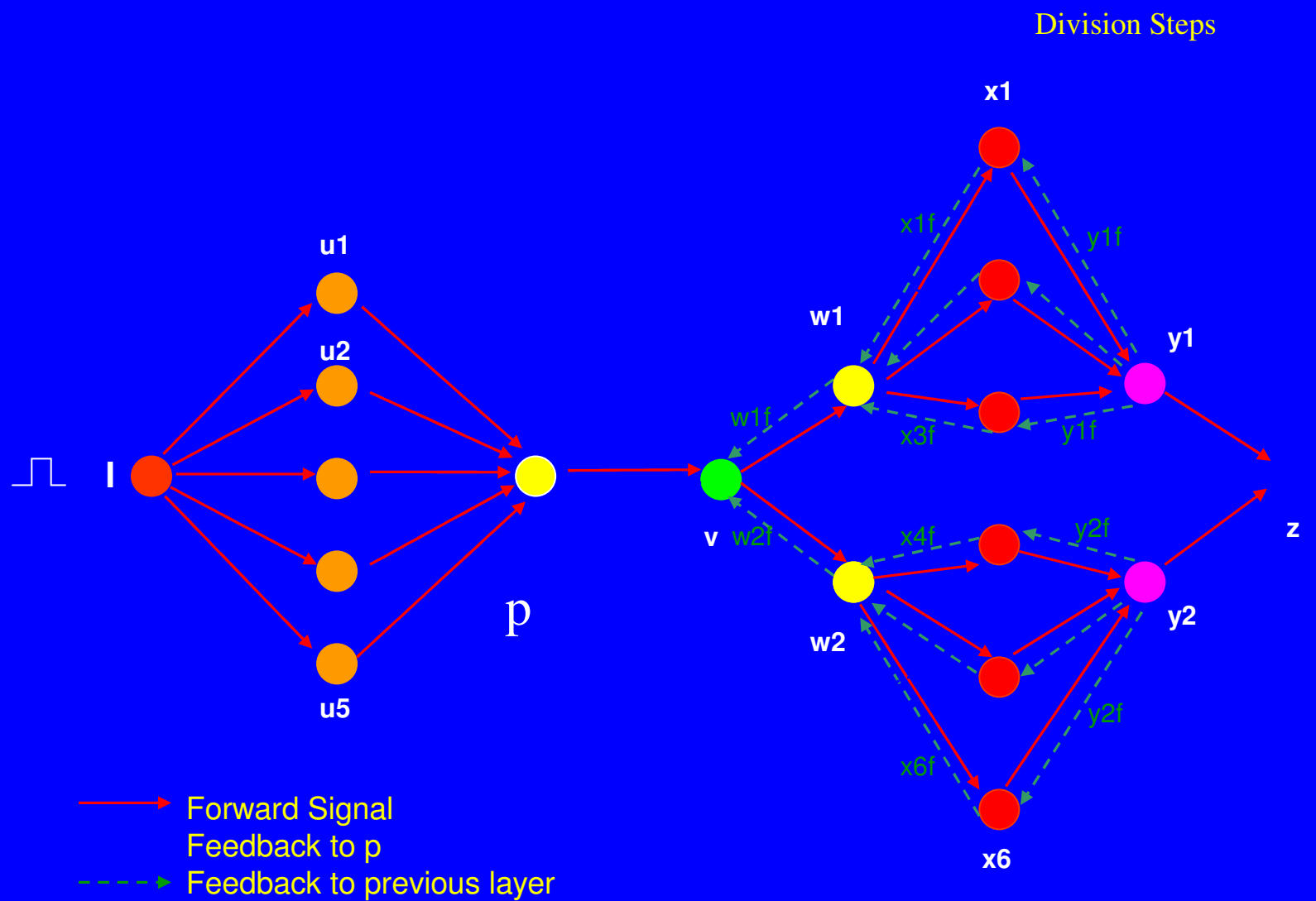
Results from Fuzzy c-means

Gene Regulation Networks

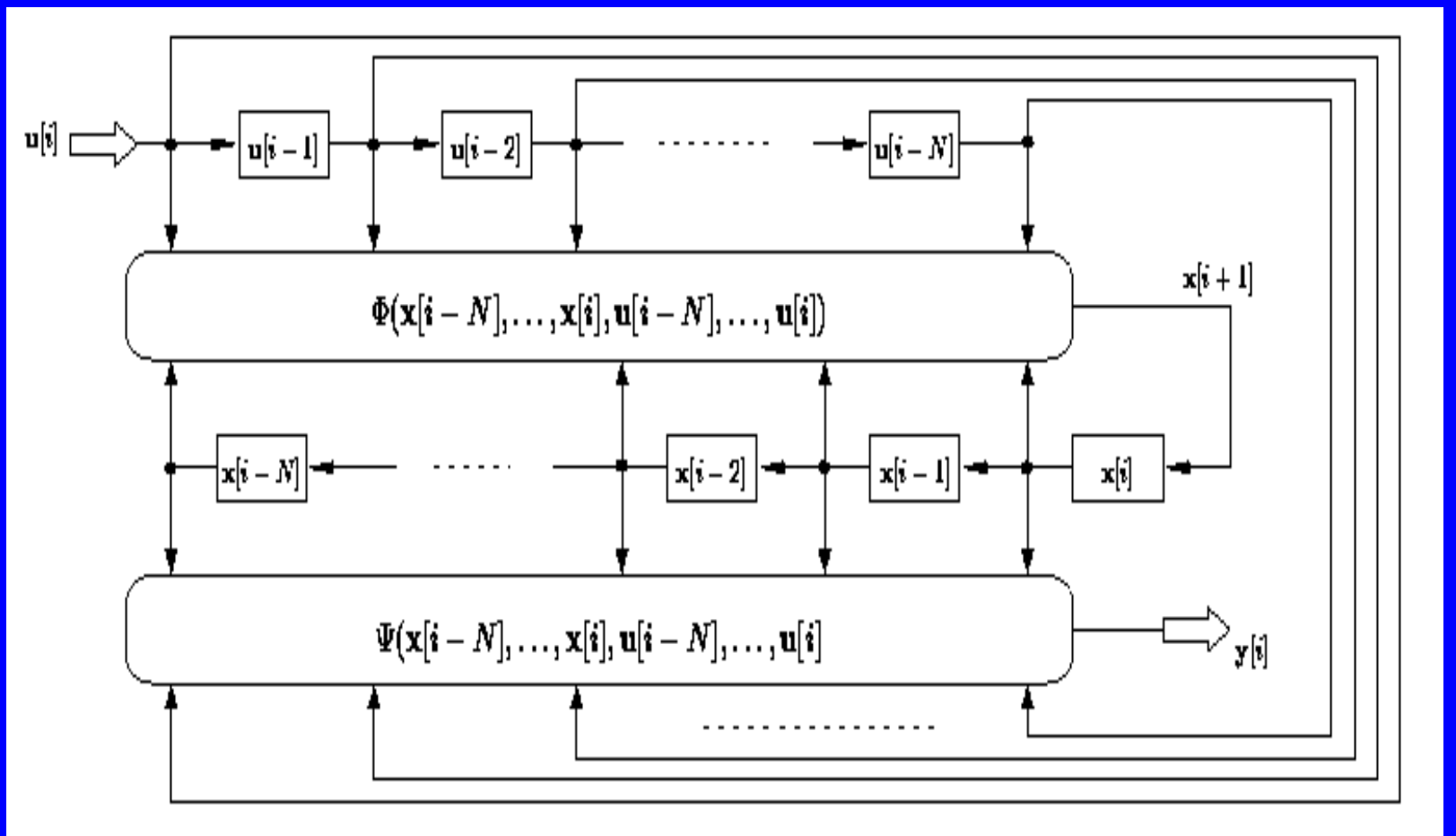
Cell


peptide
other signals
mRNA
GENES
NETWORK
Transcr.
Translat.
Proteins
Pool
Metabolic
Pathways

