Data mining in genetics

Junior Barrera

BIOINFO-USP

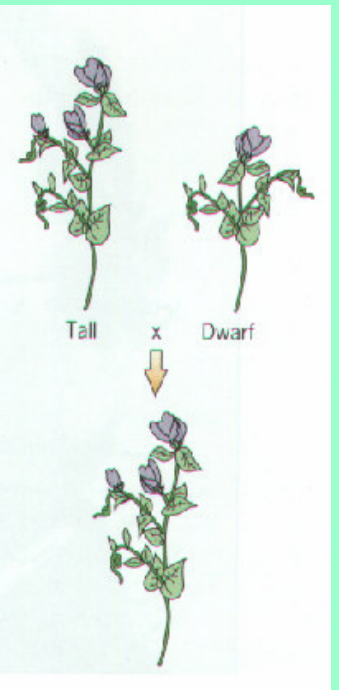
Layout

- Introduction
- Data mining
- Mapping of rare genes
- Expression analysis: measure, genes differentially expressed; clustering expression signals; identification of gene regulation networks

Introduction

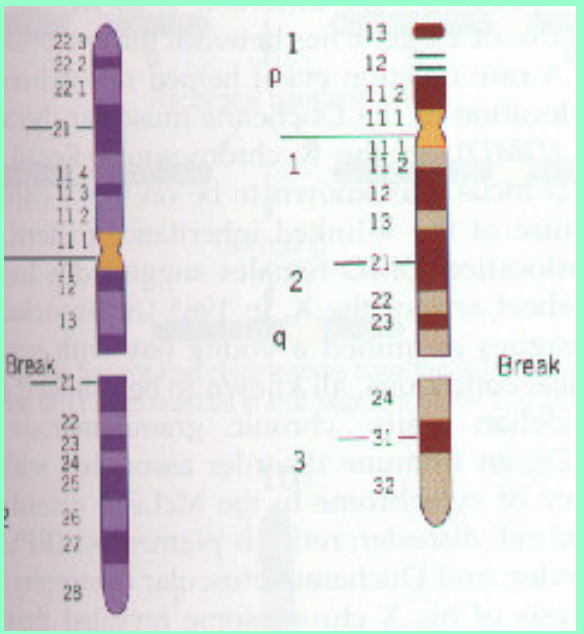
Knowledge evolution in genetics

- Heredity - Mendel (1866)
- The phenotypes of an individual depends on genes of his parents.



Knowledge evolution in genetics

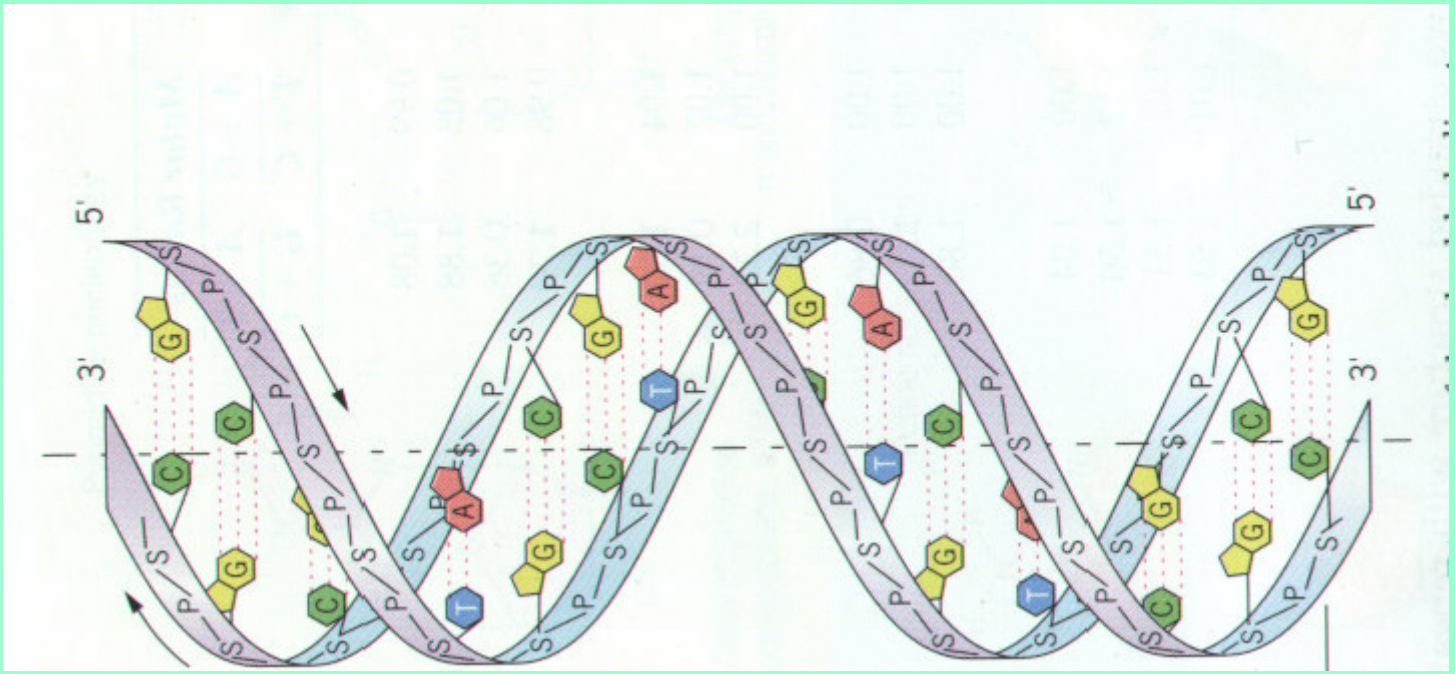
- Chromosome Theory - Morgan (1910)
- Genes were situated in chromosomes



Knowledge evolution in genetics

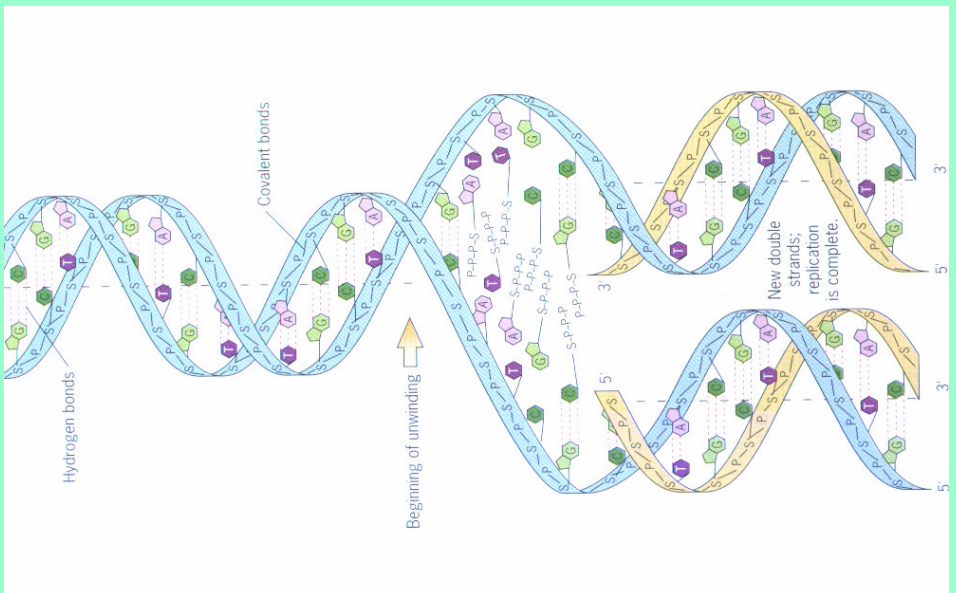
- The molecular structure of chromosomes (Watson and Crick - 1953)
- DNA structure: the double helix
- Four basis: adenine(A), guanine(G), thymine(T), cytosine(C)
- genes are sequences of nucleotides

Knowledge evolution in genetics



Knowledge evolution in genetics

- DNA manipulation
- cut, replication and decoding



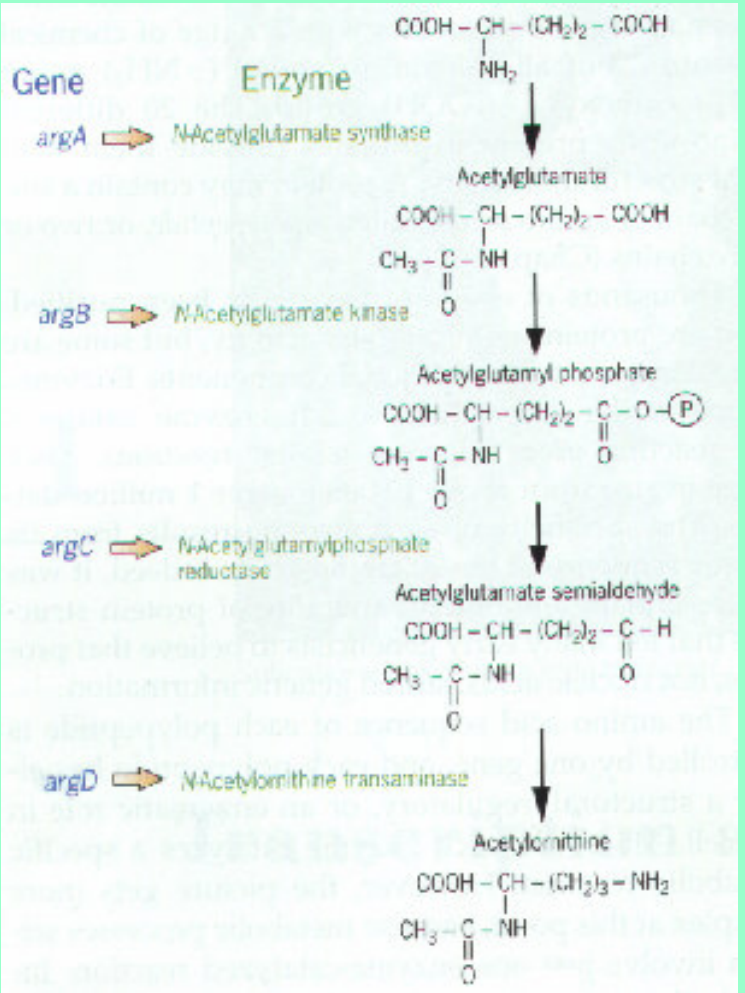
Knowledge evolution in genetics

- Genetic engineering
- species modification, drug production

Knowledge evolution in genetics

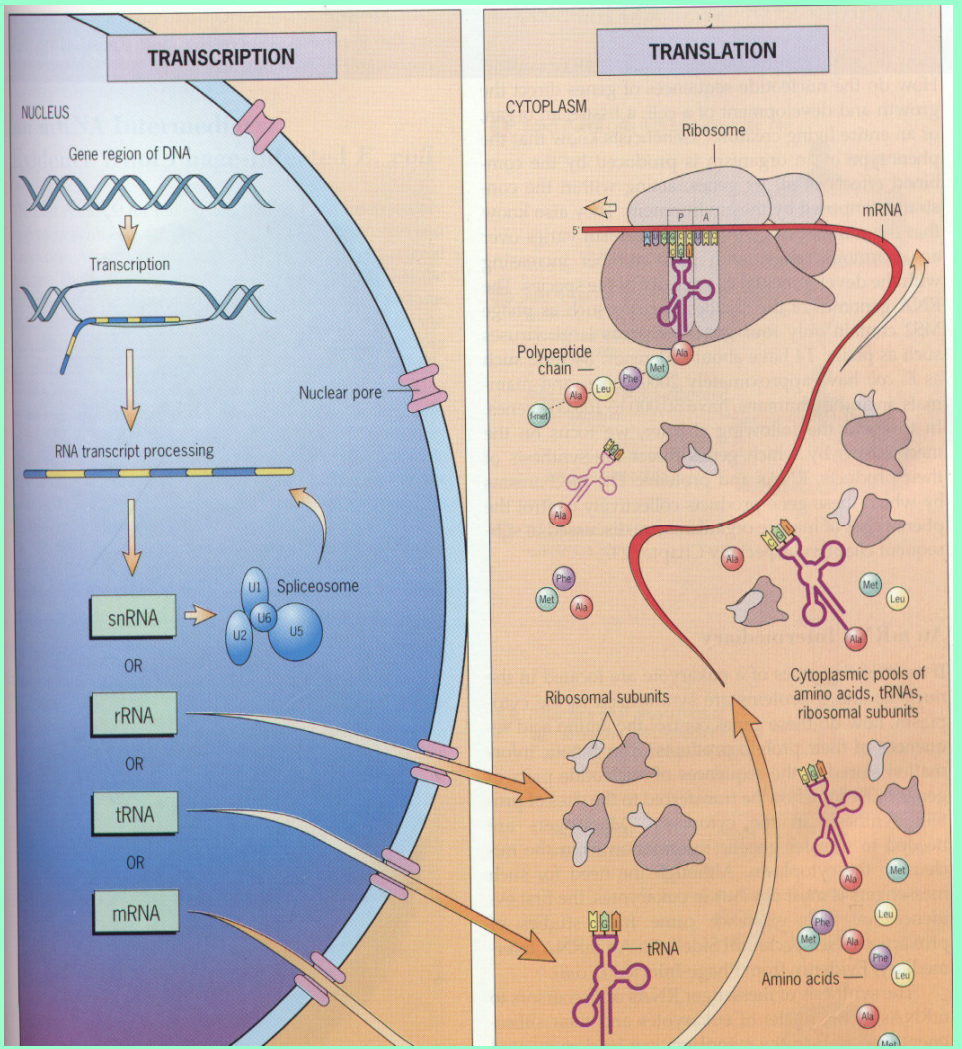
- Genes control the metabolism
- Metabolism occurs by sequences of enzyme-catalyzed reactions.
- Enzymes are specified by one or more genes