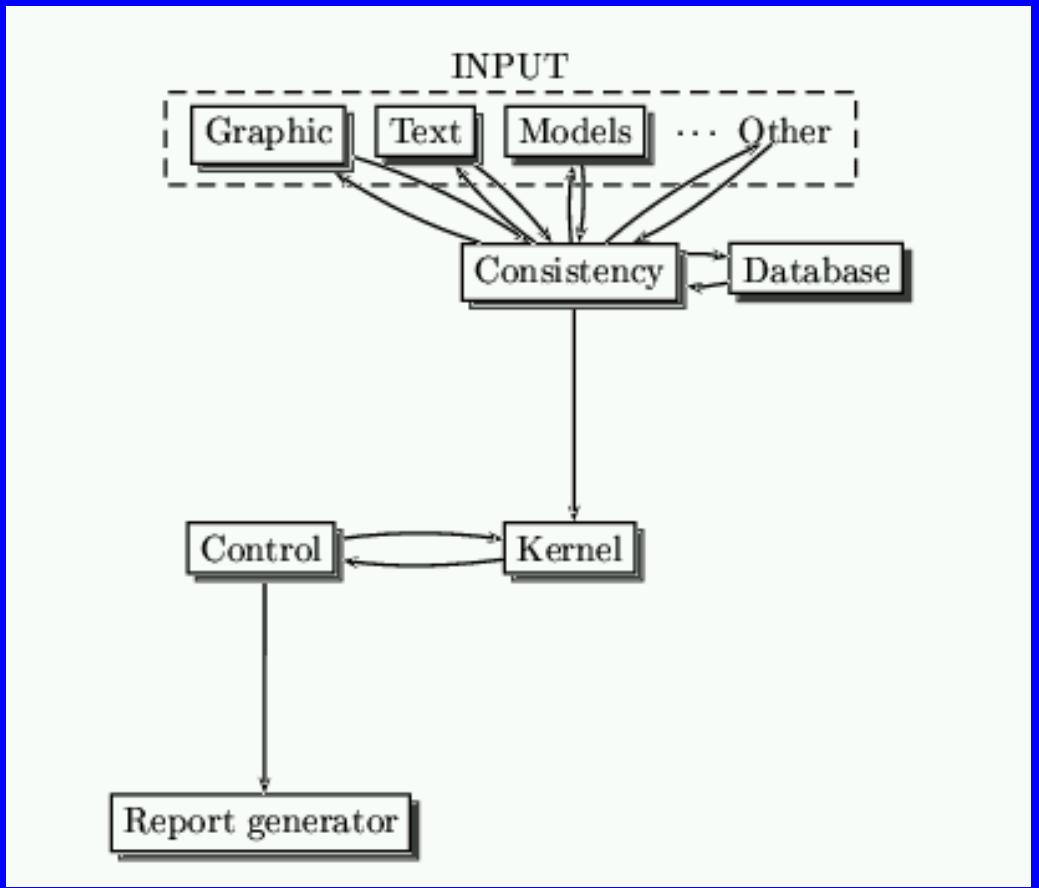
Cell Cycle Modeling

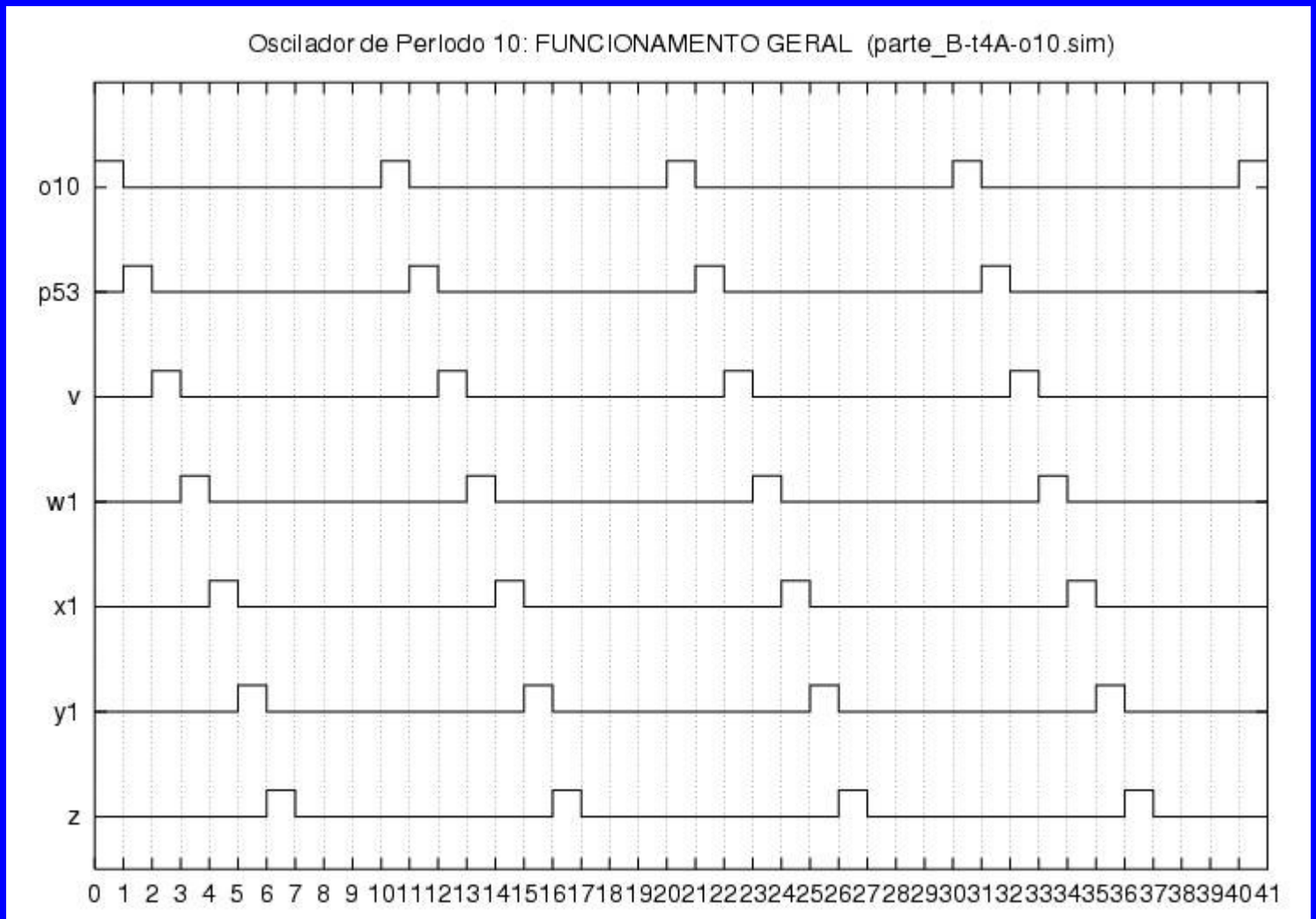