Knowledge evolution in genetics

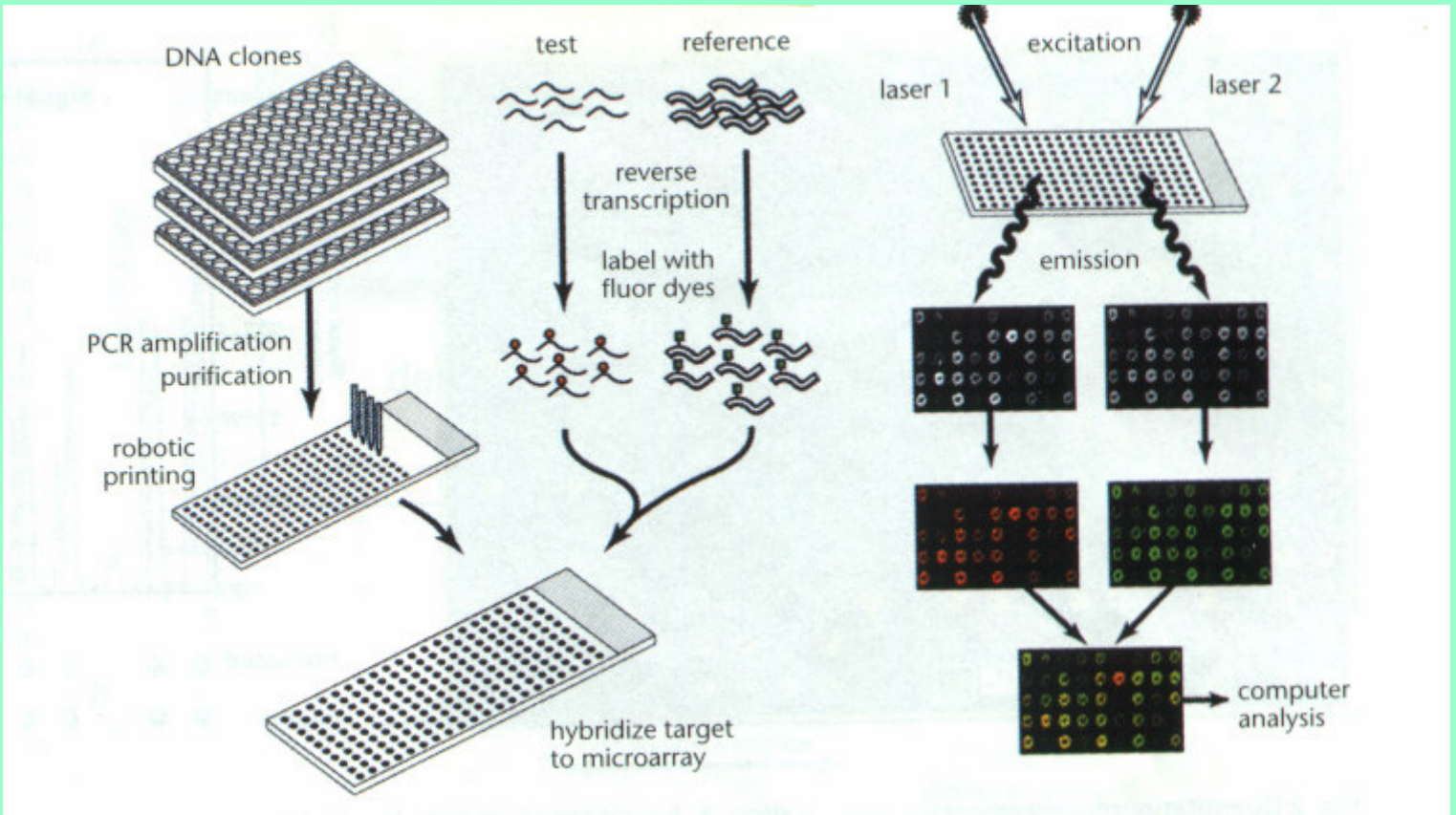


Knowledge evolution in genetics

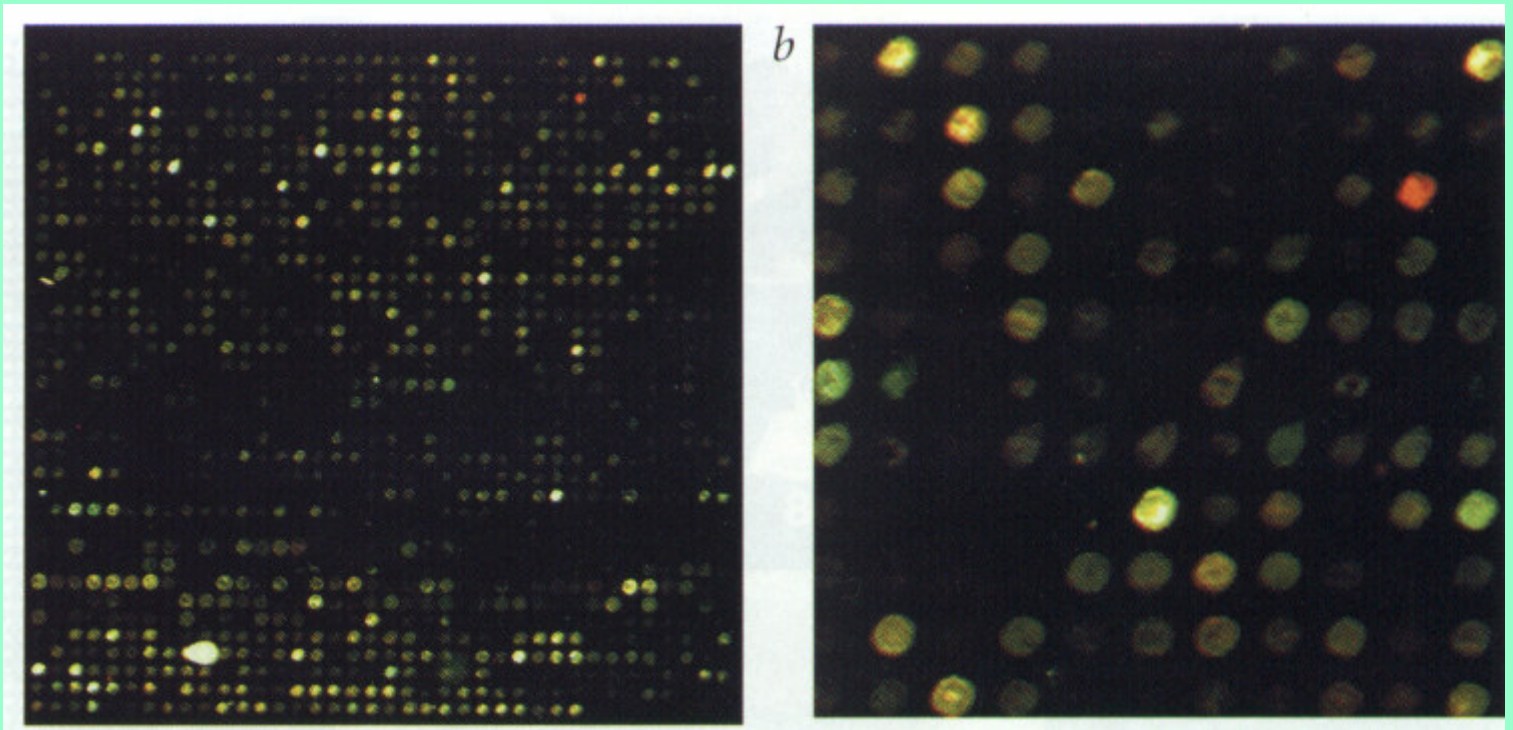
- Gene expression



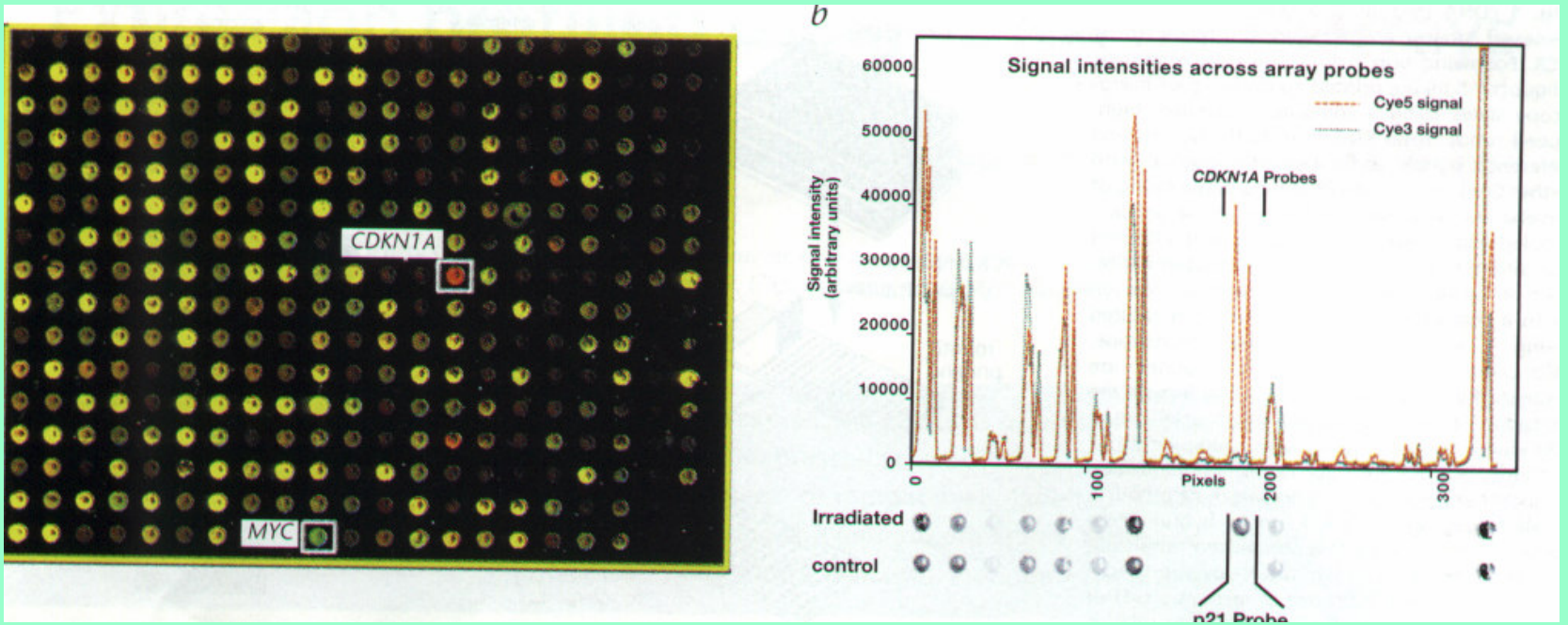
Data acquisition



Data acquisition



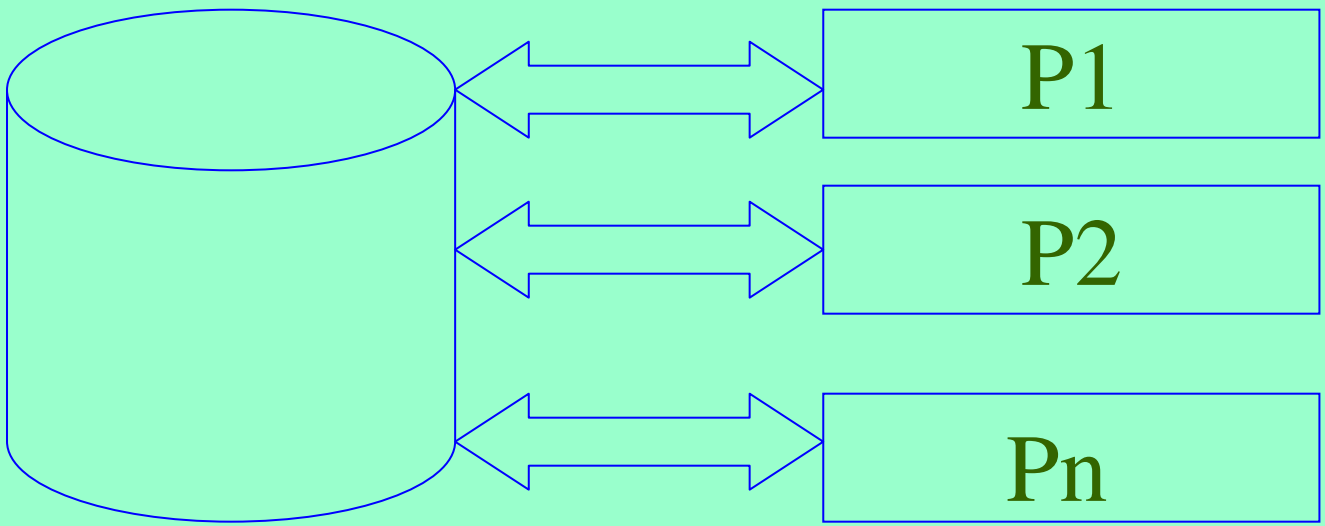
Data acquisition



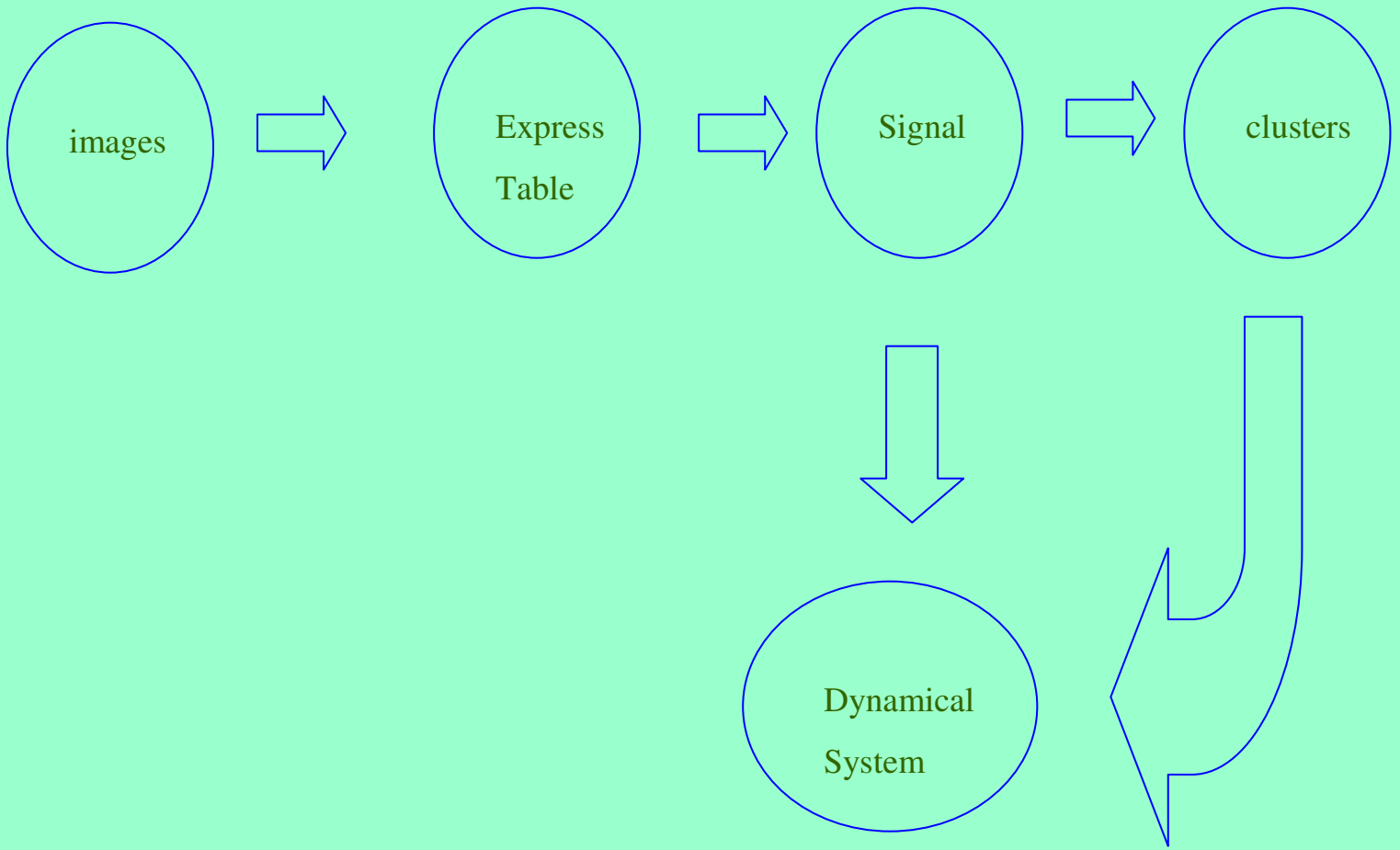
Quantization - {-1,0,1}

Data mining

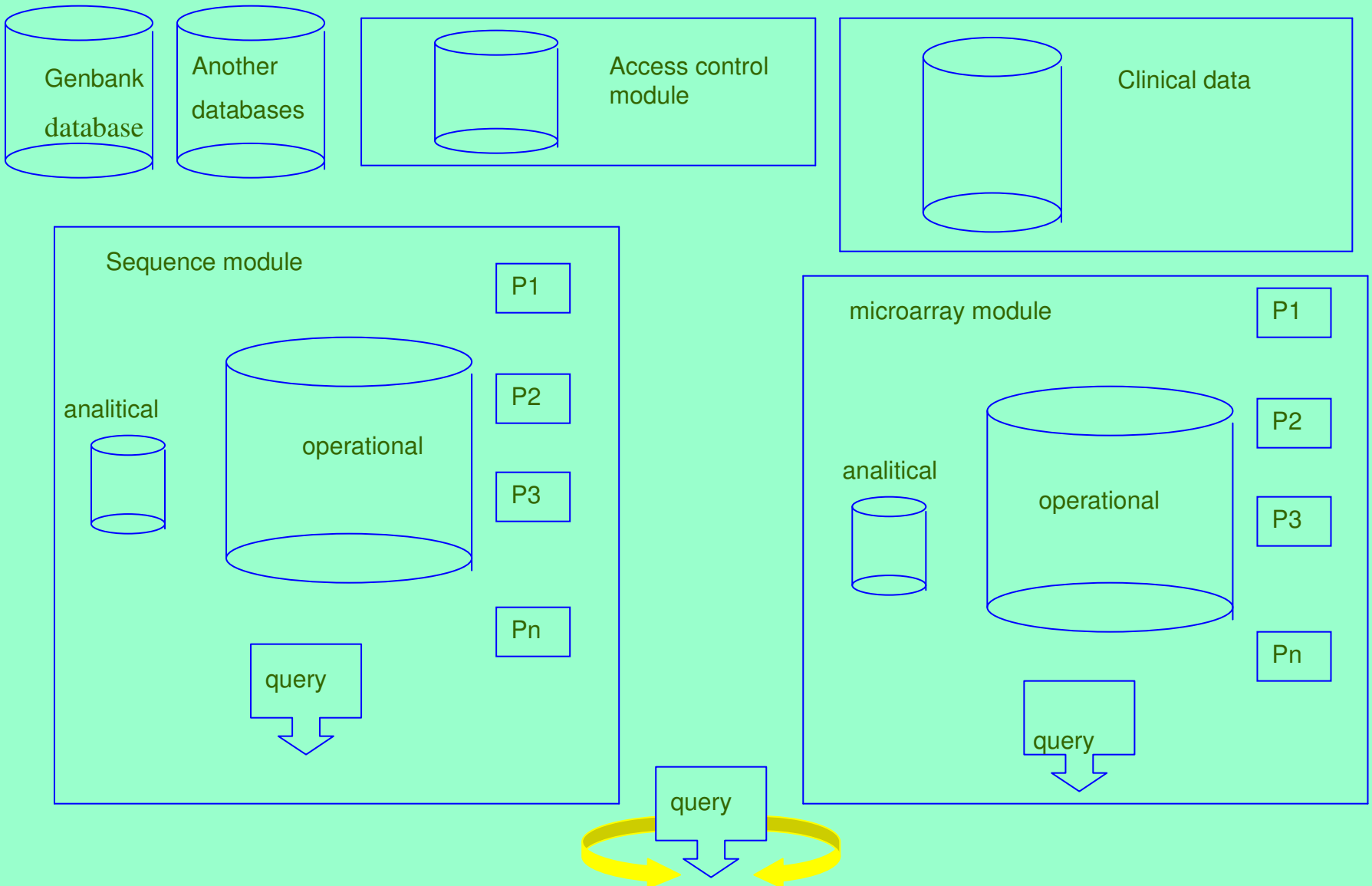
Objected oriented database

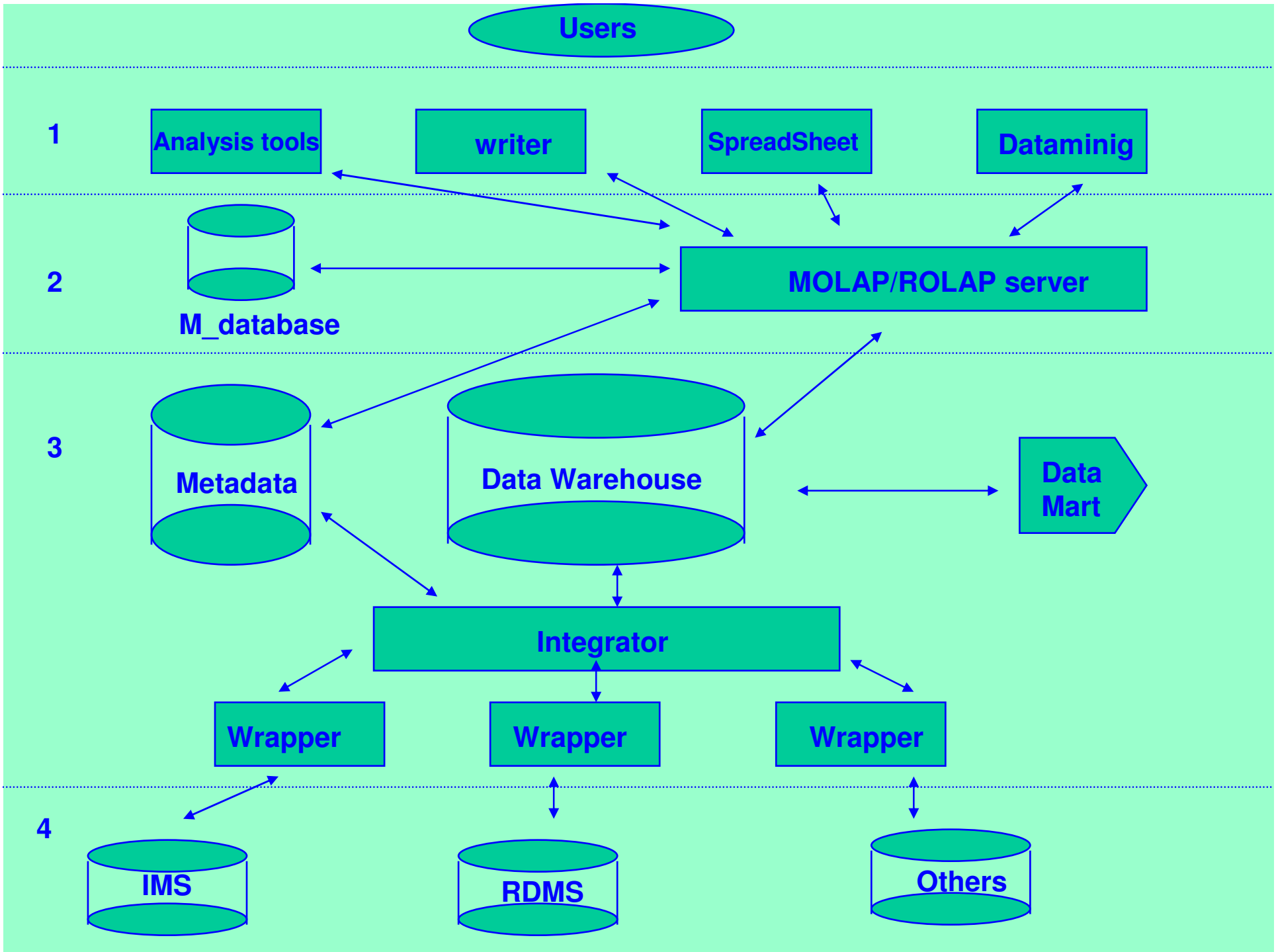


Pi : analytical and mining procedures (kernel parallel)

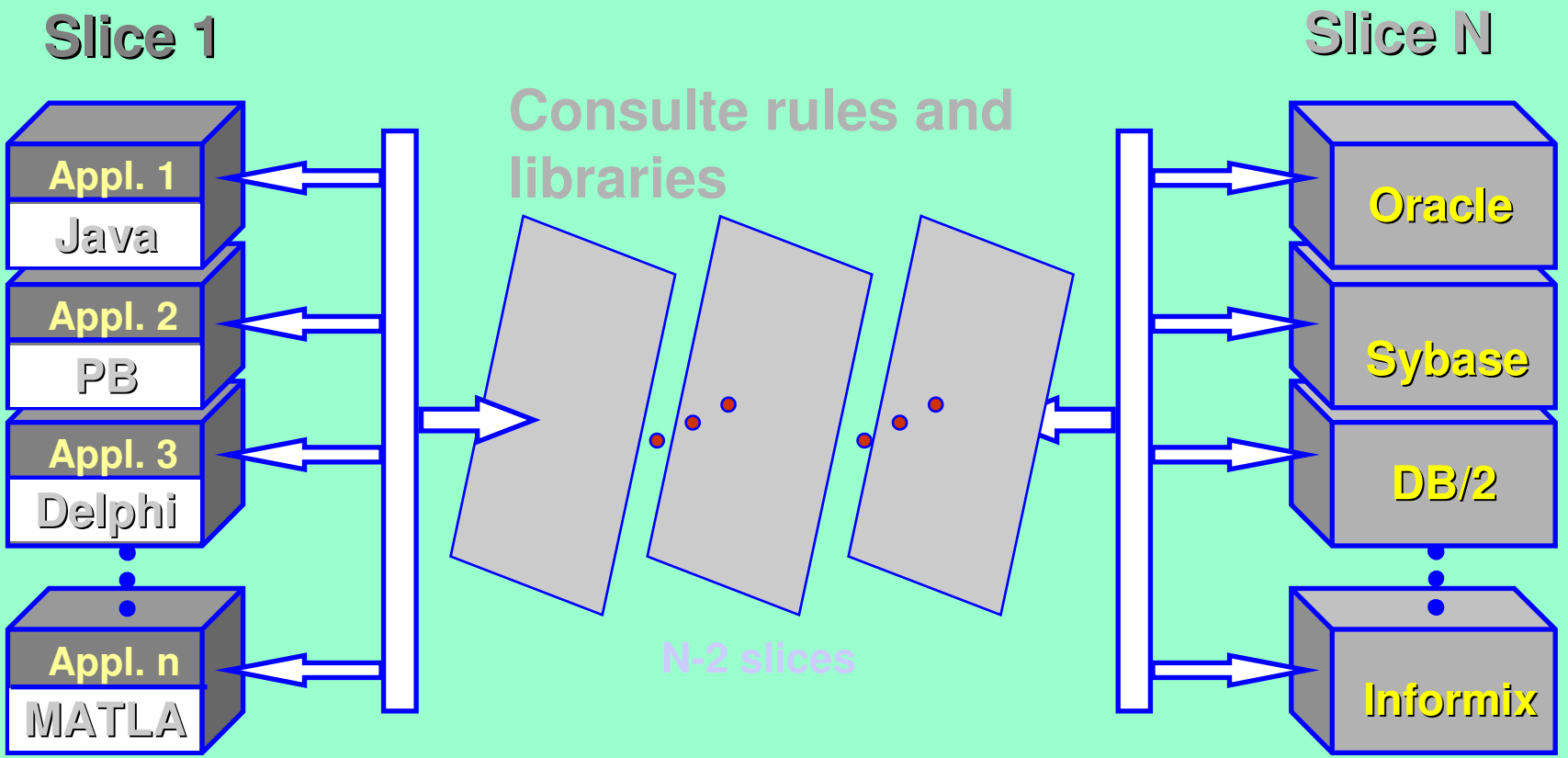

images
Express
Table
Signal
clusters
Dynamical
System

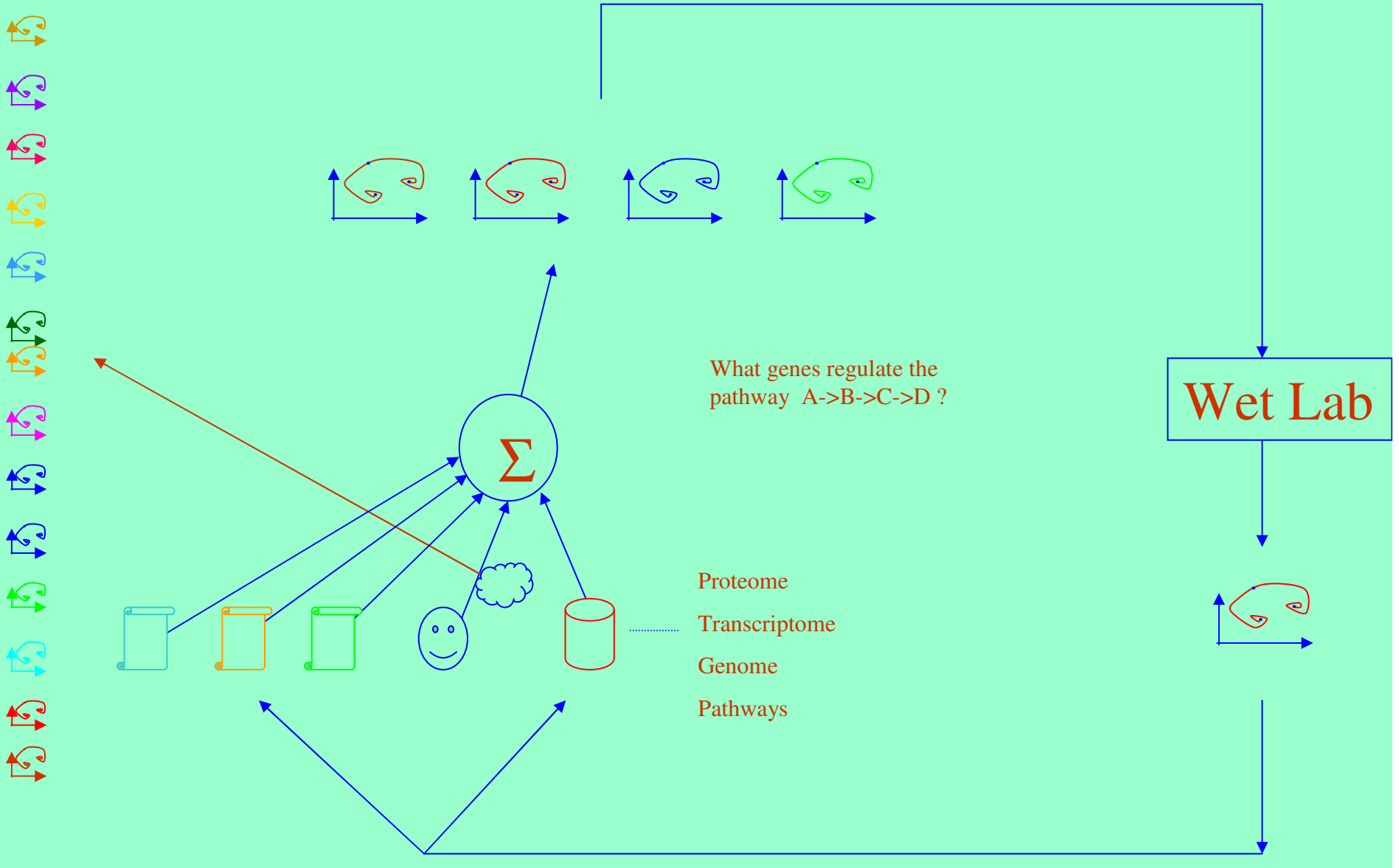
Integrated Environment

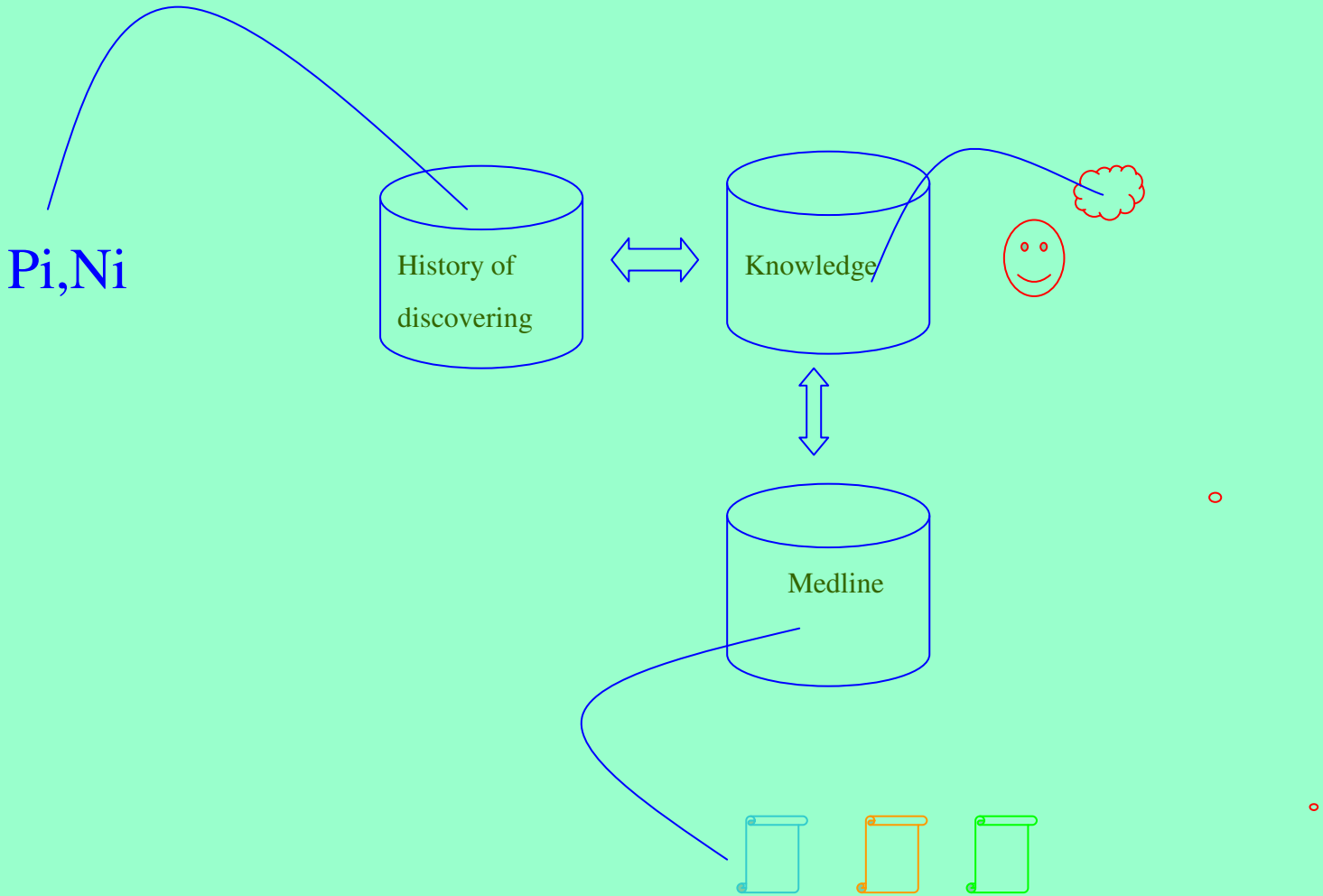



Users
1
Analysis tools
writer
SpreadSheet
Dataminig
2
M_database
MOLAP/ROLAP server
3
Metadata
Data Warehouse
Data Mart
Integrator
Wrapper
Wrapper
Wrapper
4
IMS
RDMS
Others