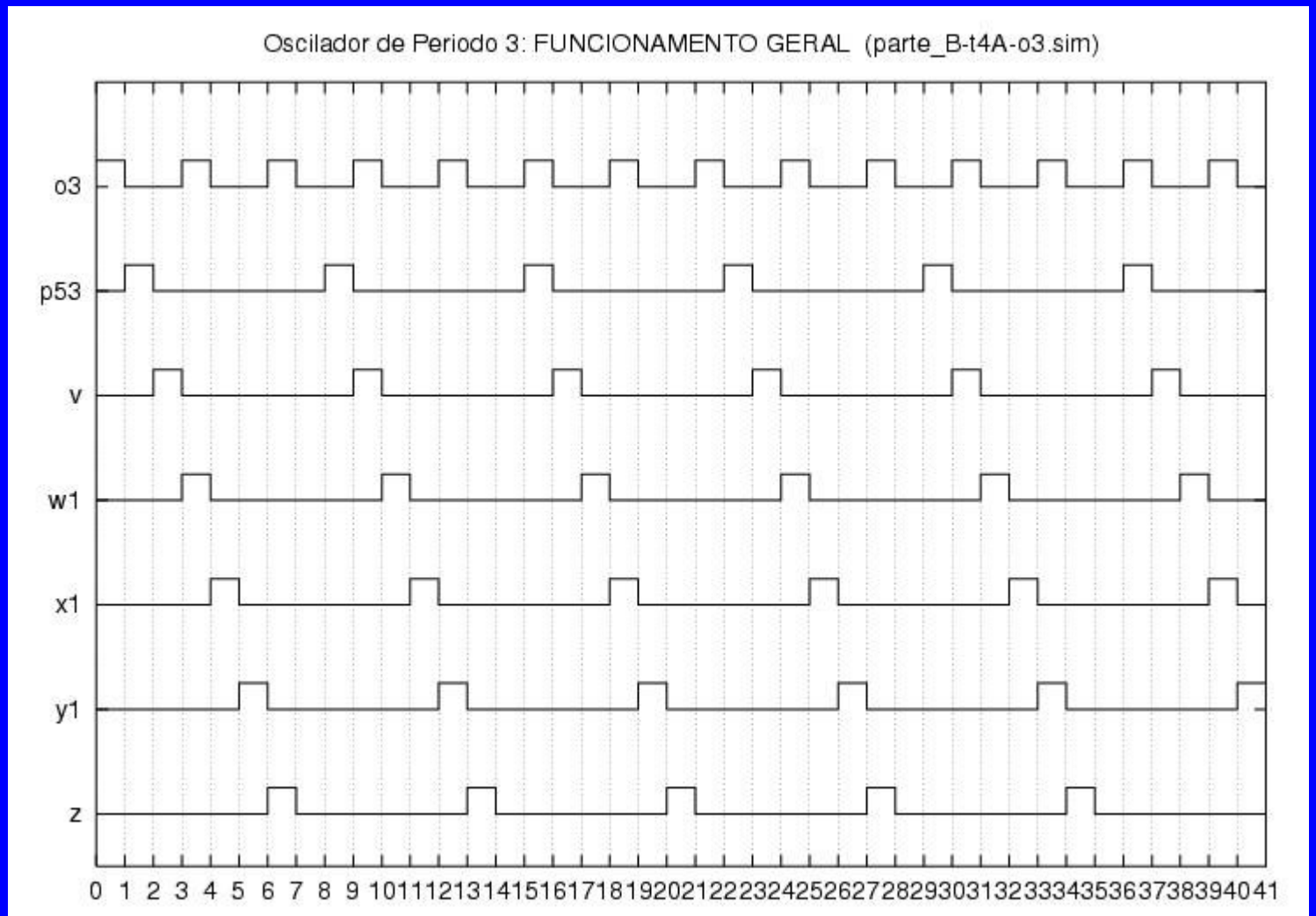
Modeling Dynamical Systems

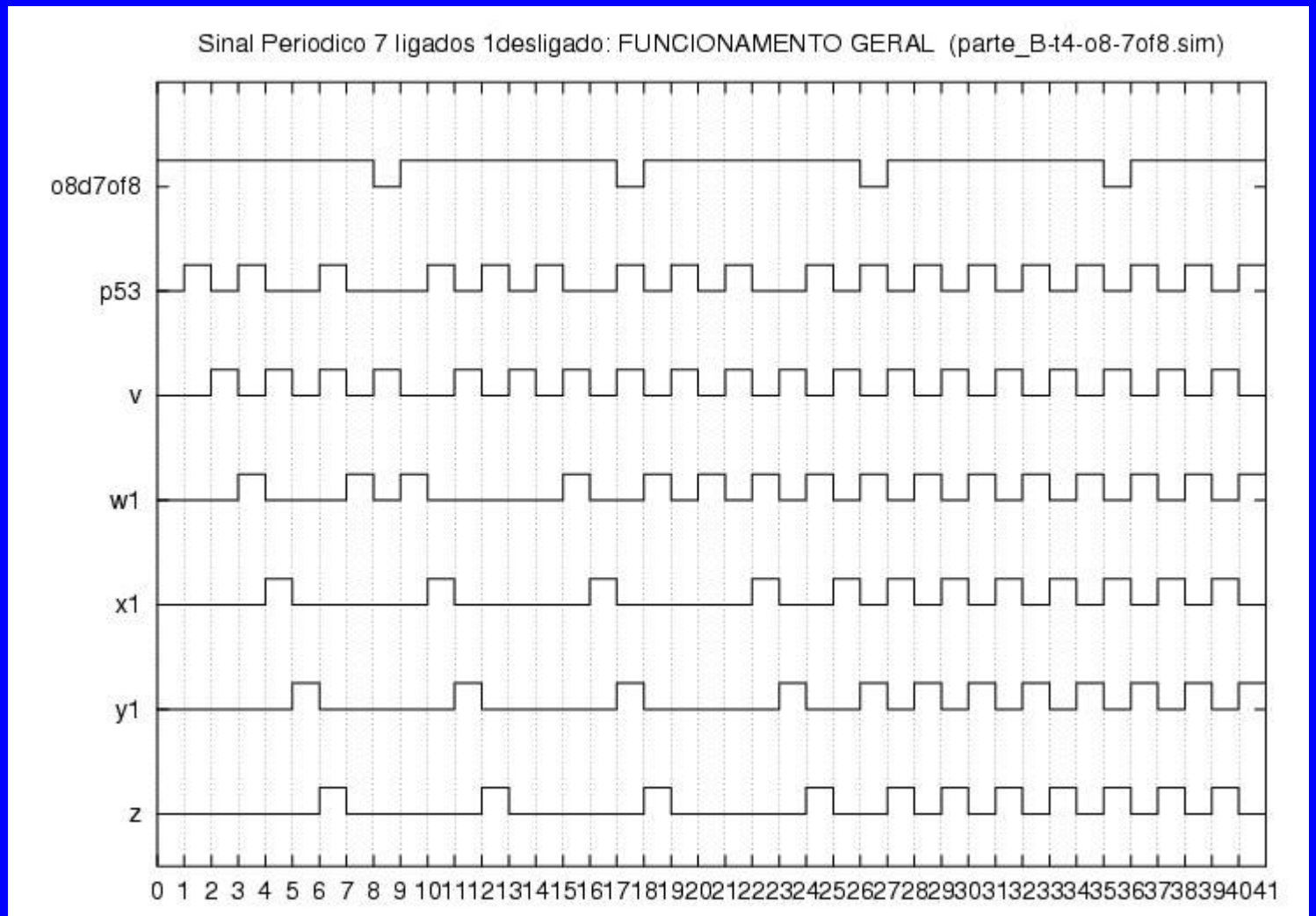


Simulator Architecture