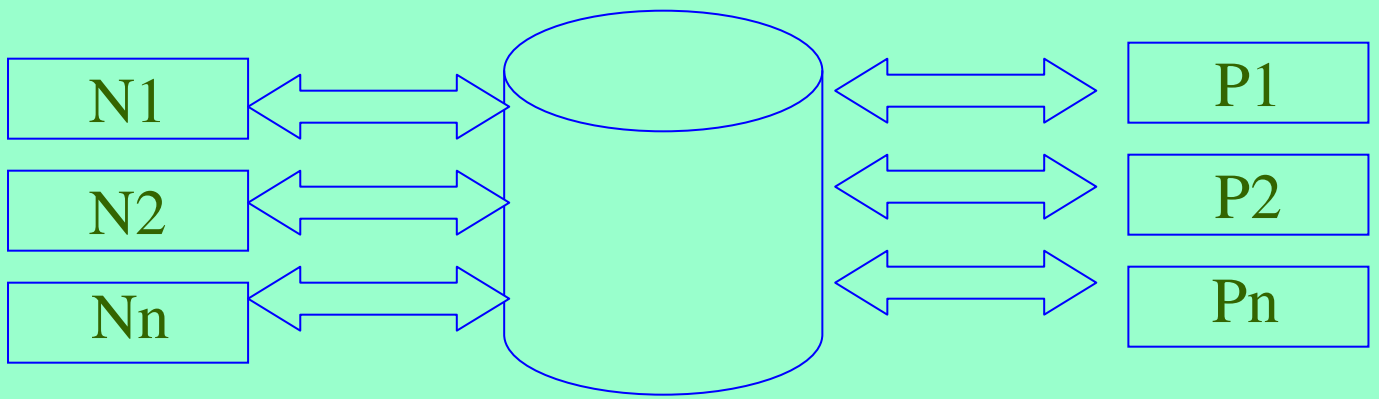
System Architecture




What genes regulate the pathway A->B->C->D ?
Wet Lab
Proteome
Transcriptome
Genome
Pathways

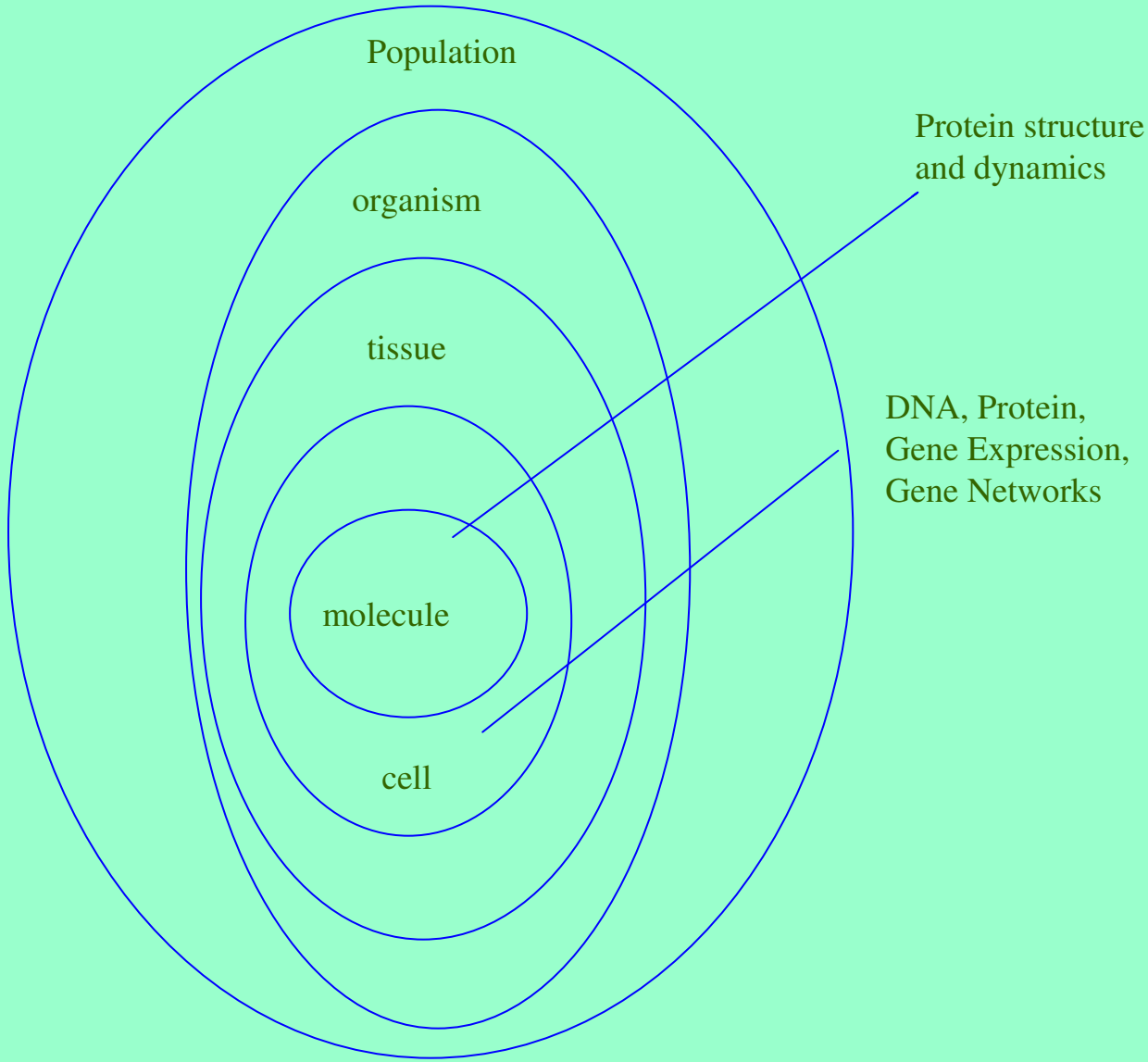

Pi,Ni
History of
discovering
Knowledge
Medline

Objected oriented database

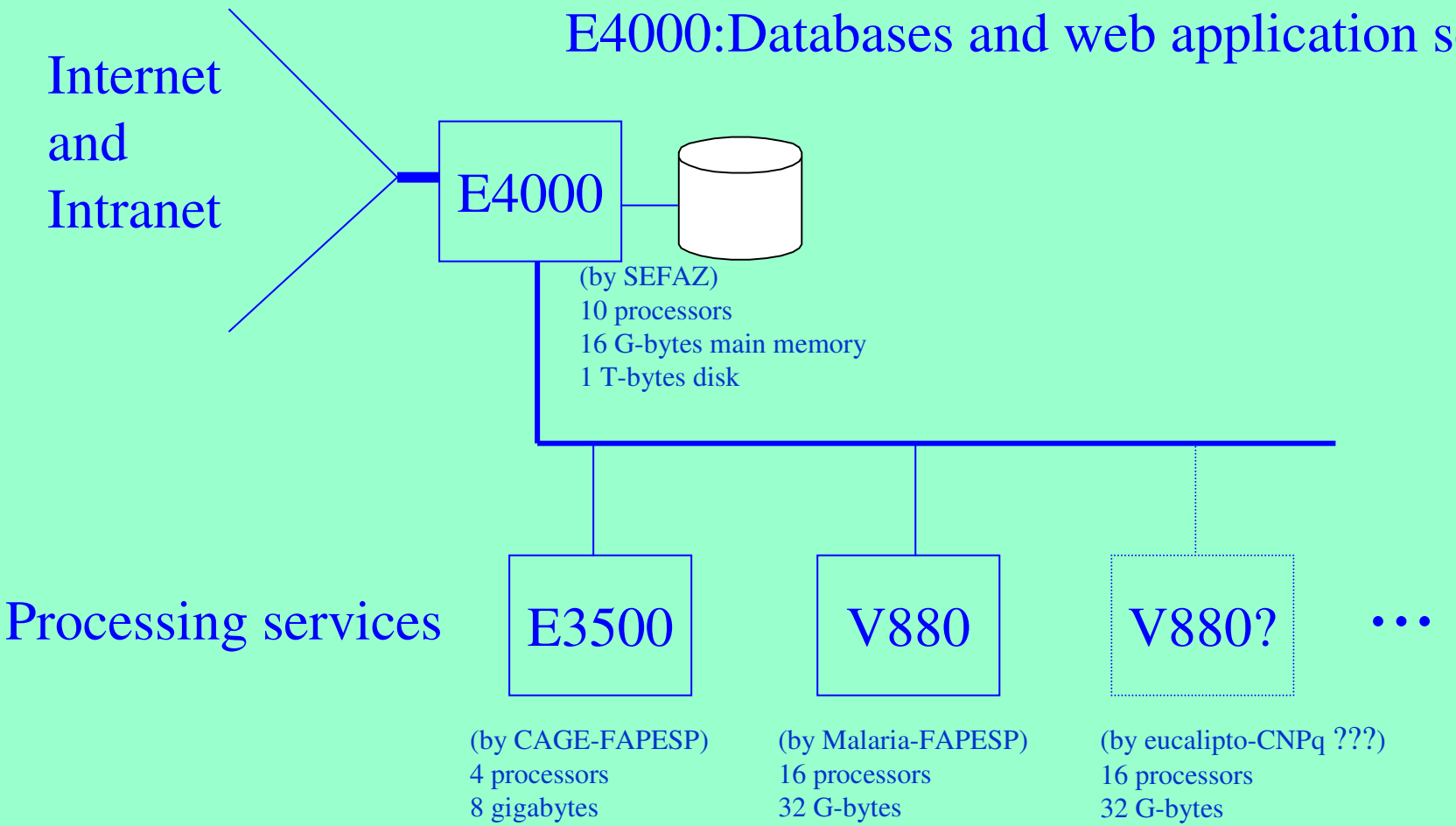


Pi : analytical and mining procedures (kernel parallel)

Ni : knowledge discovering procedures (kernel parallel)


Population
organism
tissue
molecule
cell
Protein structure
and dynamics
DNA, Protein,
Gene Expression,
Gene Networks

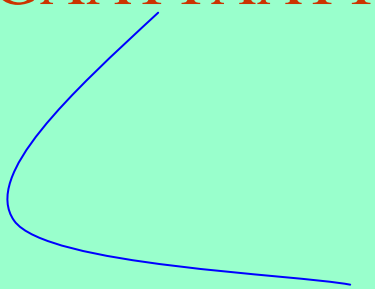
GRID Computer - DCC-IME-USP



Mapping of rare genes

ACGAATCTAGAGAATTAATTAACCGAGTTAAGA

ACGAATCTAGAGAATTAATTAACCGAGTTAAGA



Exon

Training from known genes

Expression Analysis

Analysis Phases

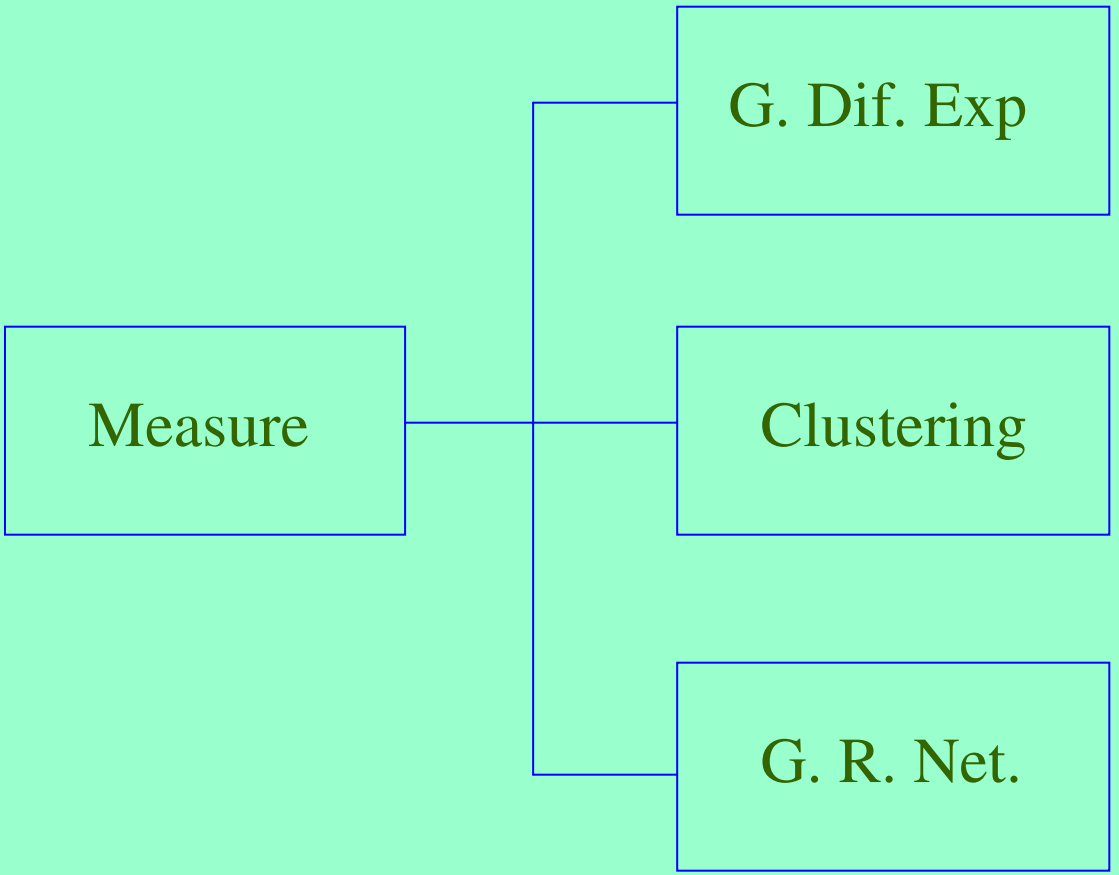
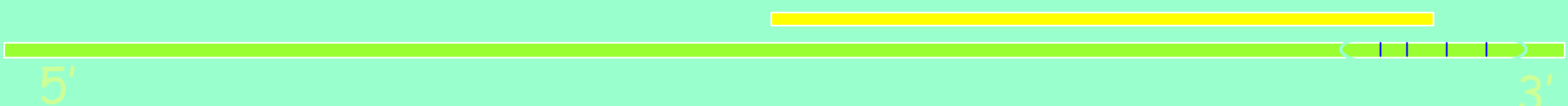


Image Analysis

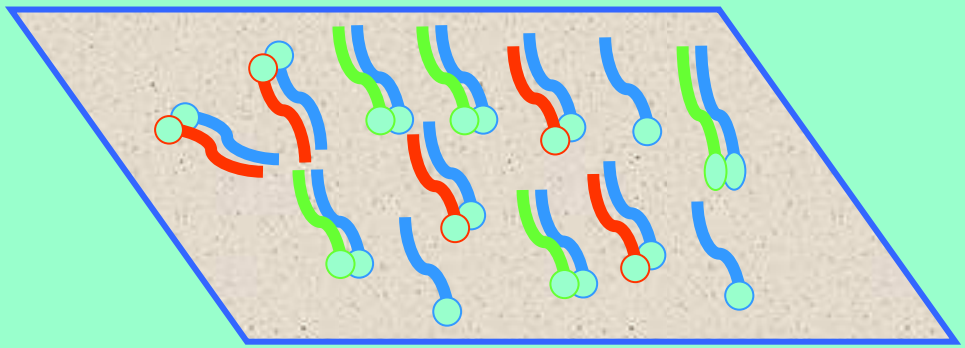
Selection of clones

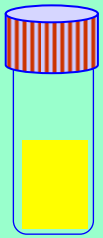
1. Clusters of the same gene

2. Choice of a representation for the gene

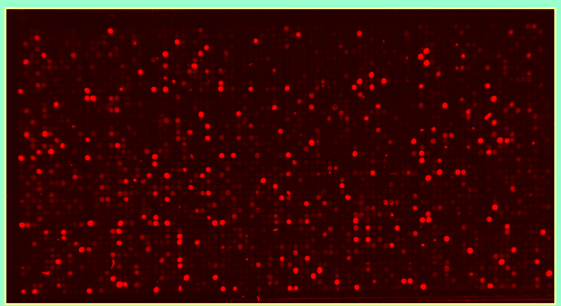


Hibridization

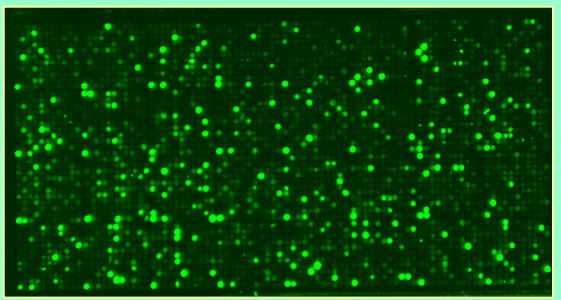




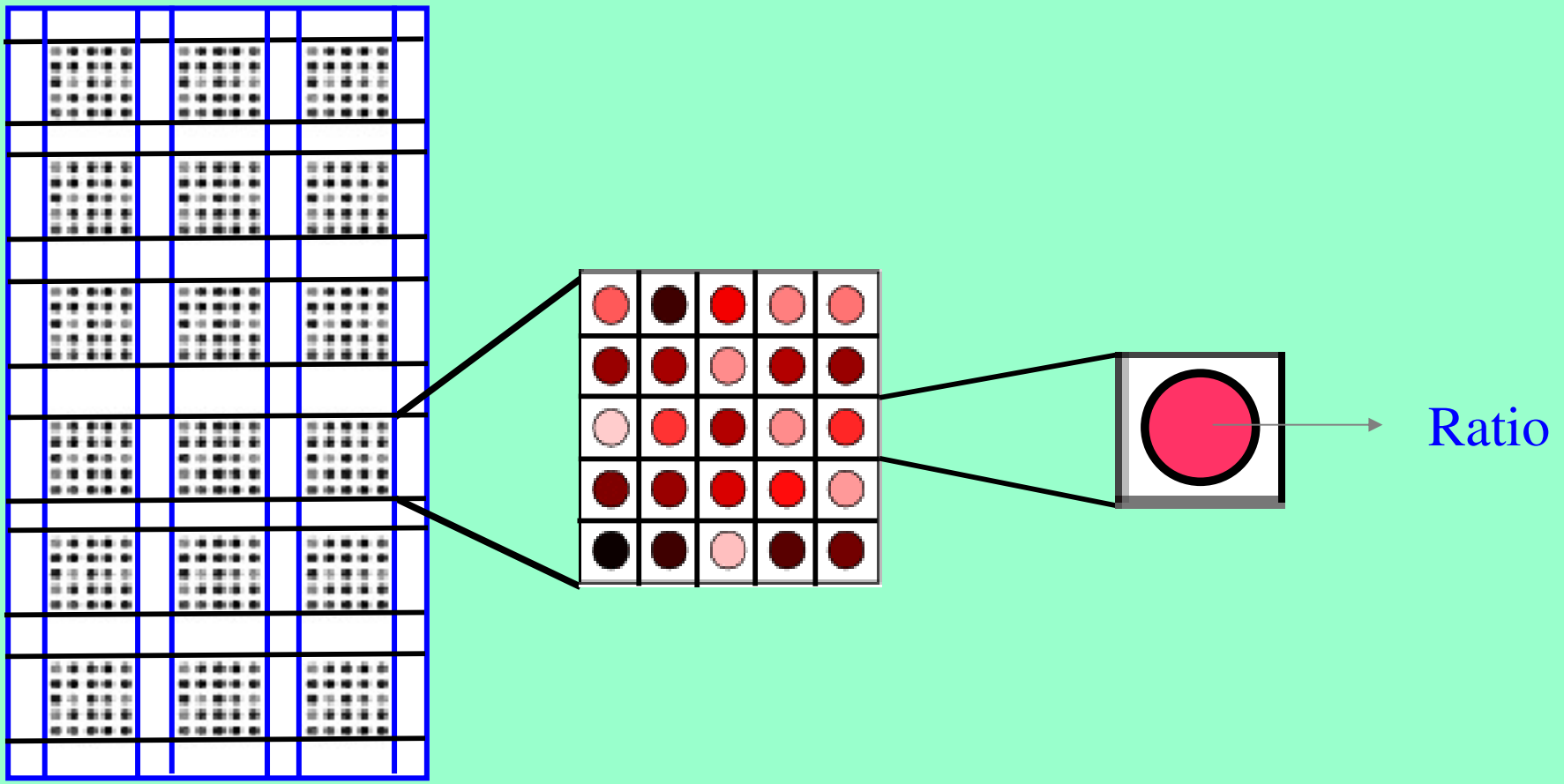
Cy5

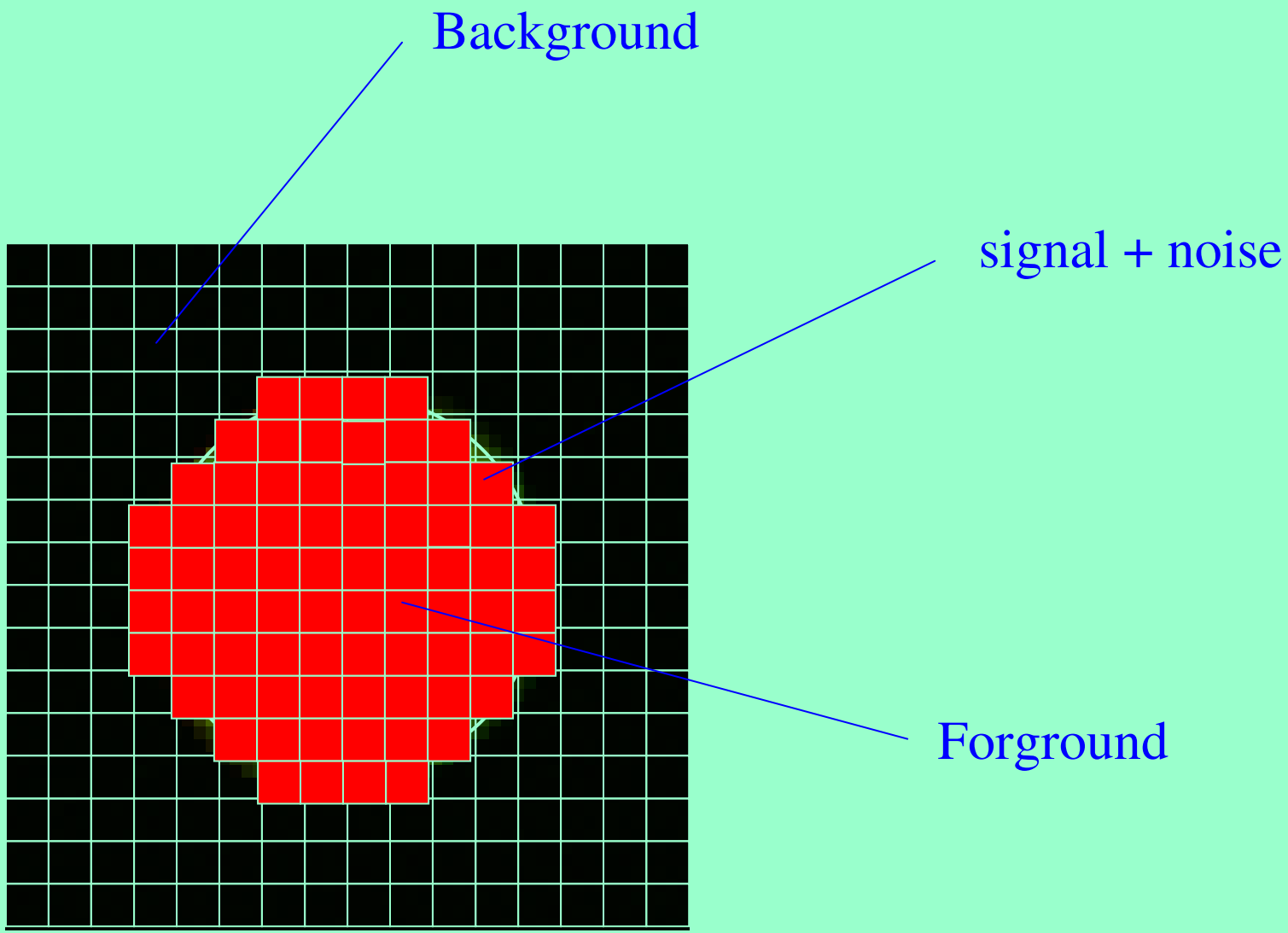


Cy3

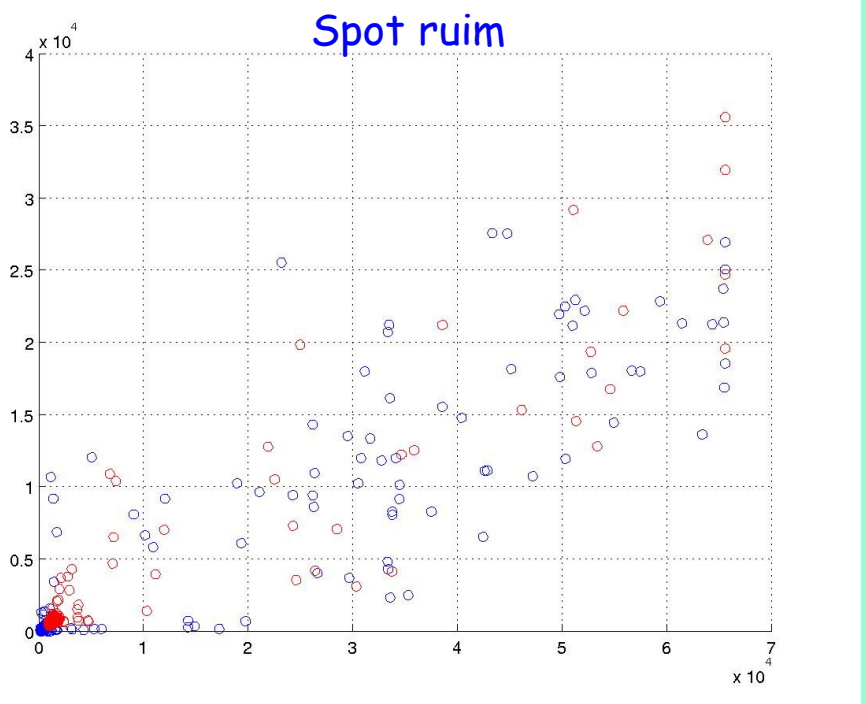


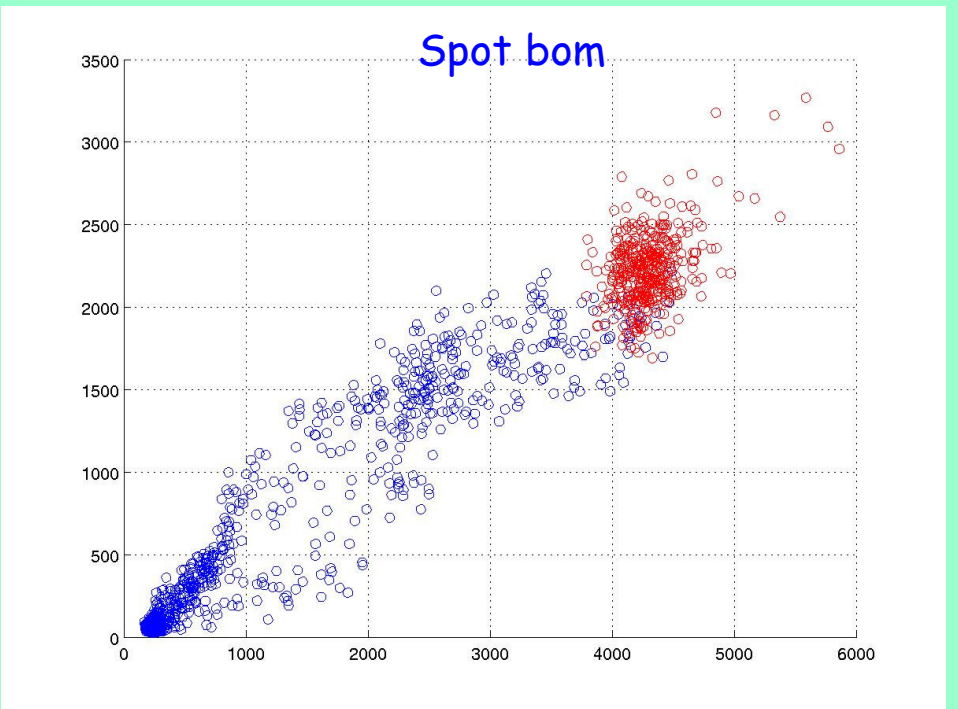
Expression Calculus




Background
signal + noise
Forground

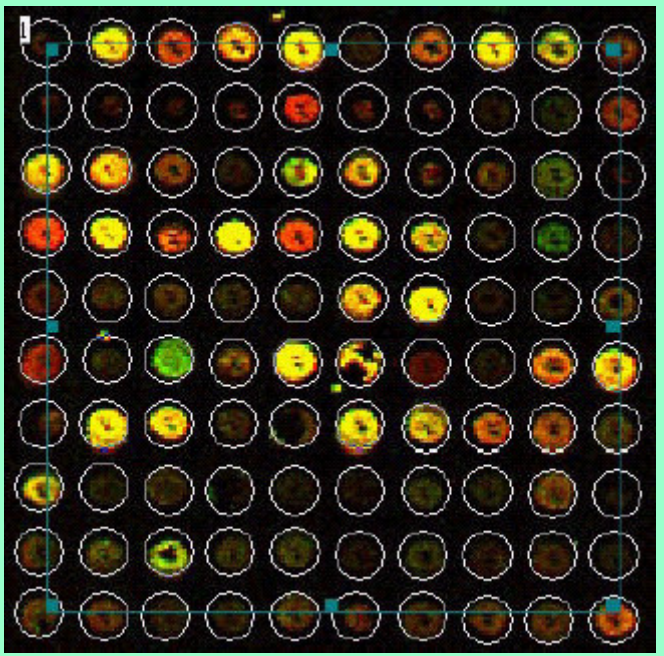
Dispersion of cy3 and cy5 give an idea of the spot quality

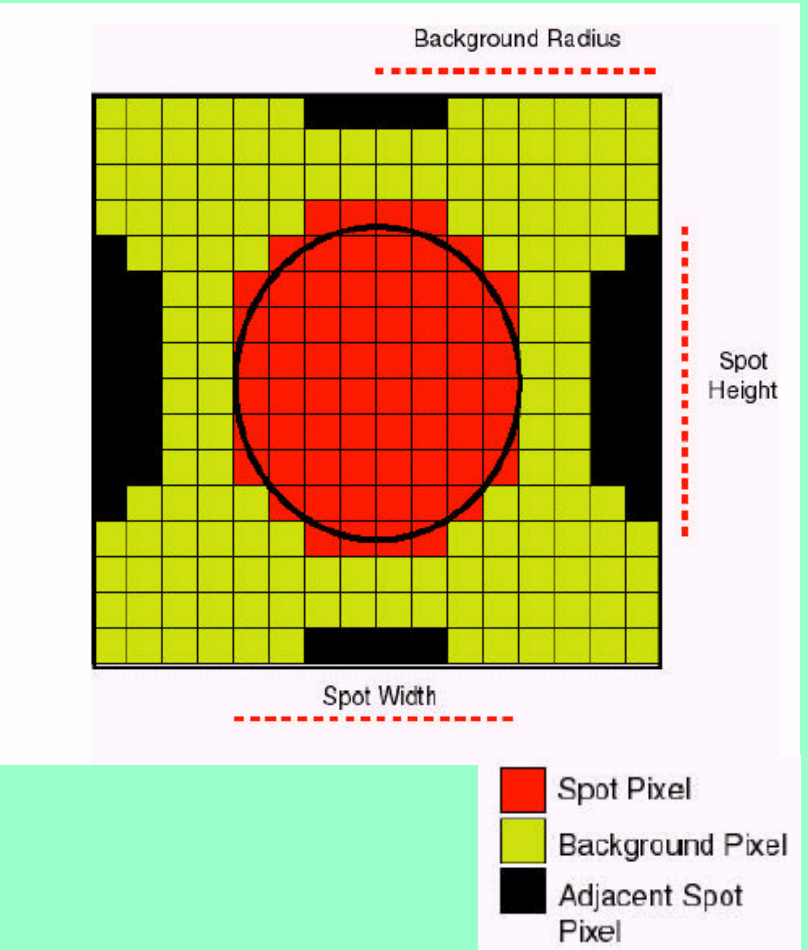




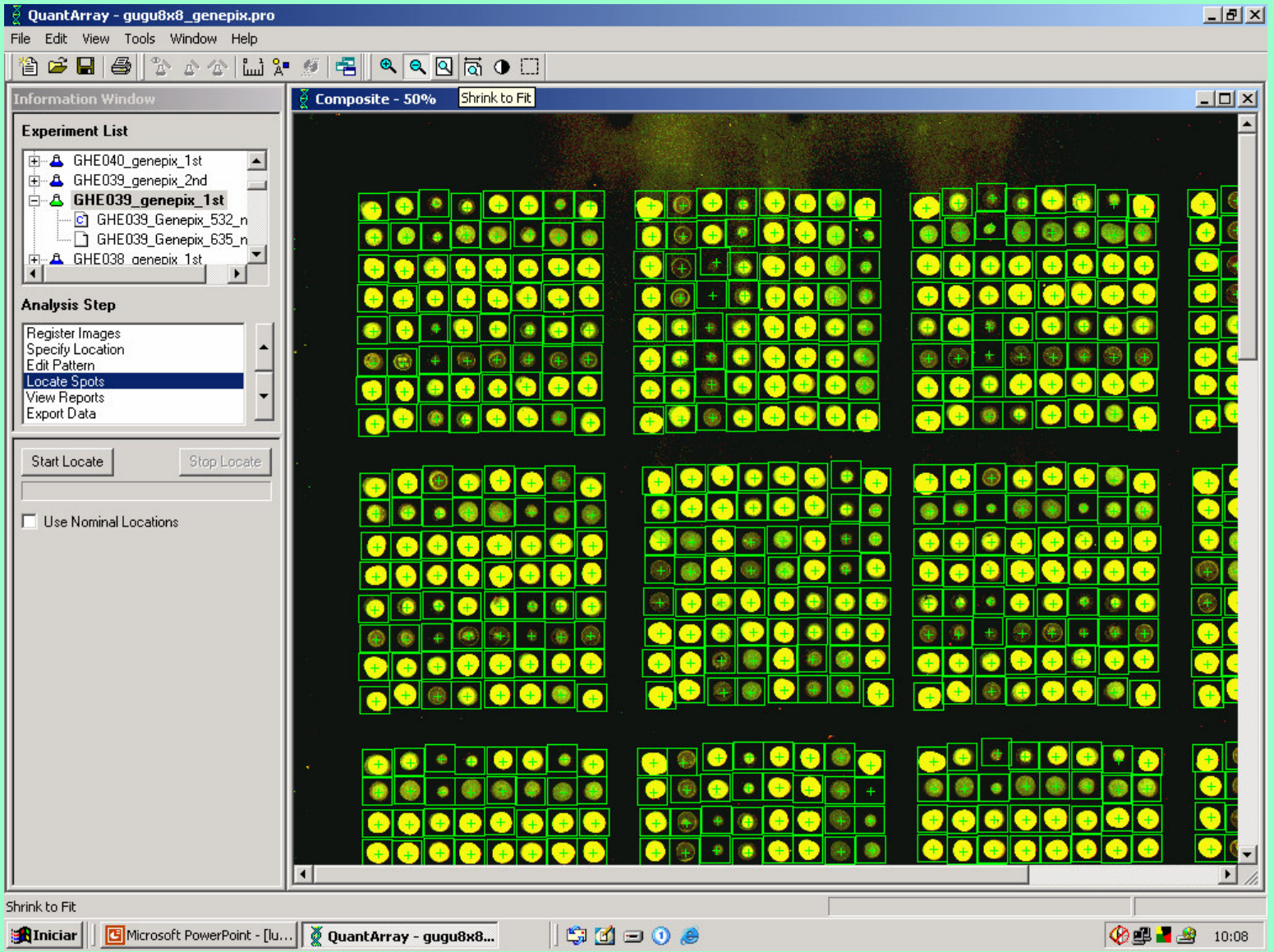
- *Pixels* do *background*
- *Pixels* do *foreground*

ScanAlyze

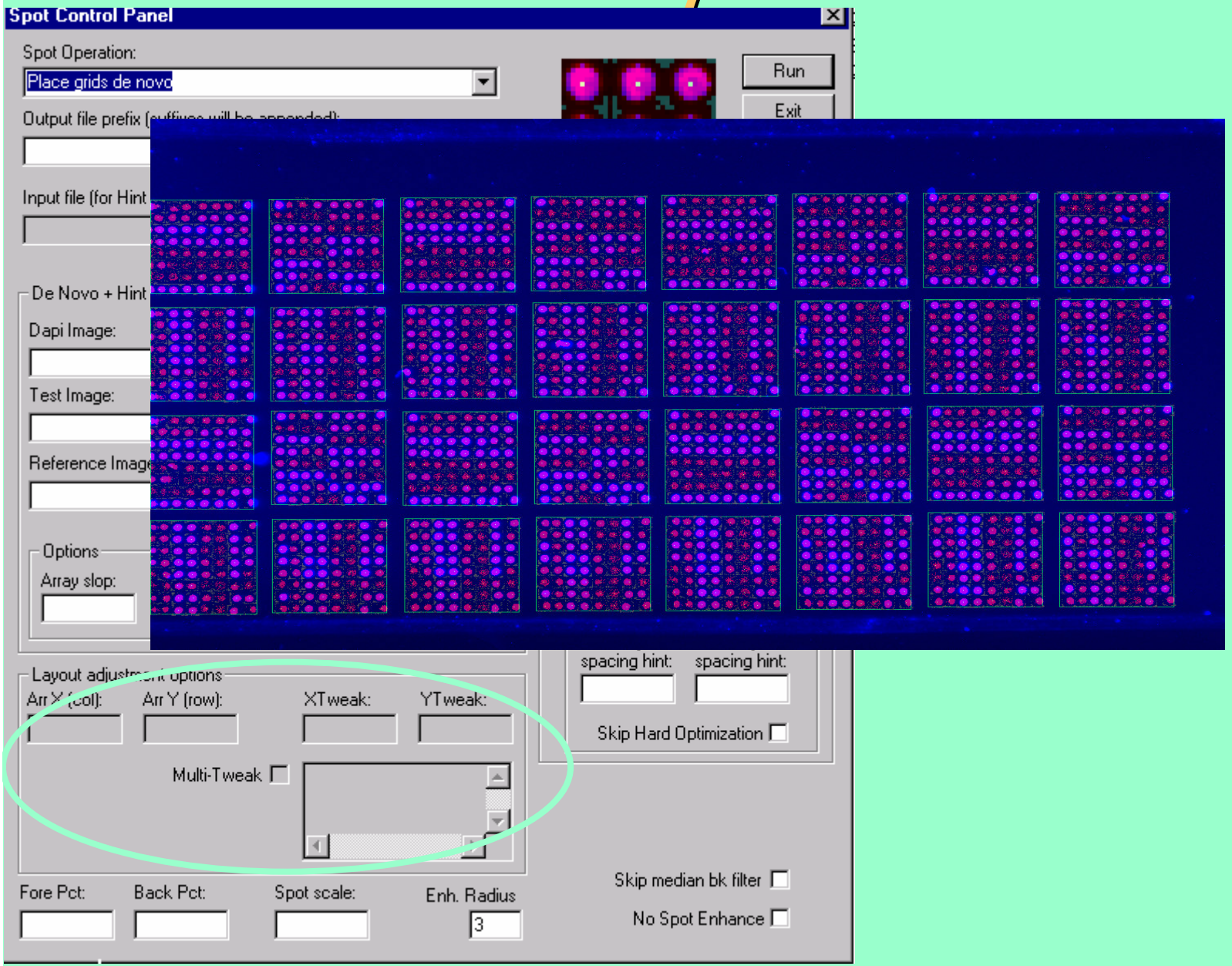



Background Radius
Spot Height
Spot Width
Spot Pixel
Background Pixel
Adjacent Spot Pixel

QuantArray®


QuantArray - gugu8x8_genepix.pro
File Edit View Tools Window Help
Information Window
Experiment List
GHE040_genepix_1st
GHE039_genepix_2nd
GHE039_genepix_1st
GHE039_Genepix_532_n
GHE039_Genepix_635_n
GHE038_genepix_1st
Analysis Step
Register Images
Specify Location
Edit Pattern
Locate Spots
View Reports
Export Data
Start Locate
Stop Locate
Use Nominal Locations
Composite - 50%
Shrink to Fit
Shrink to Fit
Iniciar
Microsoft PowerPoint - [lu...
QuantArray - gugu8x8...
10:08

Spot


Spot Control Panel
Spot Operation:
Place grids de novo
Run
Exit
Output file prefix
Input file (for Hint
De Novo + Hint
Dapi Image:
Test Image:
Reference Image
Options
Array slop:
spacing hint:
spacing hint:
Skip Hard Optimization
Layout adjustment options
Arr X (col):
Arr Y (row):
XTweak:
YTweak:
Multi-Tweak
Fore Pct:
Back Pct:
Spot scale:
Enh. Radius
3
Skip median bk filter
No Spot Enhance