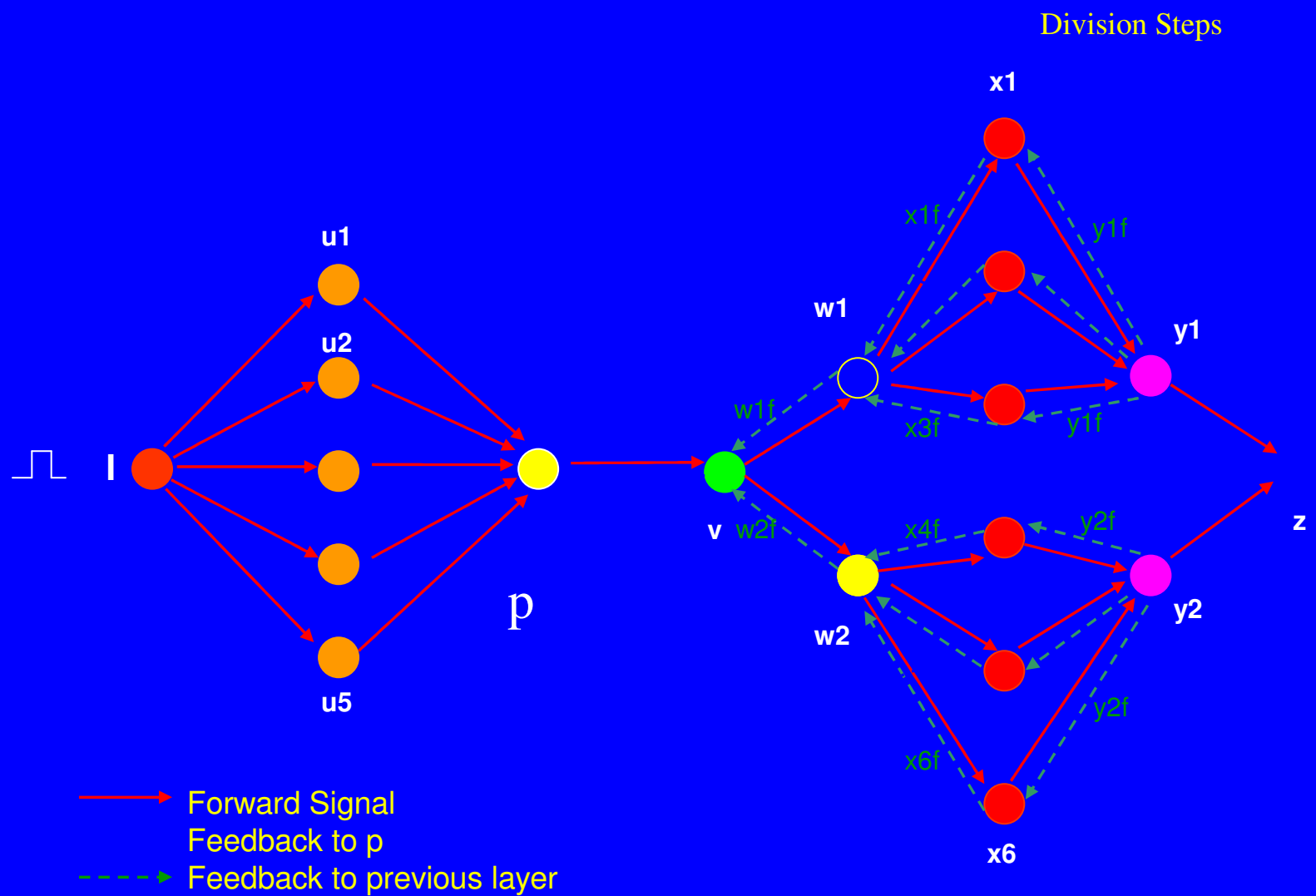


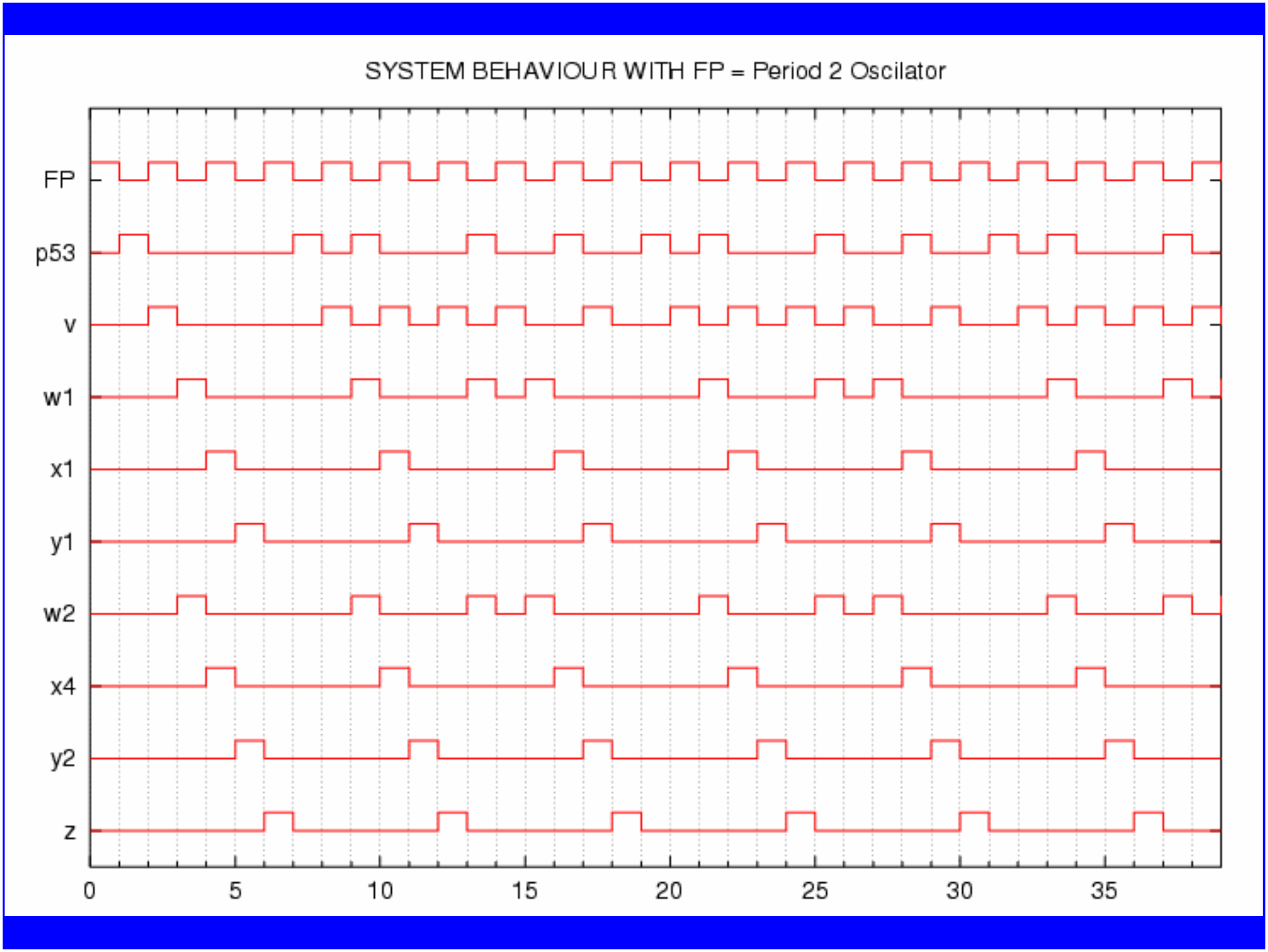

Oscilador de Perlodo 10: FUNCIONAMENTO GERAL (parte_B-t4A-o10.sim)
o10
p53
v
w1
x1
y1
z
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