BIOINFO - USP

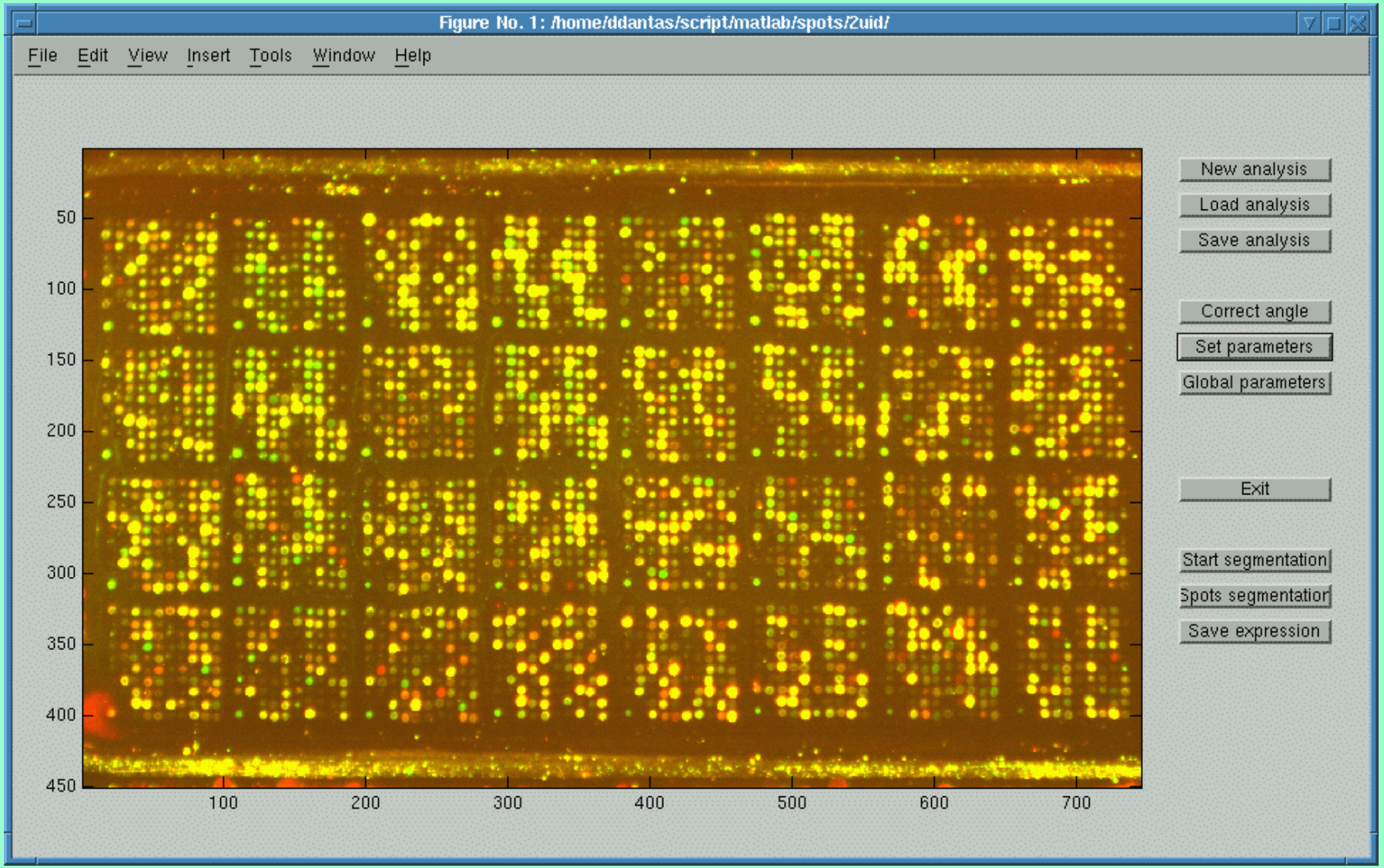
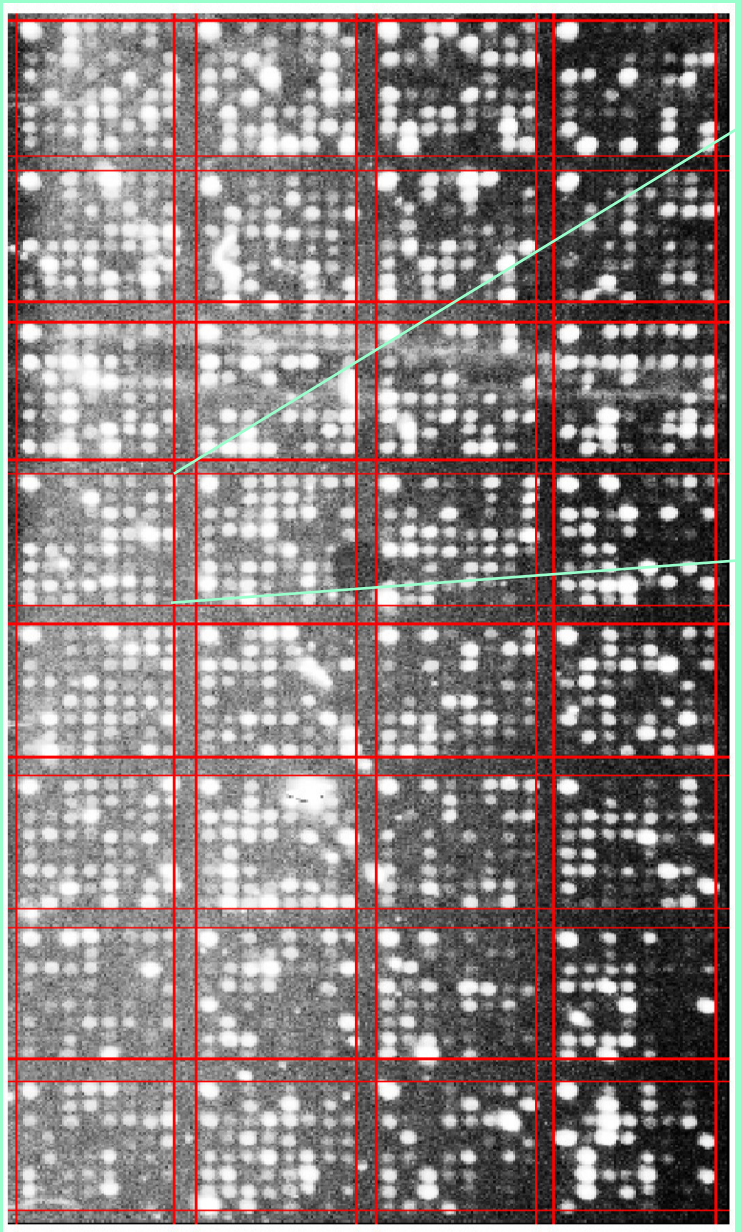
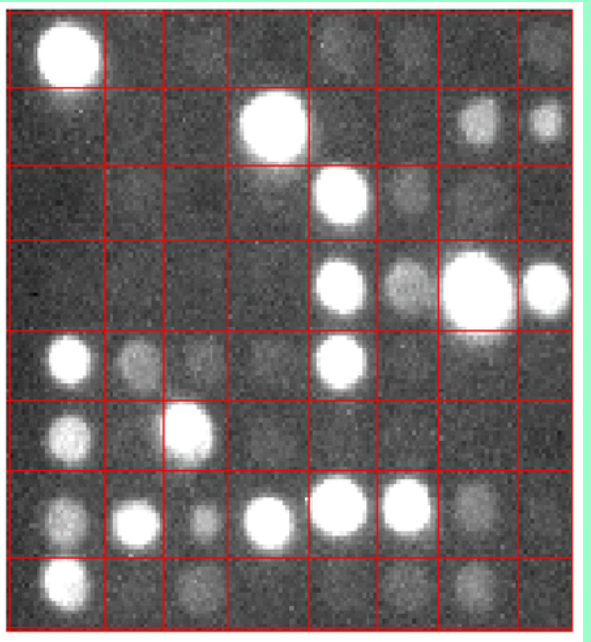
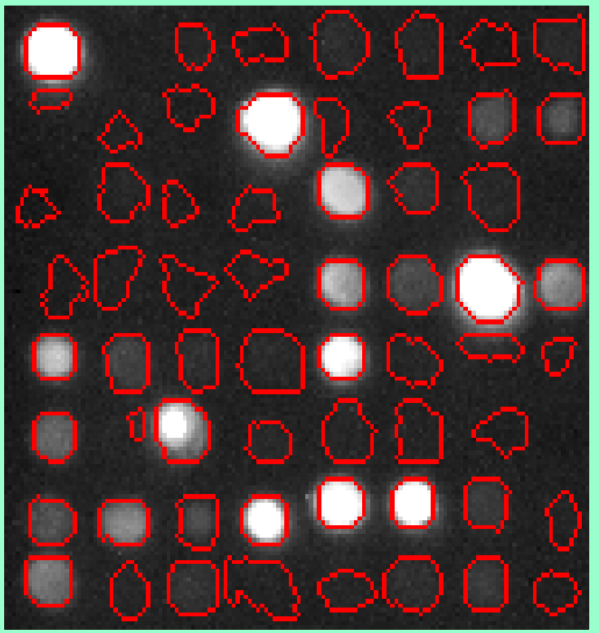


Figure No. 1: /home/ddantas/script/matlab/spots/2uid/
File Edit View Insert Tools Window Help
New analysis
Load analysis
Save analysis
Correct angle
Set parameters
Global parameters
Exit
Start segmentation
Spots segmentation
Save expression
50
100
150
200
250
300
350
400
450
100
200
300
400
500
600
700

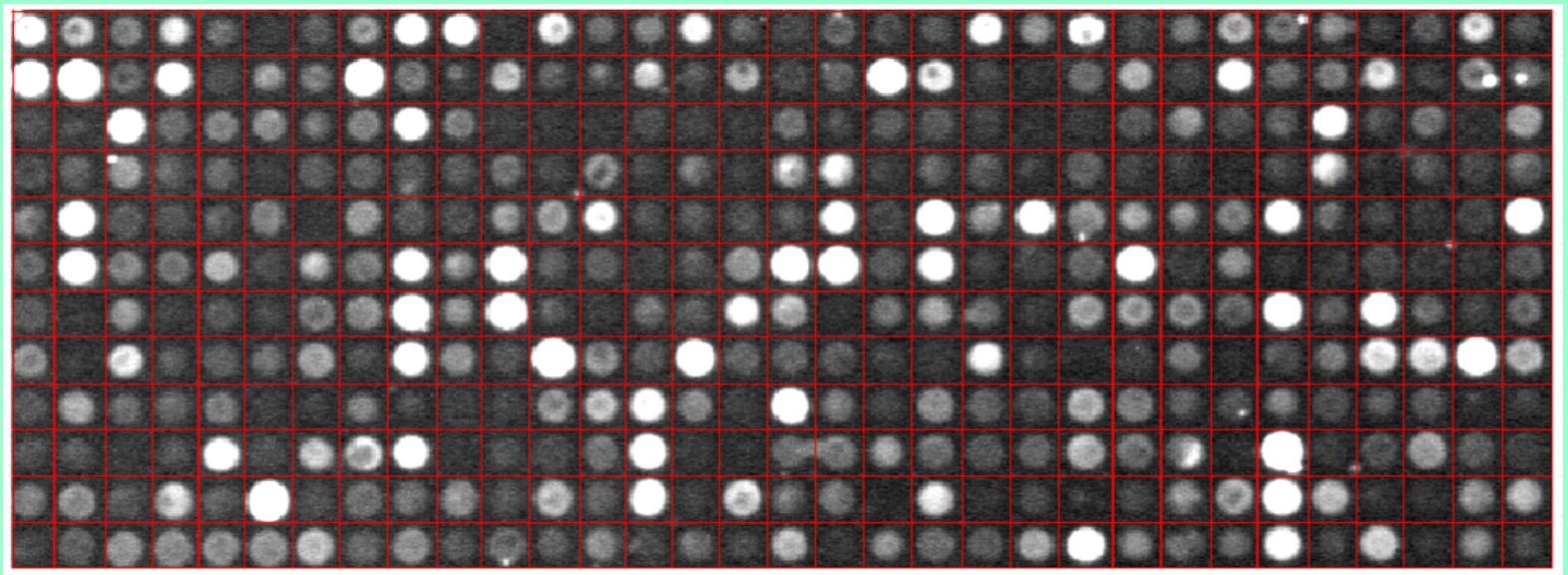
Example - Segmentation

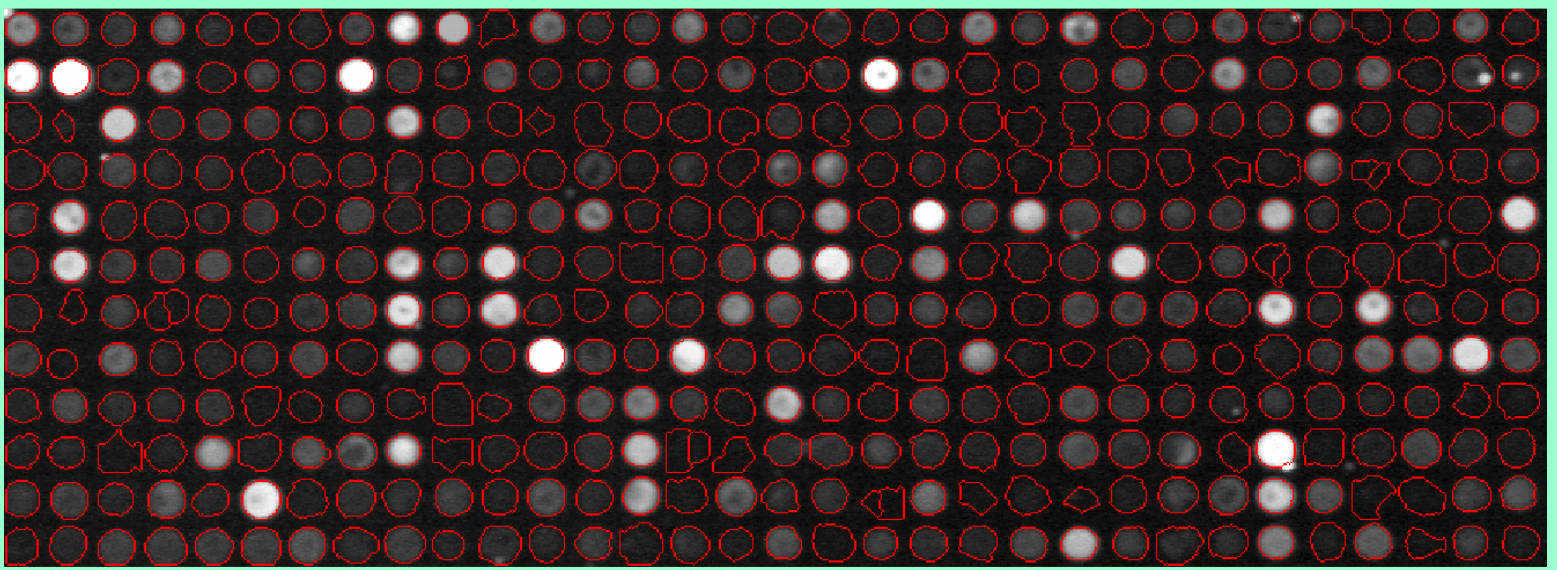






Exemplo - Segmentação





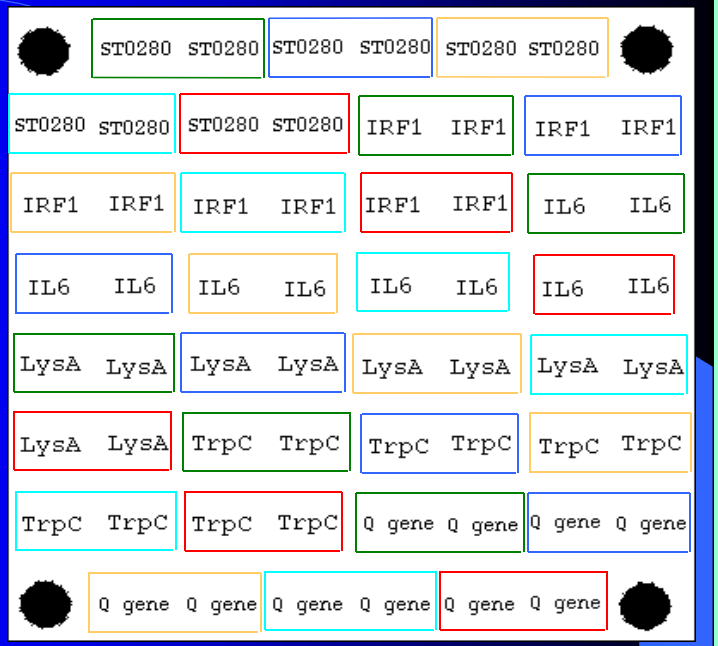
cDNAs used

- ✓ lysA - 303pb, 47,2% de C e G
- ✓ trpC - 338pb, 45,3% de C e G
- ✓ Q gene - 637pb, 52,3% de C e G
- ✓ ST0280 (ORESTES) - 659pb, 34,6% de C e G
- ✓ IL6 - 948pb, 37,7% de C e G
- ✓ IRF1 - 2069pb, 52,5% de C e G

Dilution of fixed material

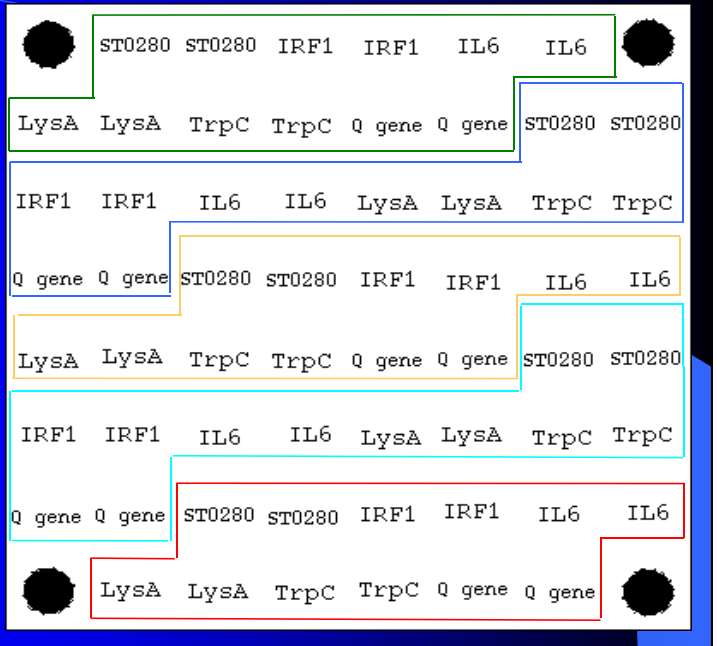
Each cDNA will be fixed in the dilutions
1/1, 1/2, 1/4, 1/8, 1/16

B1



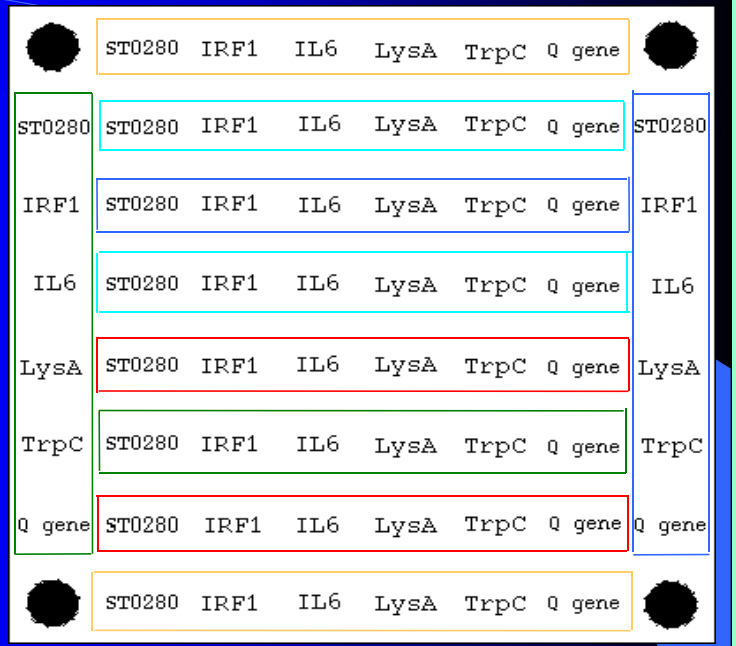
Legenda
- Diluição 1
- Diluição 2
- Diluição 3
- Diluição 4
- Diluição 5

B2



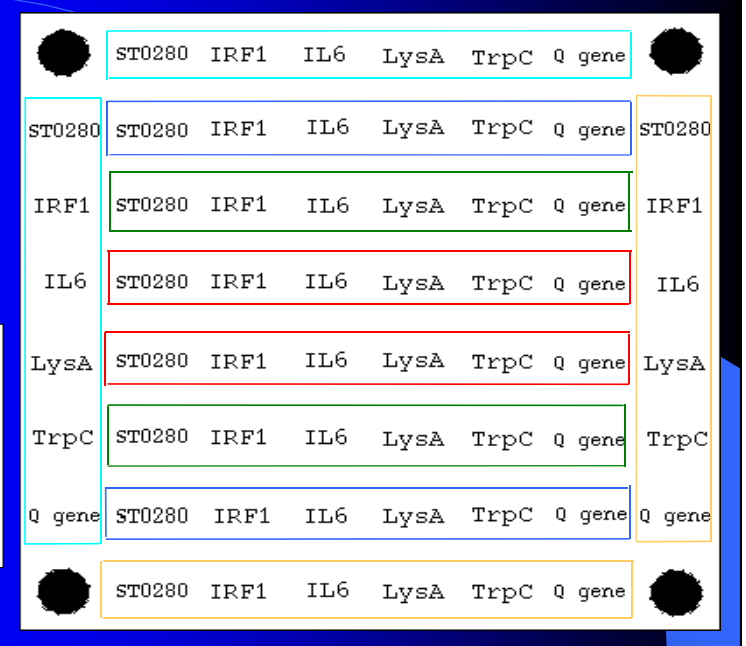
Legenda
- Diluição 1
- Diluição 2
- Diluição 3
- Diluição 4
- Diluição 5

B3

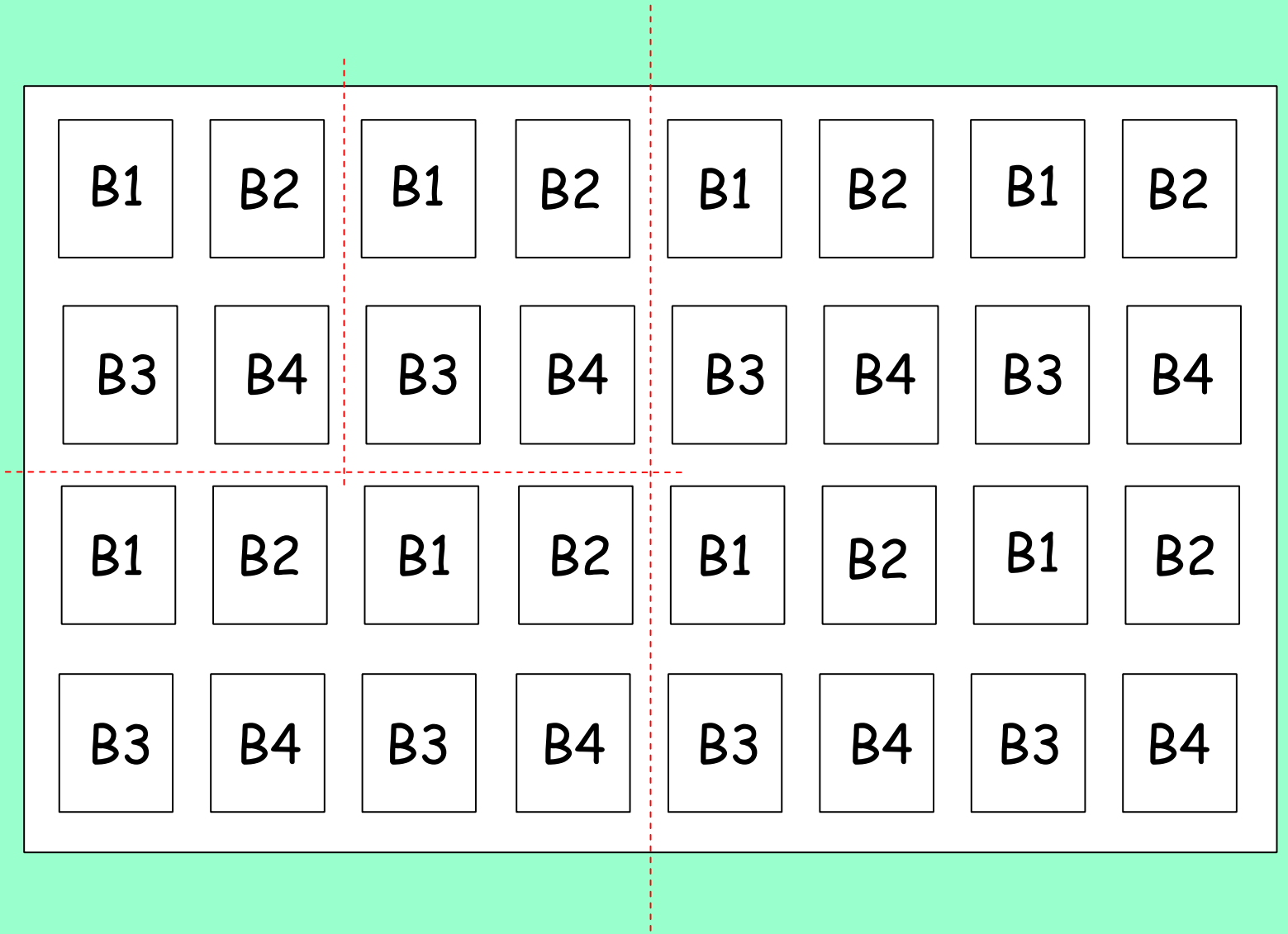


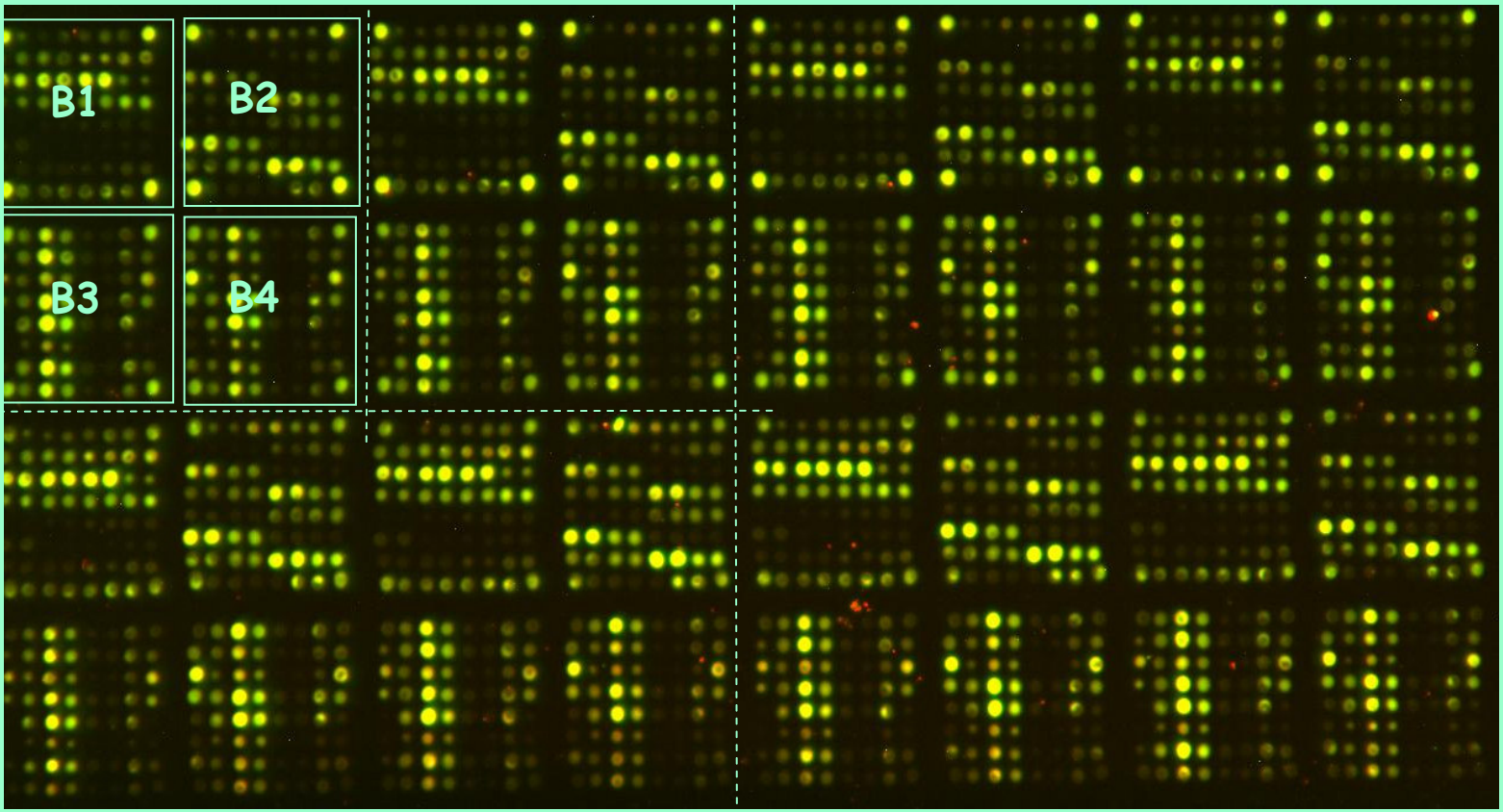
Legenda
- Diluição 1
- Diluição 2
- Diluição 3
- Diluição 4
- Diluição 5

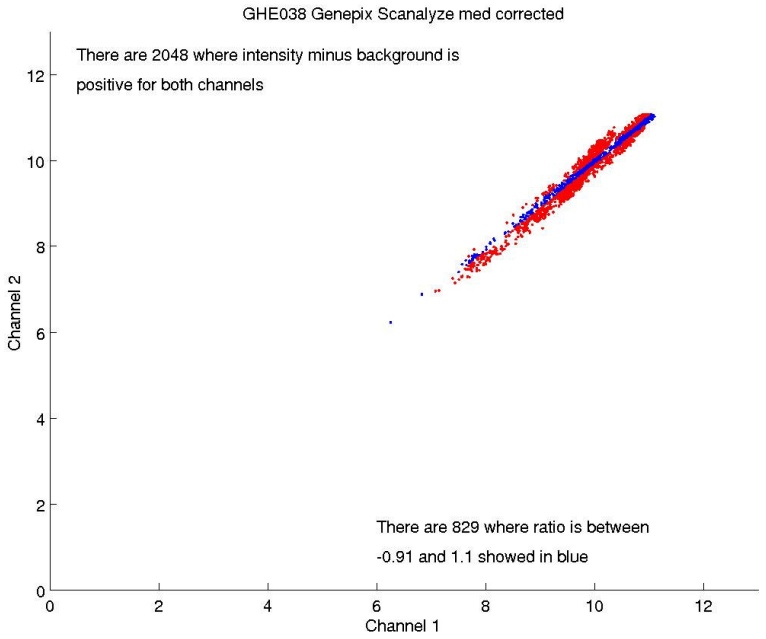
B4

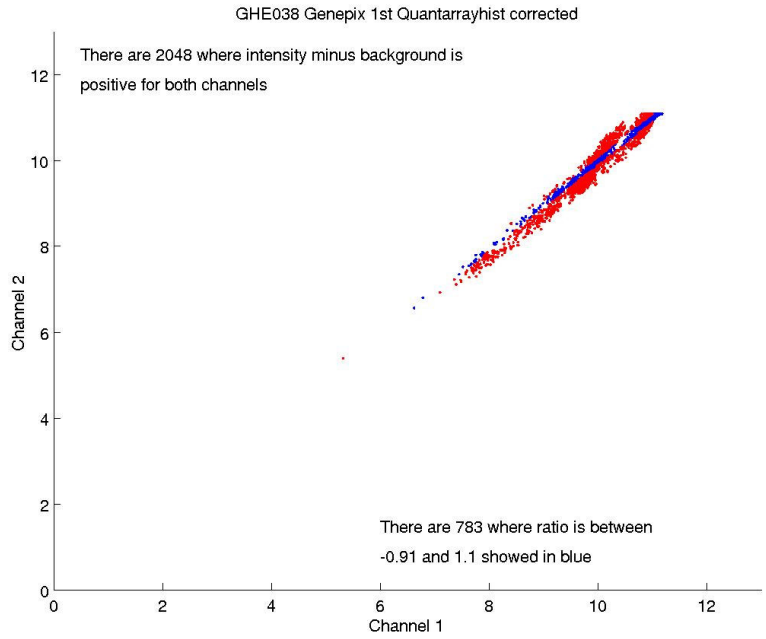


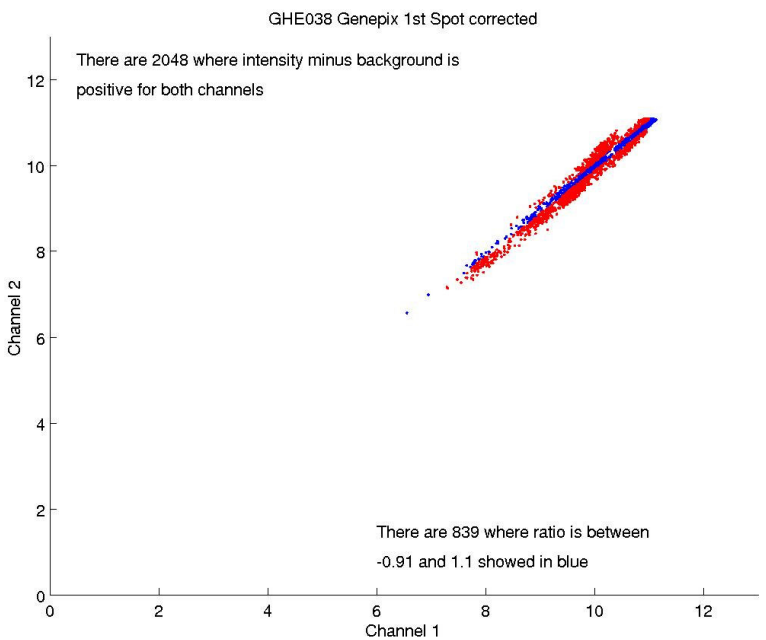
Legenda
- Diluição 1
- Diluição 2
- Diluição 3
- Diluição 4
- Diluição 5

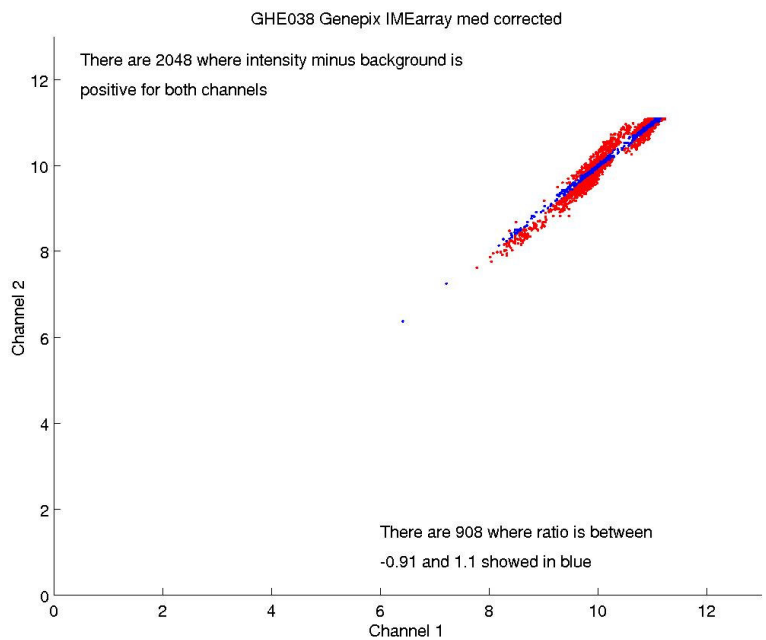

B1 B2 B1 B2 B1 B2 B1 B2
B3 B4 B3 B4 B3 B4 B3 B4
B1 B2 B1 B2 B1 B2 B1 B2
B3 B4 B3 B4 B3 B4 B3 B4

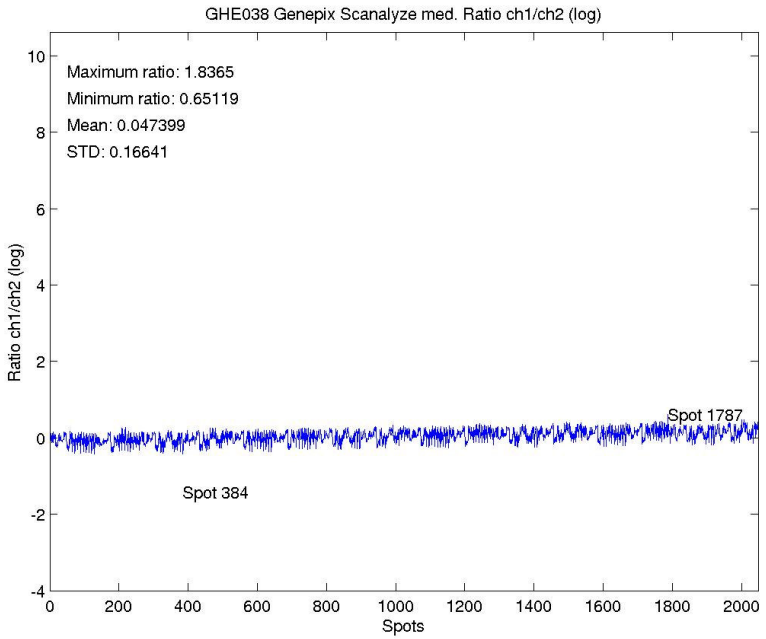

B1
B2
B3
B4

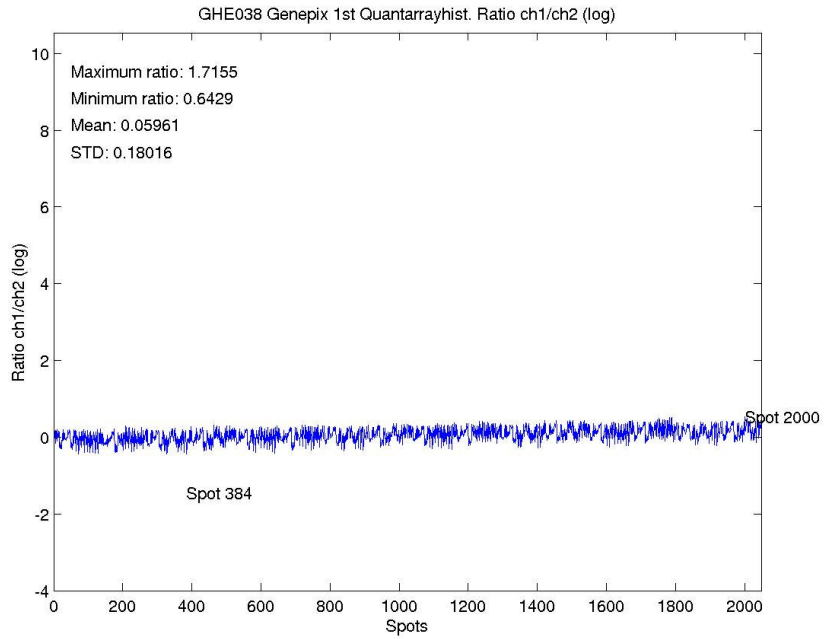

GHE038 Genepix Scanalyze med corrected
There are 2048 where intensity minus background is
positive for both channels
There are 829 where ratio is between
-0.91 and 1.1 showed in blue
Channel 2
Channel 1
0
2
4
6
8
10
12

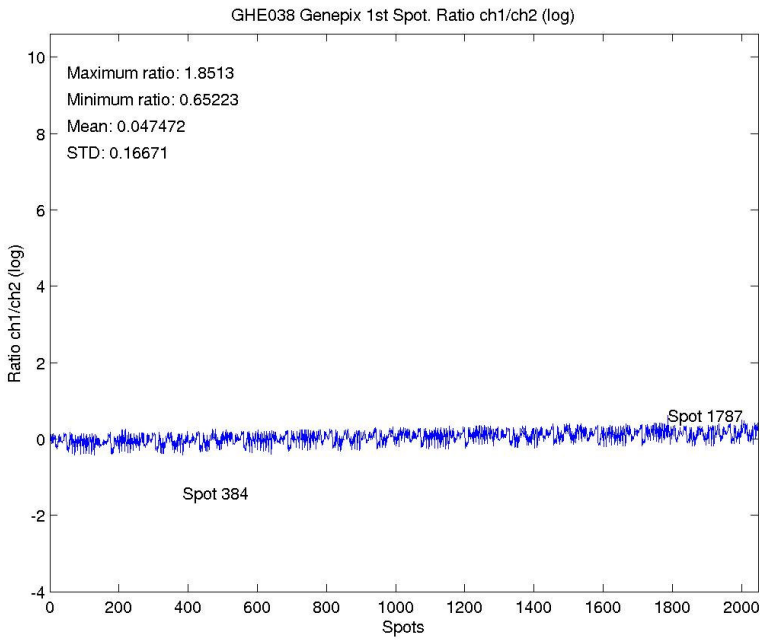

GHE038 Genepix 1st Quantarrayhist corrected
There are 2048 where intensity minus background is
positive for both channels
There are 783 where ratio is between
-0.91 and 1.1 showed in blue
Channel 2
Channel 1
0
2
4
6
8
10
12

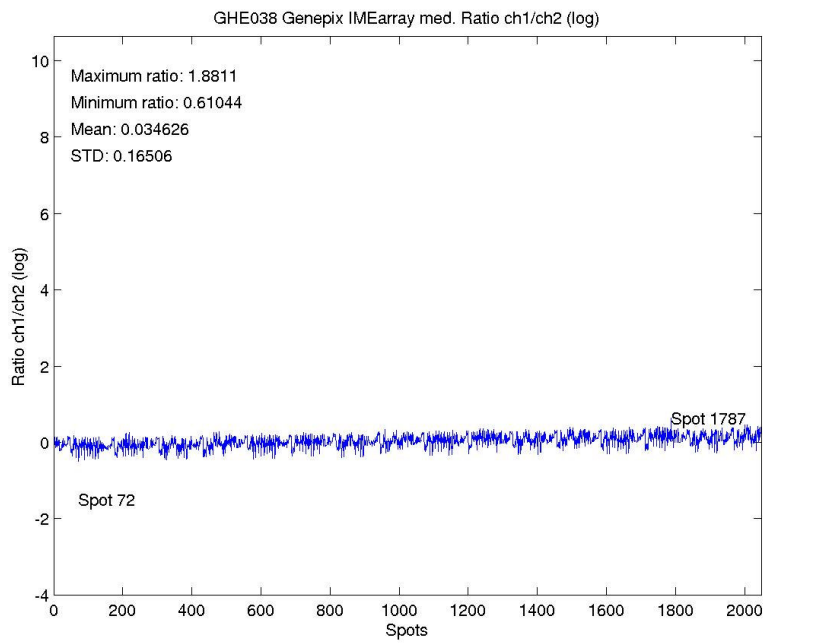

GHE038 Genepix 1st Spot corrected
There are 2048 where intensity minus background is
positive for both channels
There are 839 where ratio is between
-0.91 and 1.1 showed in blue
Channel 2
Channel 1
0
2
4
6
8
10
12


GHE038 Genepix IMEarray med corrected
There are 2048 where intensity minus background is
positive for both channels
There are 908 where ratio is between
-0.91 and 1.1 showed in blue
Channel 2
Channel 1
0
2
4
6
8
10
12


GHE038 Genepix Scanalyze med. Ratio ch1/ch2 (log)
Maximum ratio: 1.8365
Minimum ratio: 0.65119
Mean: 0.047399
STD: 0.16641
Spot 1787
Spot 384
Ratio ch1/ch2 (log)
Spots


GHE038 Genepix 1st Quantarrayhist. Ratio ch1/ch2 (log)
Maximum ratio: 1.7155
Minimum ratio: 0.6429
Mean: 0.05961
STD: 0.18016
Spot 2000
Spot 384
Ratio ch1/ch2 (log)
Spots


GHE038 Genepix 1st Spot. Ratio ch1/ch2 (log)
Maximum ratio: 1.8513
Minimum ratio: 0.65223
Mean: 0.047472
STD: 0.16671
Spot 1787
Spot 384
Ratio ch1/ch2 (log)
Spots


GHE038 Genepix IMEarray med. Ratio ch1/ch2 (log)
Maximum ratio: 1.8811
Minimum ratio: 0.61044
Mean: 0.034626
STD: 0.16506
Spot 1787
Spot 72
Ratio ch1/ch2 (log)
Spots

For a good signal:

- the linear regression is a good estimator for ratio (background estimation is avoided)
- Swap permits to normalize cy3 and cy5
- Confidence intervals increase inversely with the signal intensity

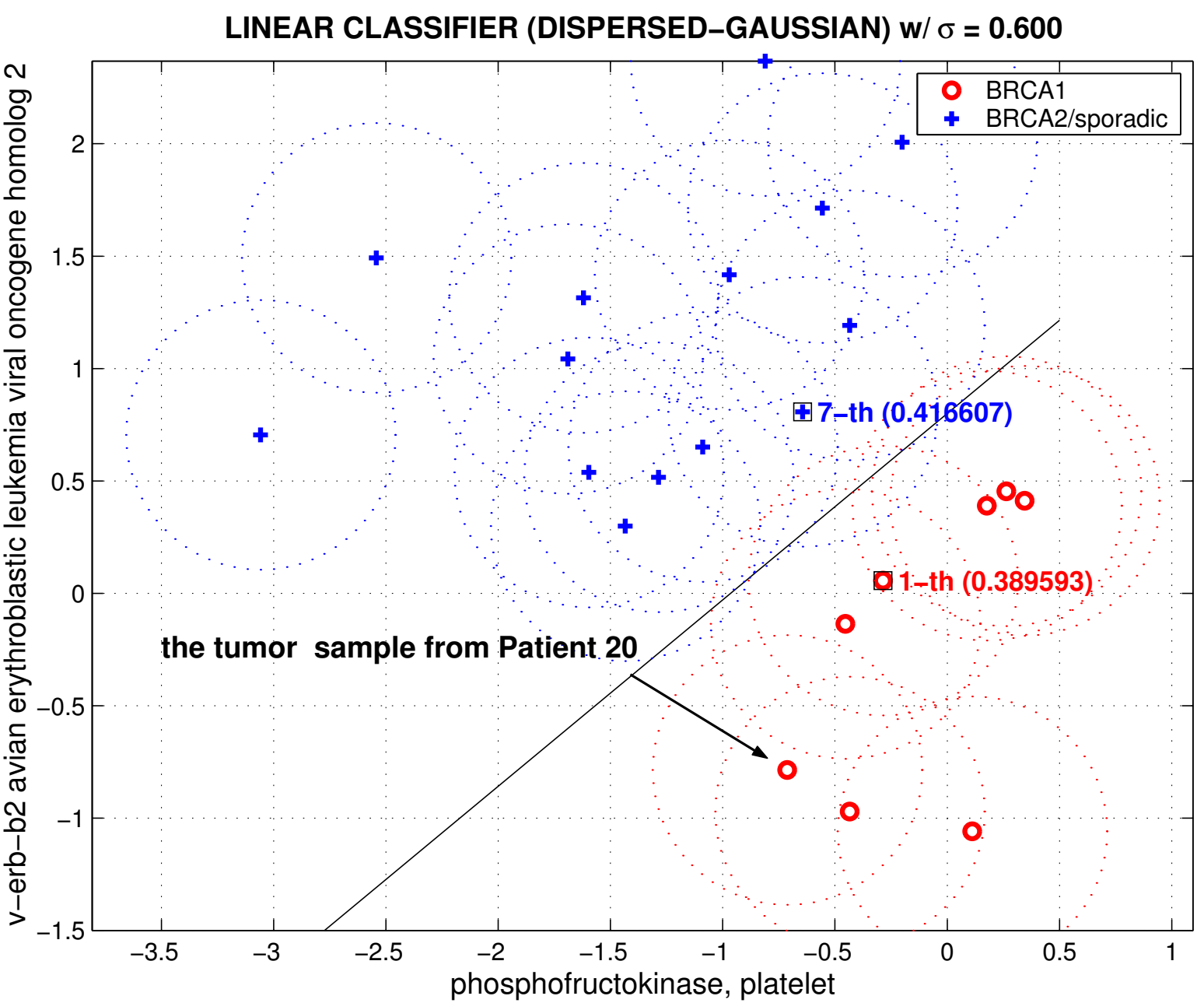
Genes differentially expressed

Experiment

- Choose a population of men and measure their physical characteristics
- Ask men to move a 150 kg object
- Separate men that succeed and the ones that do not
- Find common characteristics between men that succeed and between men that do not