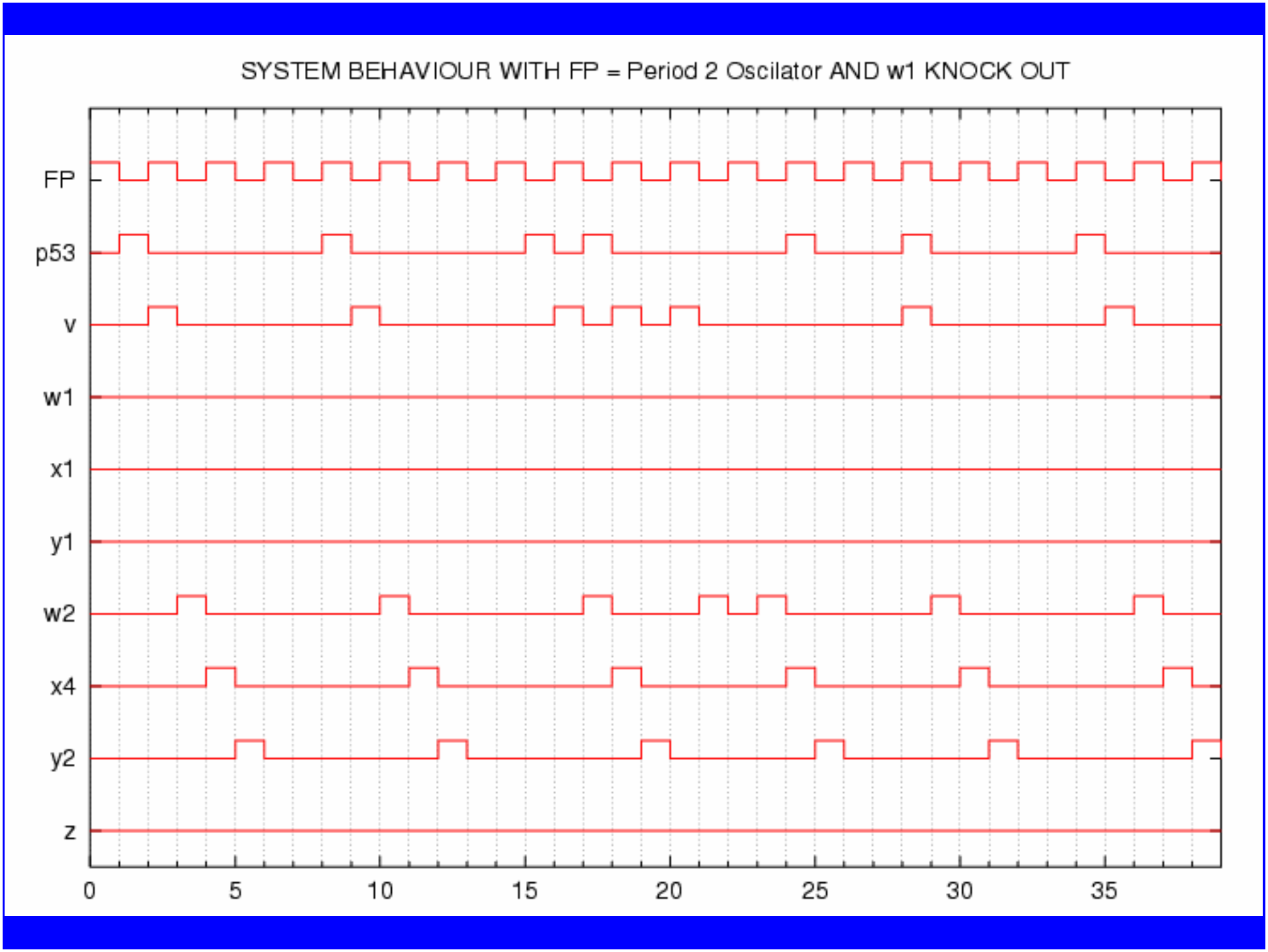

Oscilador de Periodo 3: FUNCIONAMENTO GERAL (parte_B-t4A-o3.sim)
o3
p53
v
w1
x1
y1
z
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41


Sinal Periodico 7 ligados 1desligado: FUNCIONAMENTO GERAL (parte_B-t4-o8-7of8.sim)
o8d7of8
p53
v
w1
x1
y1
z
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

Knockout

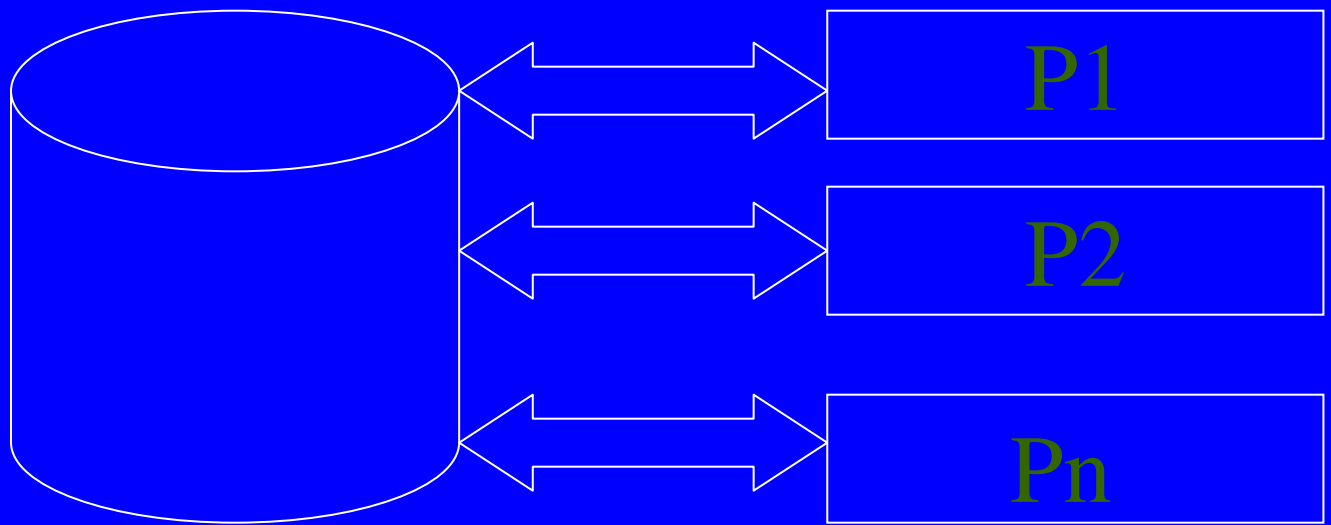

Division Steps
x1
u1
u2
u5
I
p
w1
x1f
y1f
y1
w1f
x3f
y1f
v
w2f
x4f
y2f
z
y2
w2
y2f
x6f
x6
Forward Signal
Feedback to p
Feedback to previous layer


SYSTEM BEHAVIOUR WITH FP = Period 2 Oscilator
FP
p53
v
w1
x1
y1
w2
x4
y2
z
0
5
10
15
20
25
30
35


SYSTEM BEHAVIOUR WITH FP = Period 2 Oscilator AND w1 KNOCK OUT
FP
p53
v
w1
x1
y1
w2
x4
y2
z
0
5
10
15
20
25
30
35