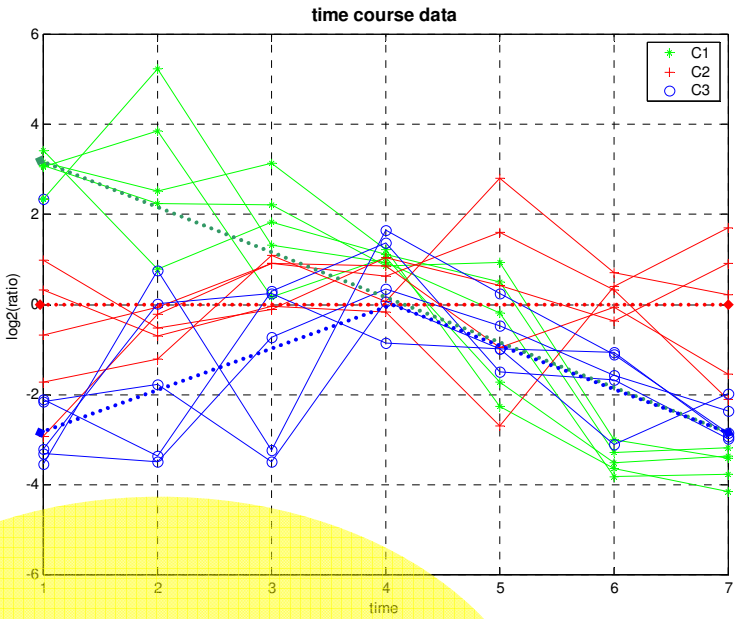
Genes differentially expressed

- Choose a set of genes
- Measure the expression of these genes on two different Biological states
- Choose subsets of genes that are enough to characterize each Biological state


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

Clustering

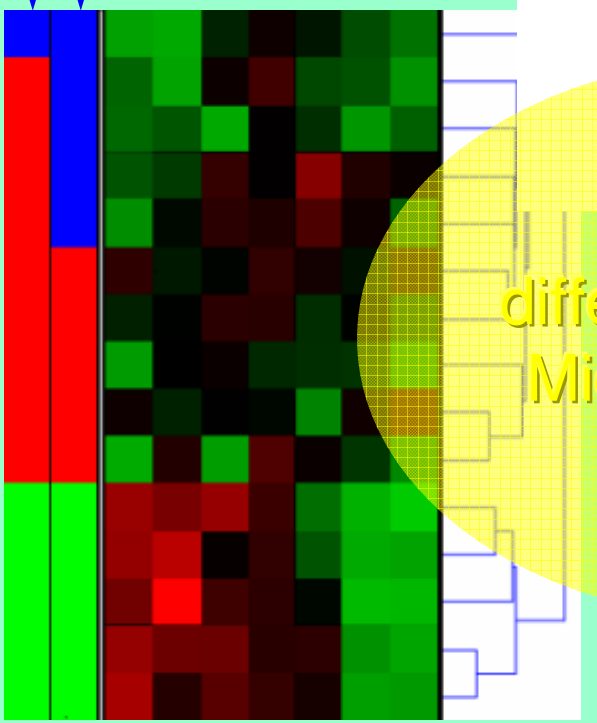
Time course data



Clustered by dendrogram

Original clusters

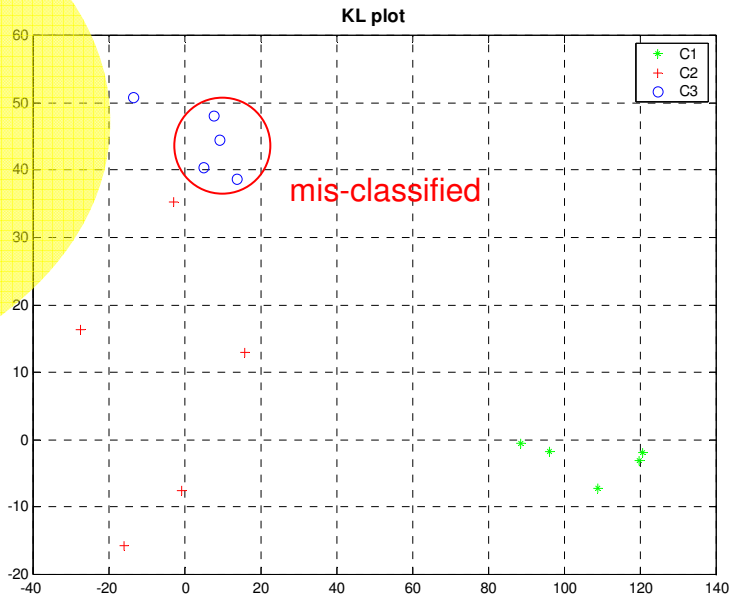
Dendrogram



different views of Microarray data

KL plot

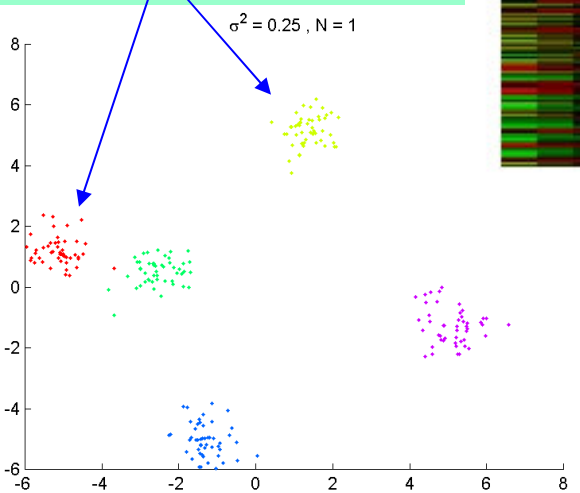
multidimensional space

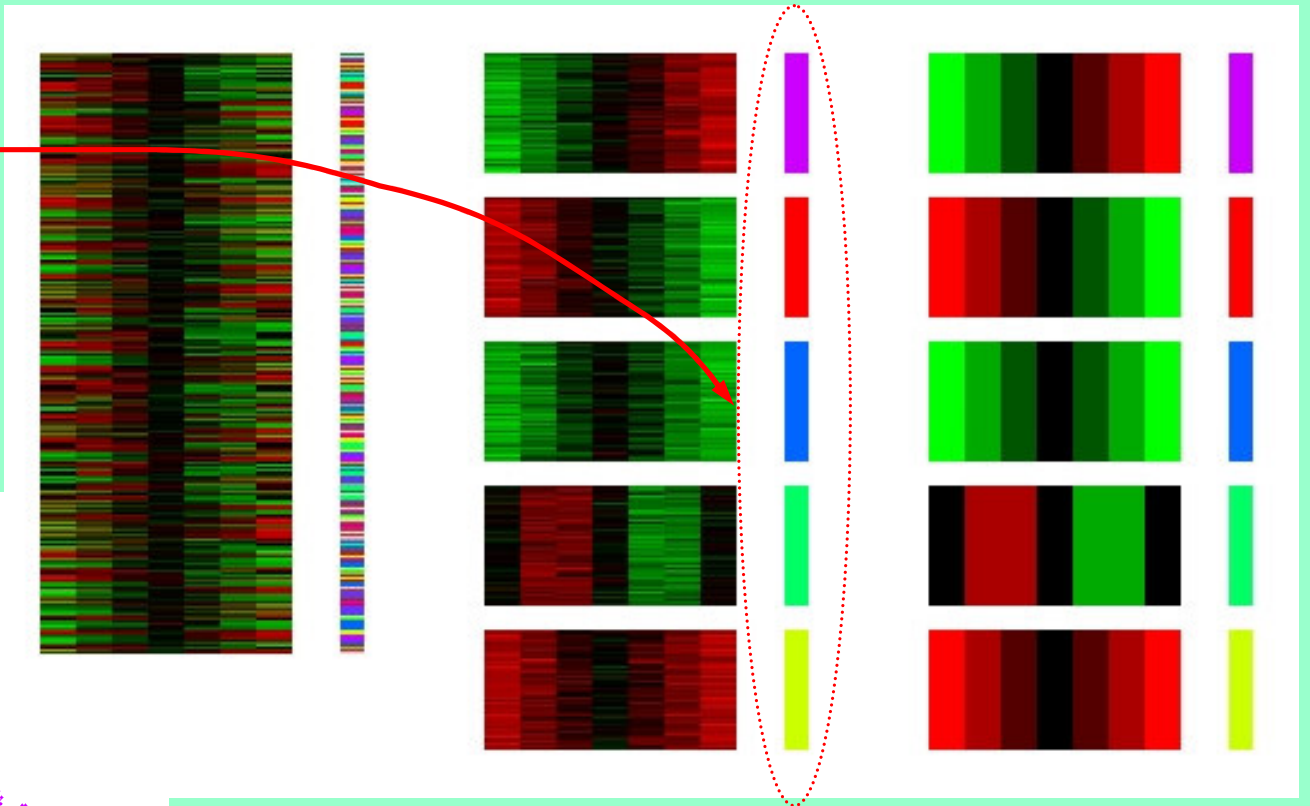


Example

No error!

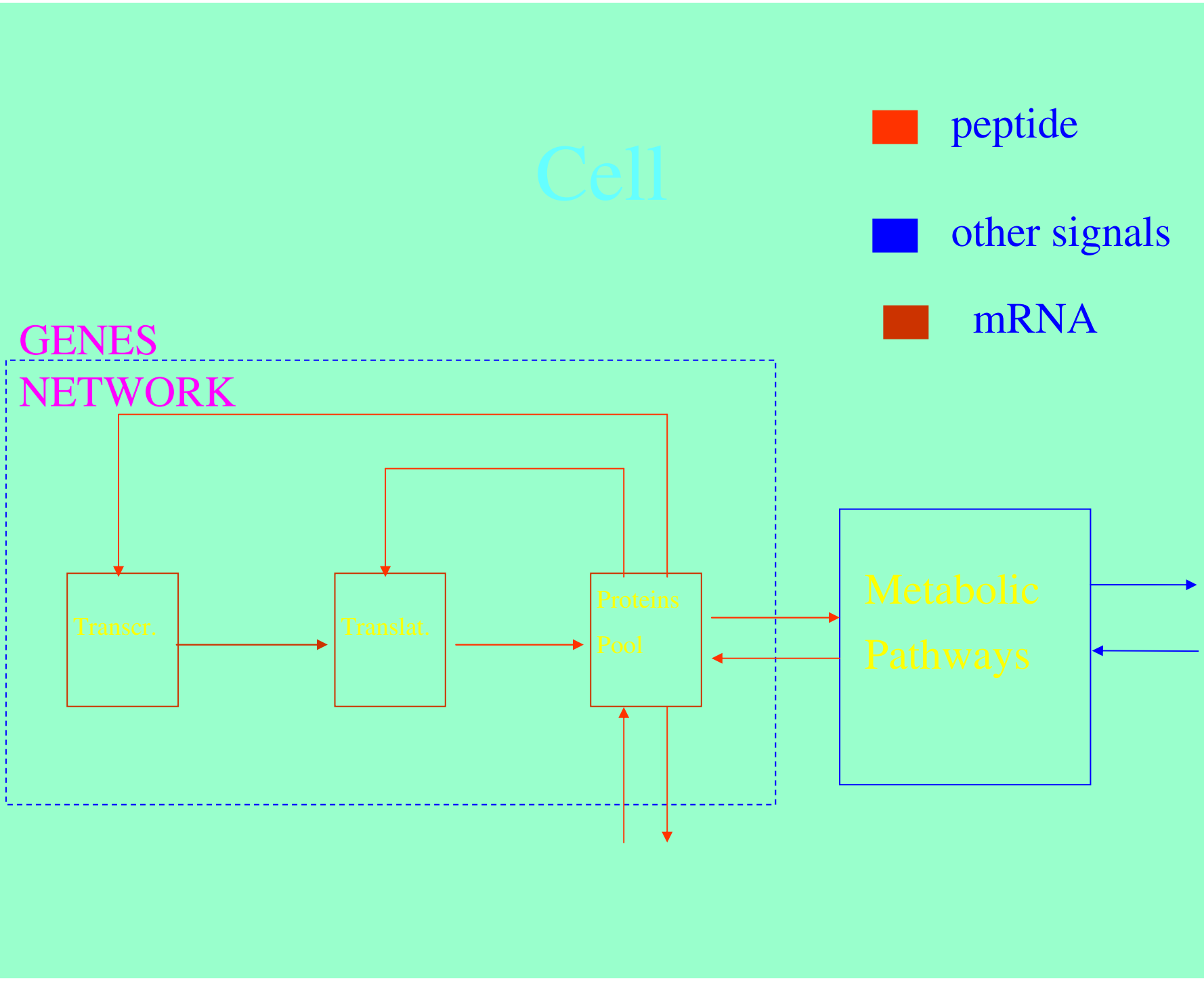
Tighter clusters due to small variance



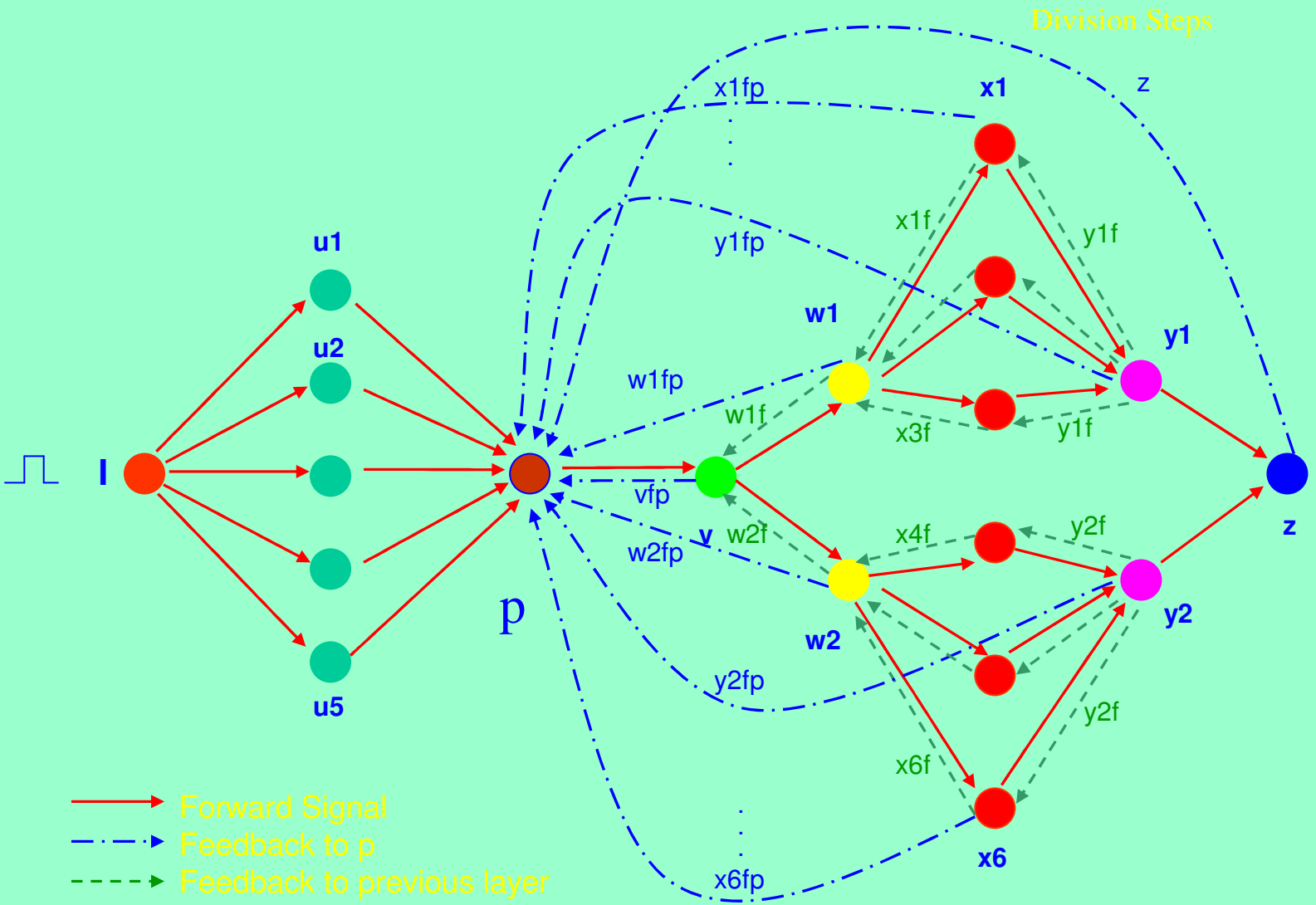


Results from Fuzzy c-means

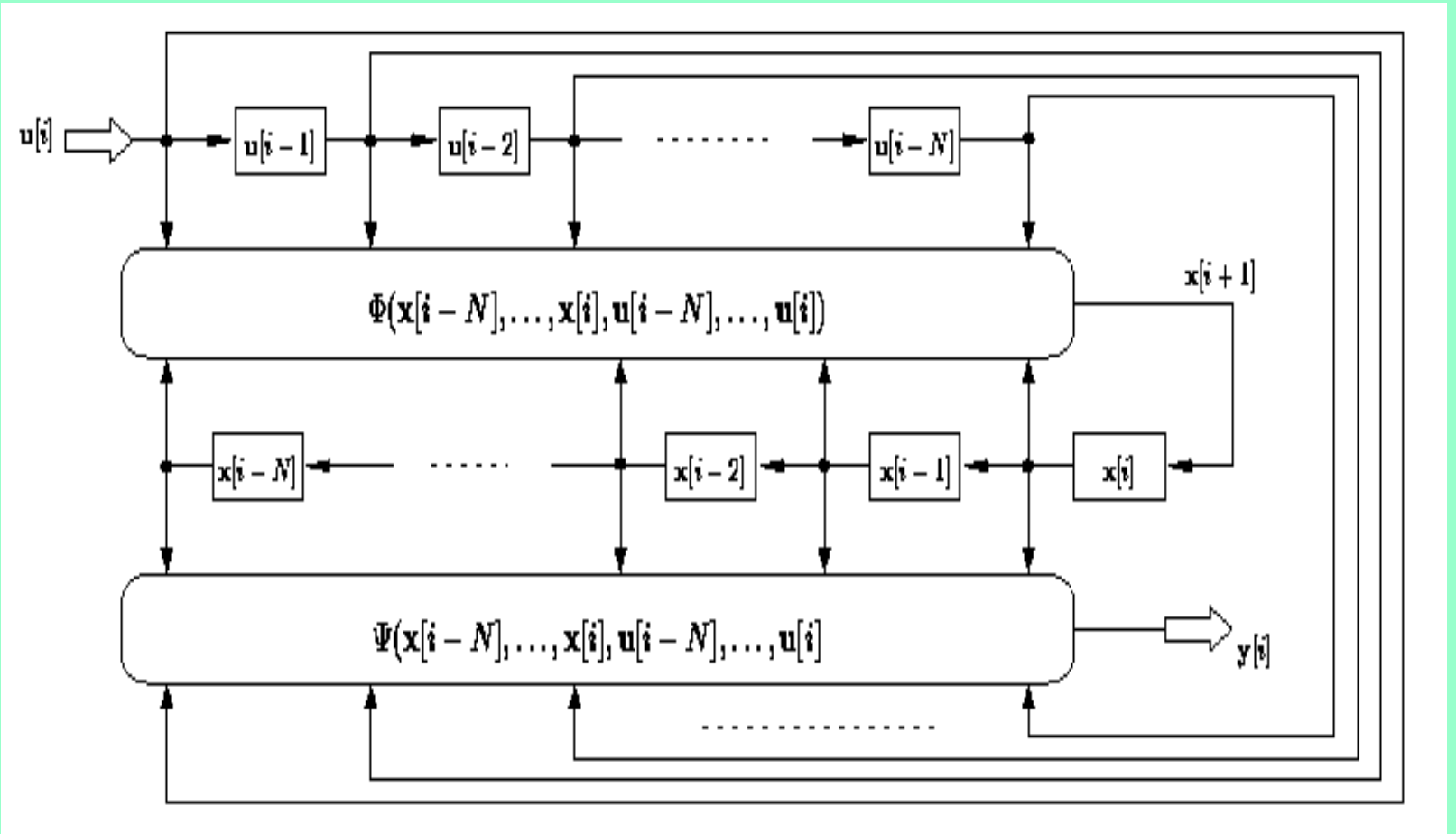
Gene Regulation Networks


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

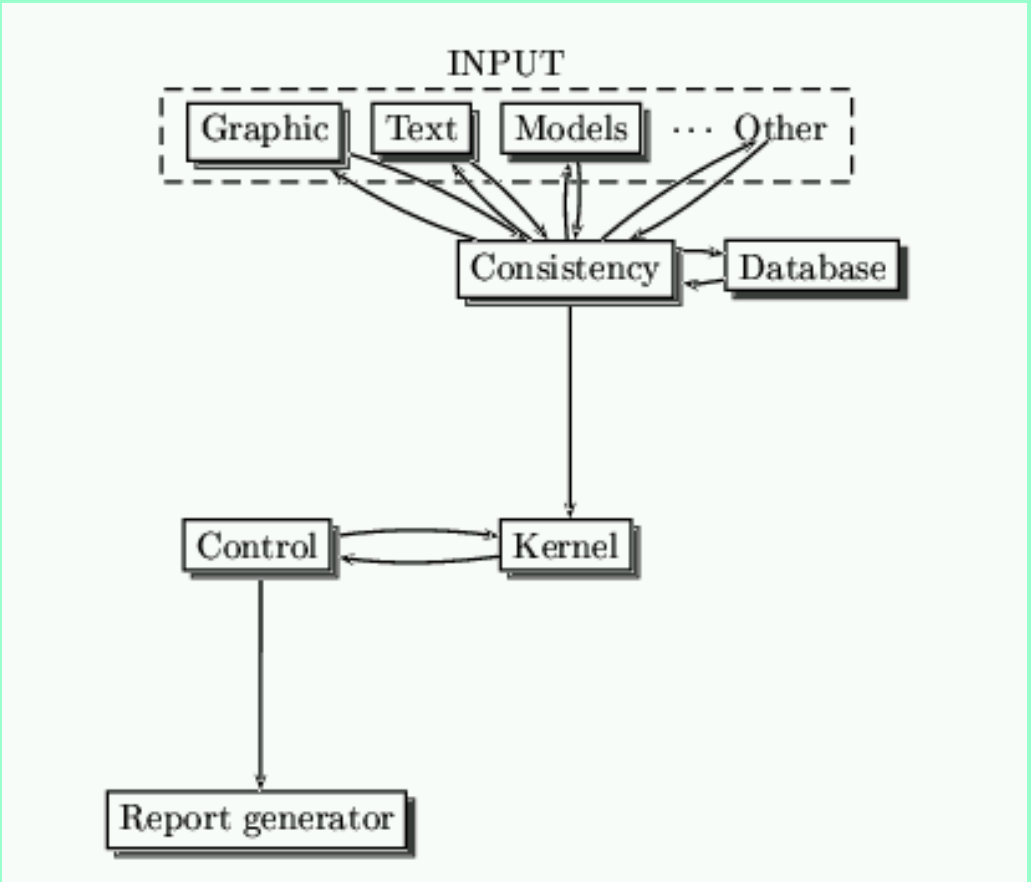
Cell Cycle Modeling

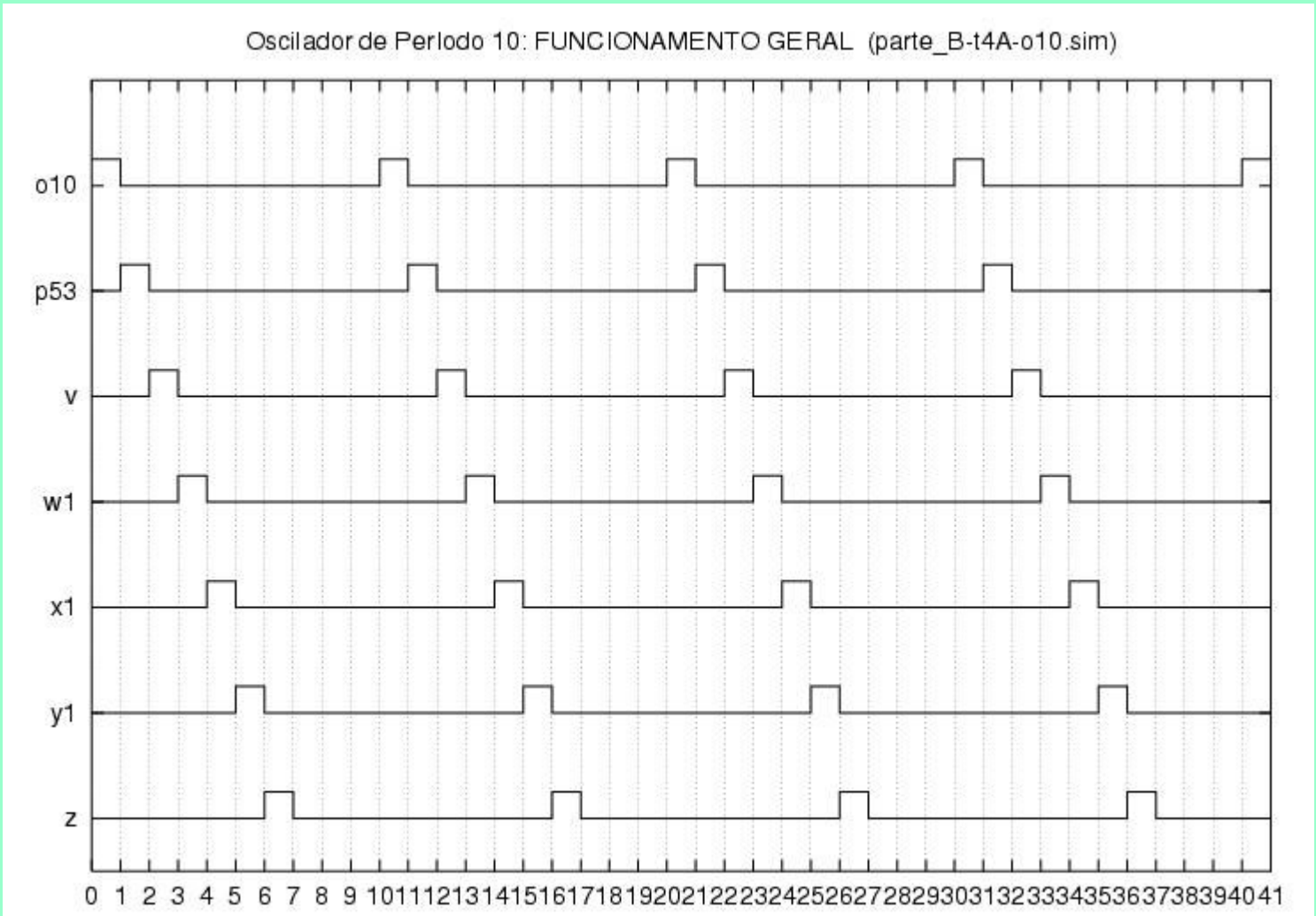


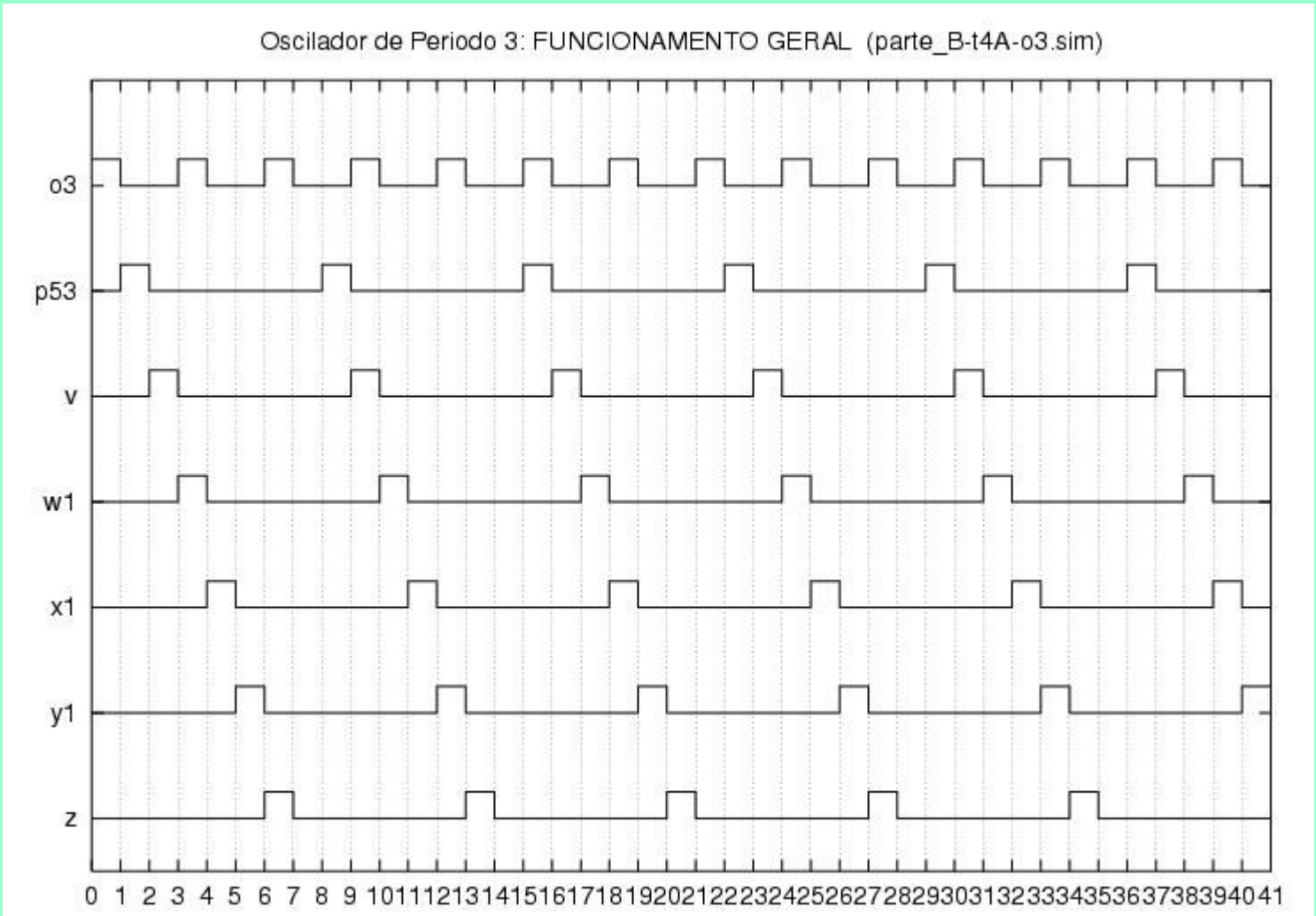
Modeling Dynamical Systems

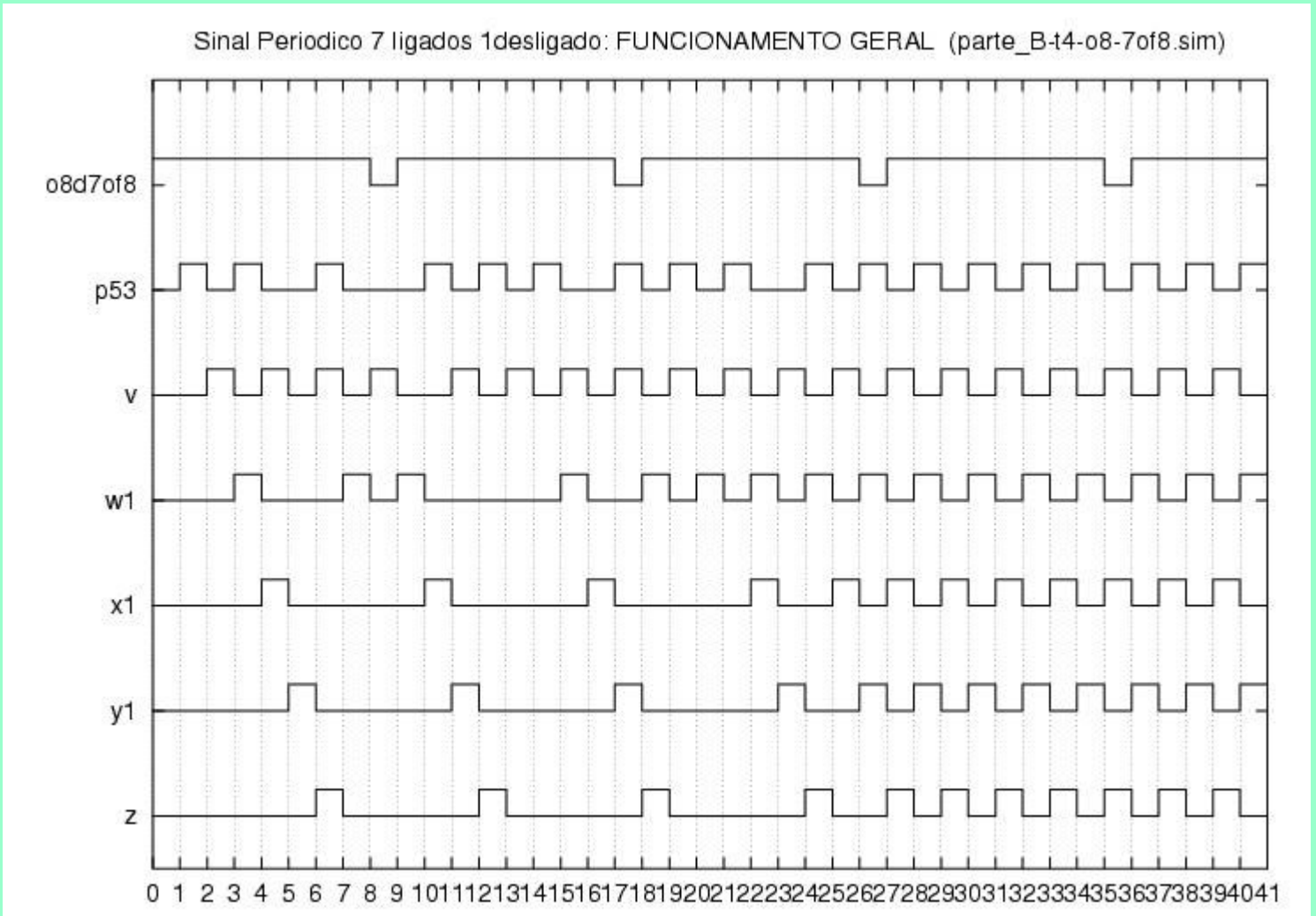


Simulator Architecture

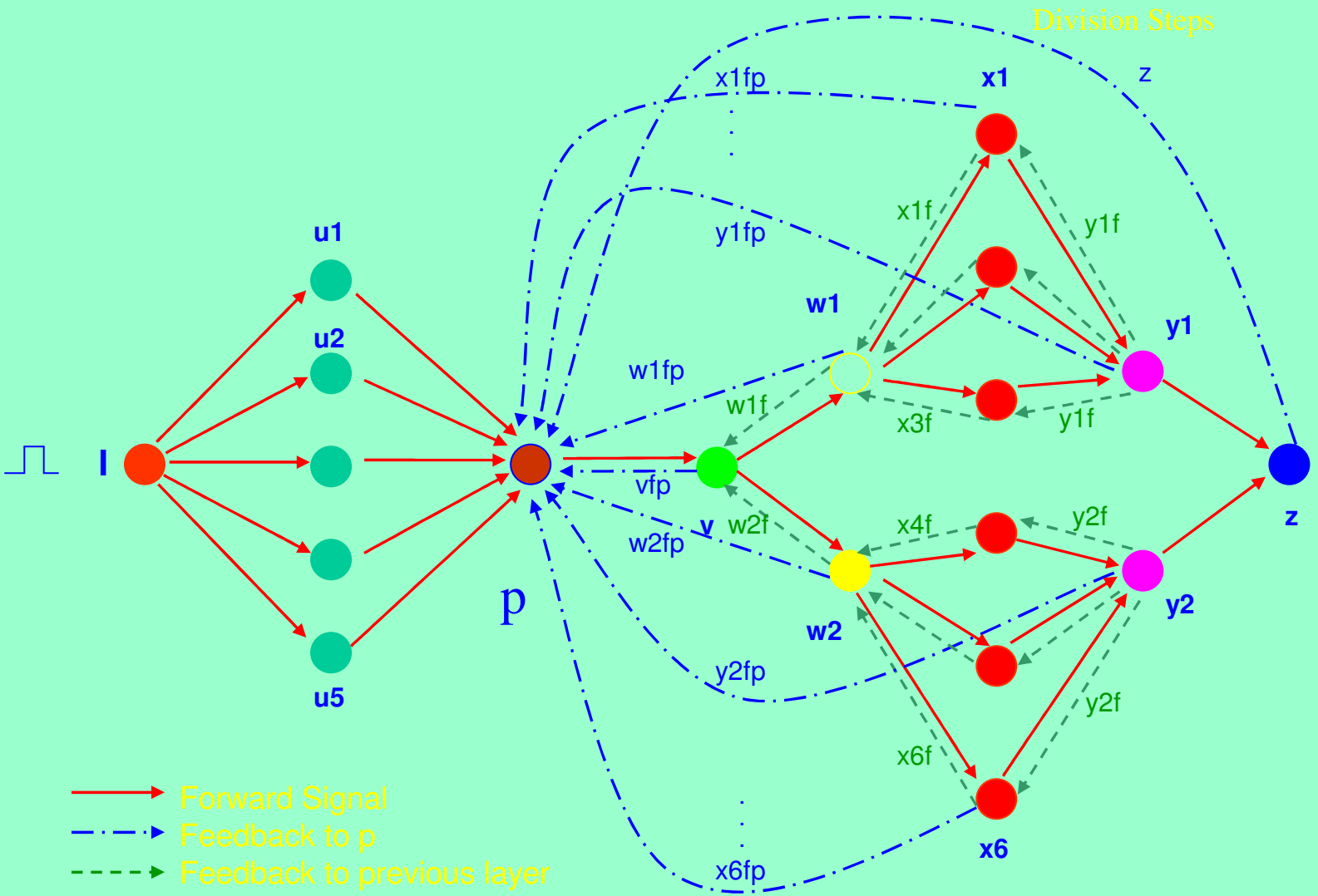


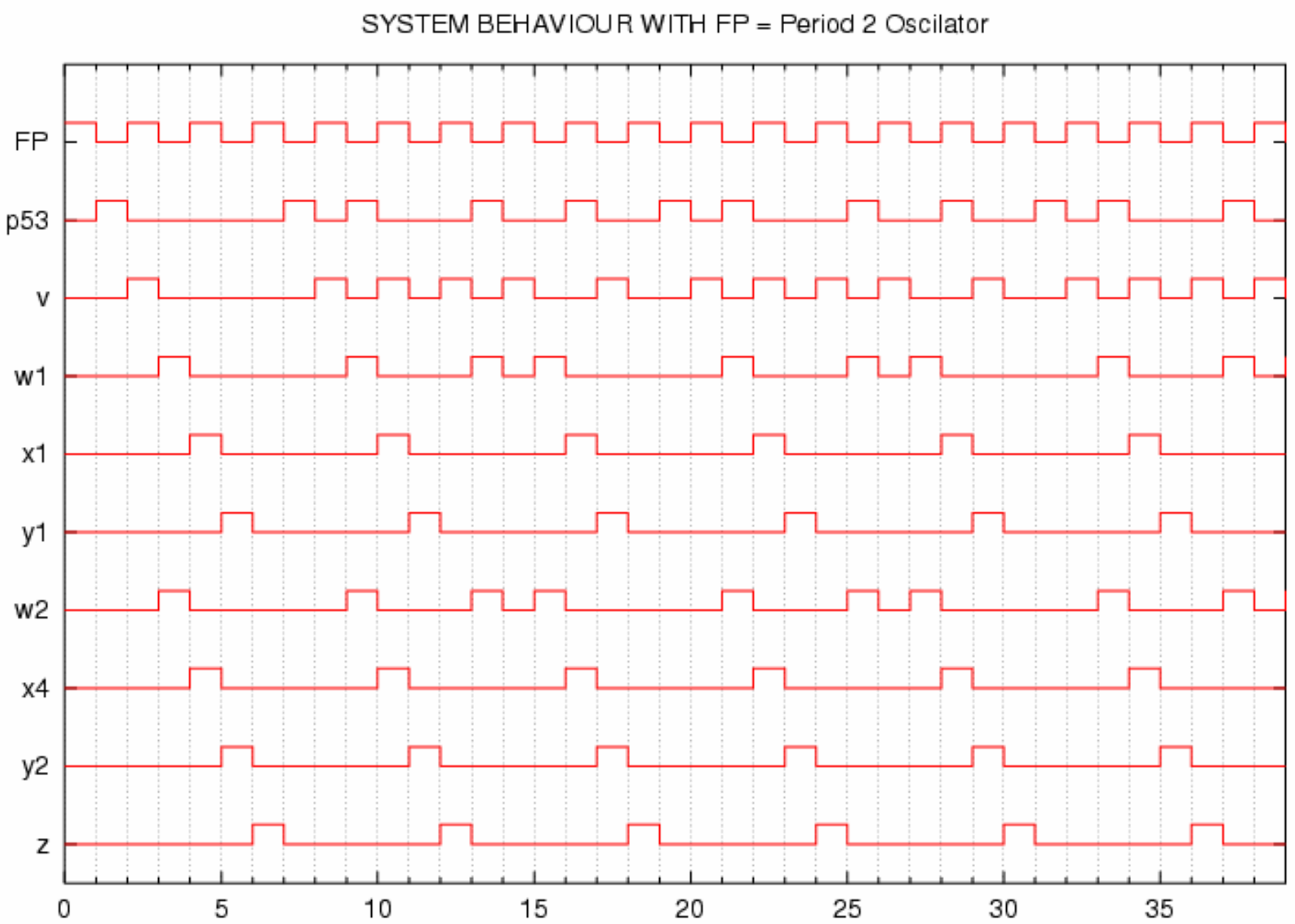

Oscilador de Perlodo 10: FUNCIONAMENTO GERAL (parte_B-t4A-o10.sim)
o10
p53
v
w1
x1
y1
z
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

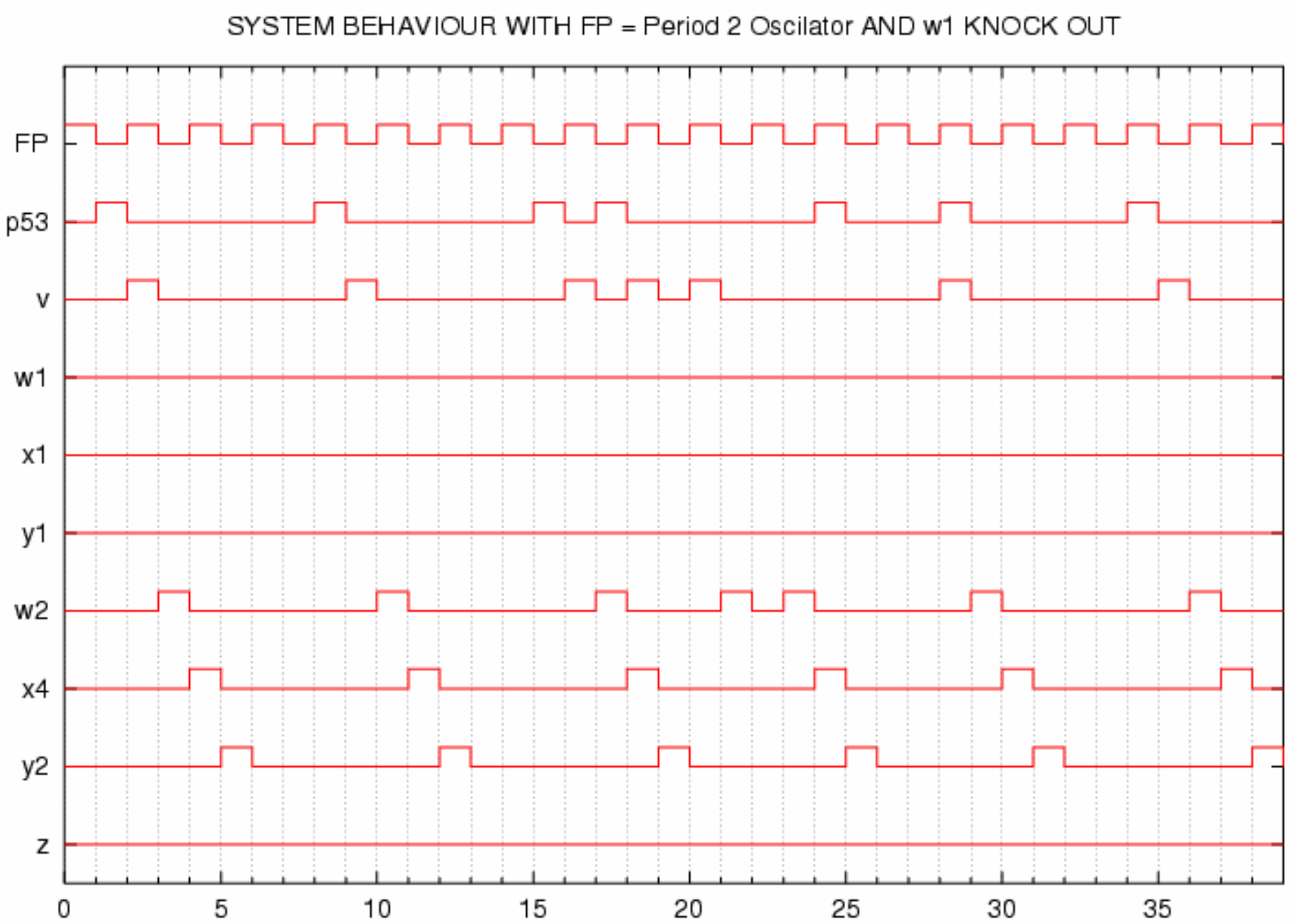

Oscilador de Periodo 3: FUNCIONAMENTO GERAL (parte_B-t4A-o3.sim)
o3
p53
v
w1
x1
y1
z
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41


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
x1fp
x1
z
x1f
y1f
y1fp
u1
u2
w1
y1
w1fp
w1f
x3f
y1f
I
vfp
v
w2f
x4f
y2f
z
w2fp
p
w2
y2
y2fp
y2f
x6f
u5
x6
x6fp
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

Challanges

- Architecture identification
- Dynamics transition function identification