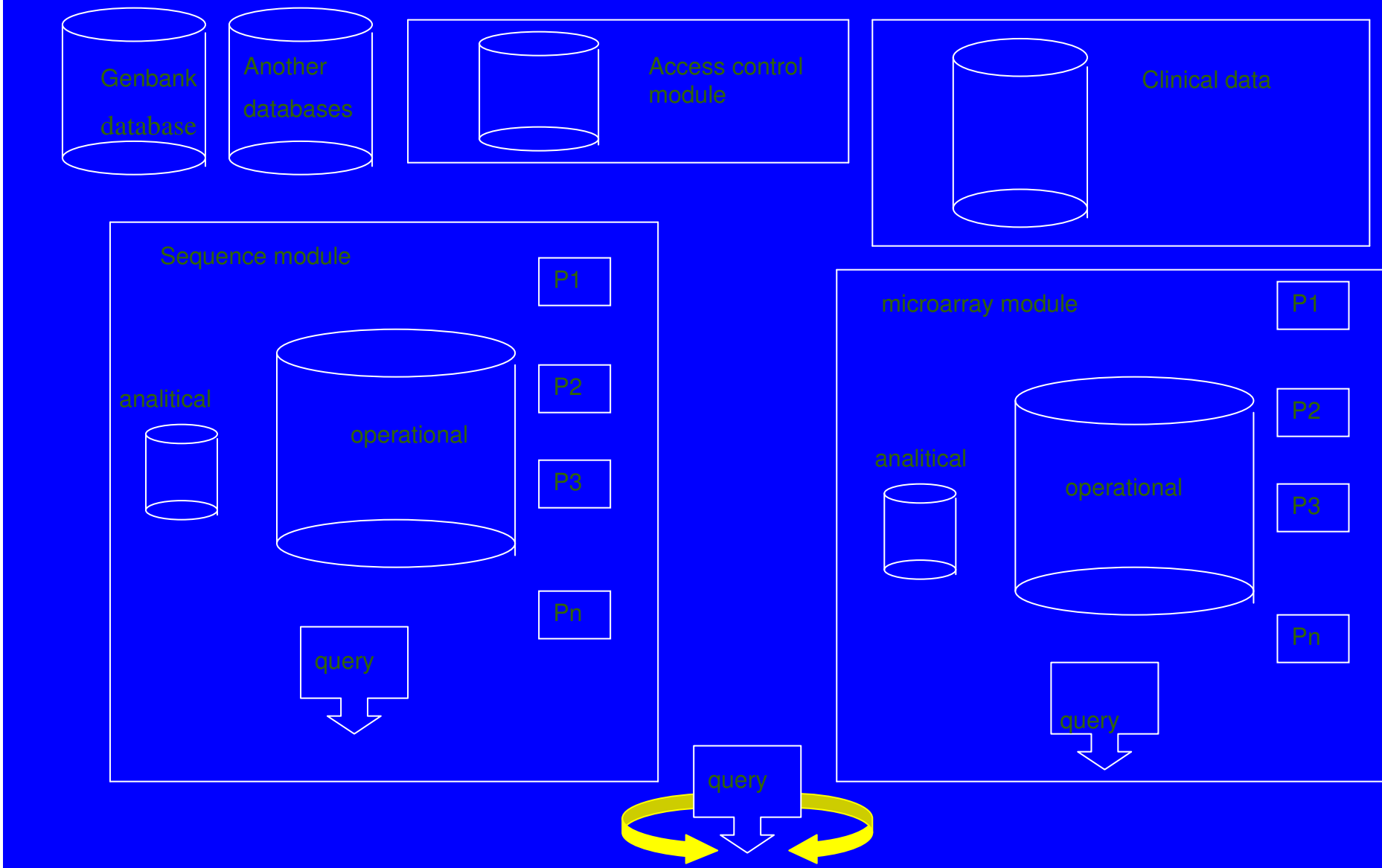
An environment for knowledge discovery

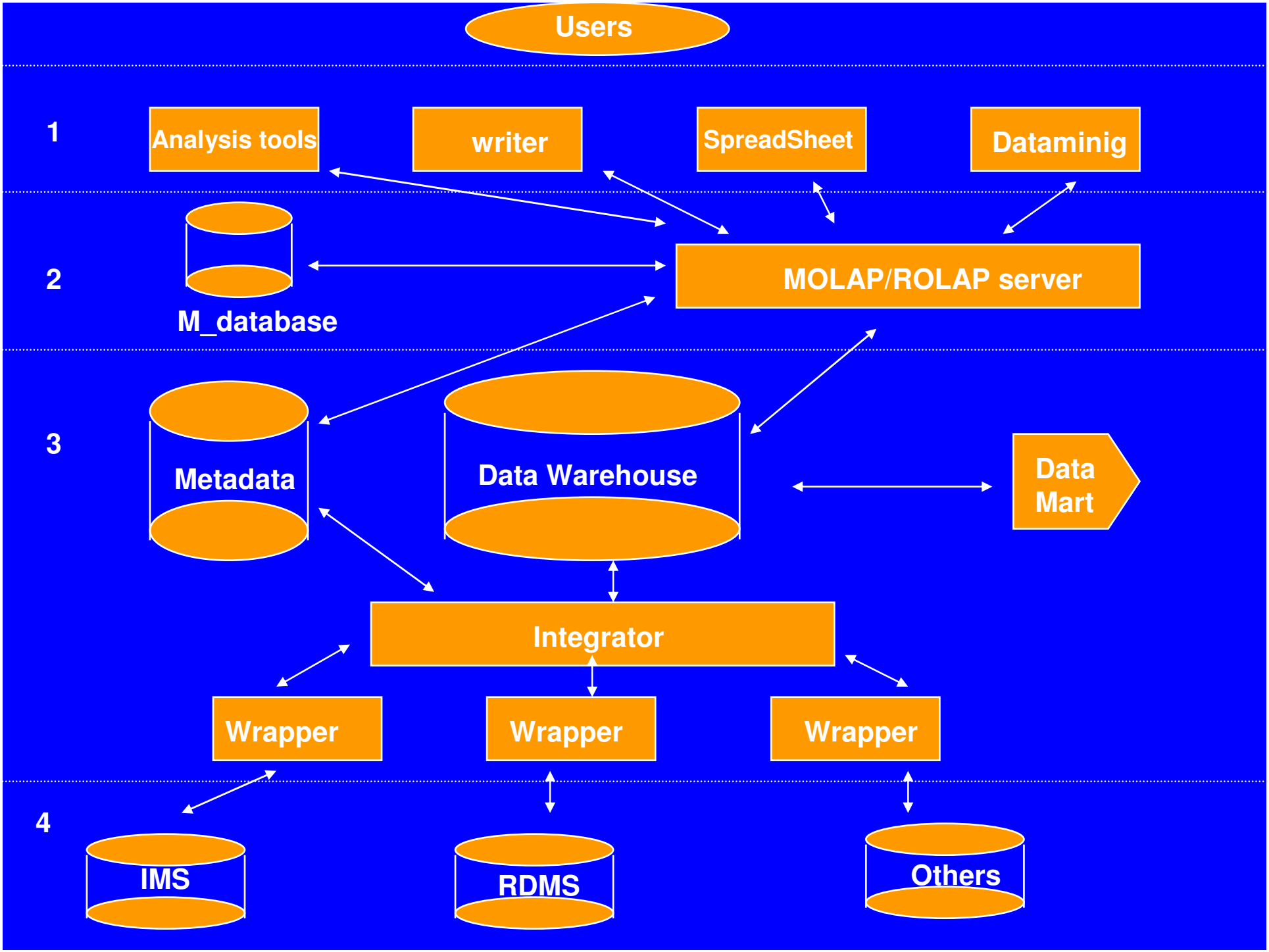
Objected oriented database



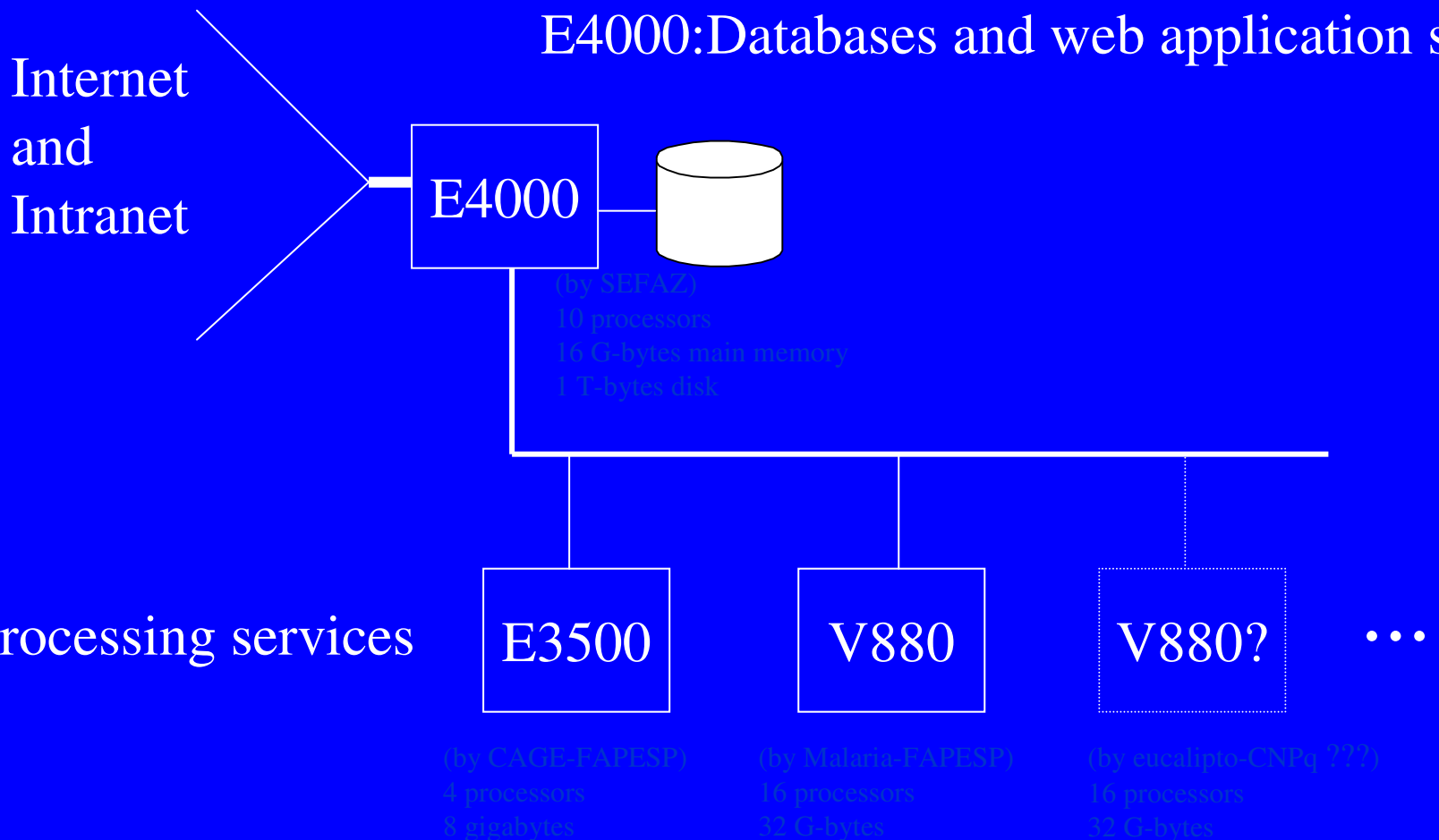
Pi : analytical and mining procedures (kernel parallel)

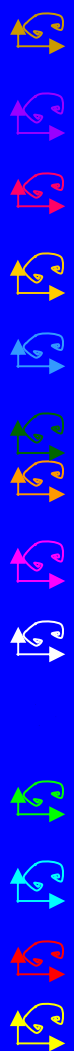
Integrated Environment

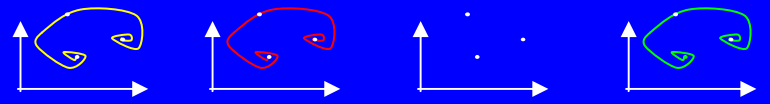


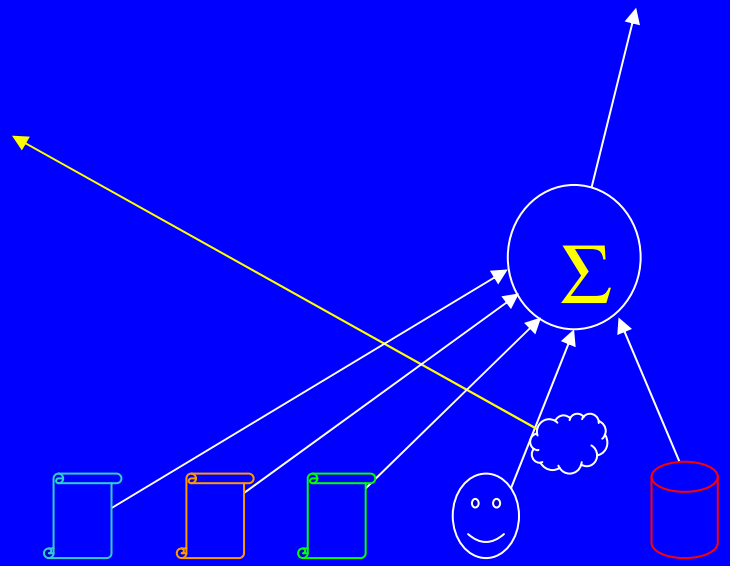

Users
1
Analysis tools
writer
SpreadSheet
Dataminig
2
M_database
MOLAP/ROLAP server
3
Metadata
Data Warehouse
Data Mart
Integrator
Wrapper
Wrapper
Wrapper
4
IMS
RDMS
Others

GRID Computer - DCC-IME-USP









What genes regulate the pathway A->B->C->D ?

Wet Lab

Proteome

Transcriptome

Genome

Pathways