Molecular Biology:

from sequence analysis to signal processing

Junior Barrera

University of Sao Paulo

Layout

- Introduction
- Knowledge evolution in Genetics
- Data acquisition
- Data Analysis
- A system for genetic data analysis
- Applications

Introduction

- Some medical signals are: EEG, ECG, ultra sound, tomography, etc.
- These signals are a great source of information about the human body
- For fully exploration of these data, Digital Signal Processing Techniques are necessary
- DSP: Algebra + Statistics + Computation

Introduction

- Techniques for the identification of genetic code are well known
- Soon the code of all genes will be known
- This knowledge open the way for one of the greatest challenges of science: the understanding of genes functionality

Introduction

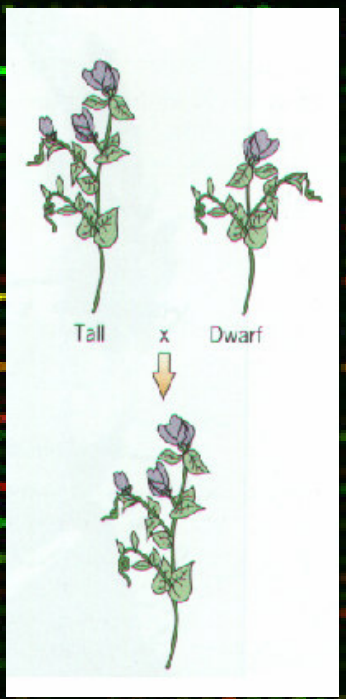
- Sets of genes constitute dynamical systems that control sequences of Biochemical reactions, called pathway
- The pathways in a cell define its activities
- States of large sets of genes can be observed by the microarray technology
- Gene states observed in time are digital signals

Introduction

- Gene states may describe properties of tissues. Pattern Recognition techniques permits to predict tissue properties. Example: cancer classification
- System Identification techniques permit to estimate net architectures and dynamics. Example: control of cell division

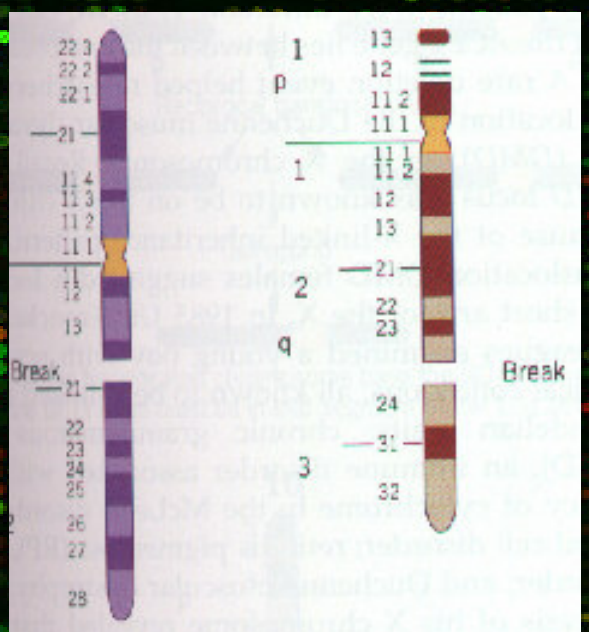
Knowledge evolution in genetics

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



Knowledge evolution in genetics

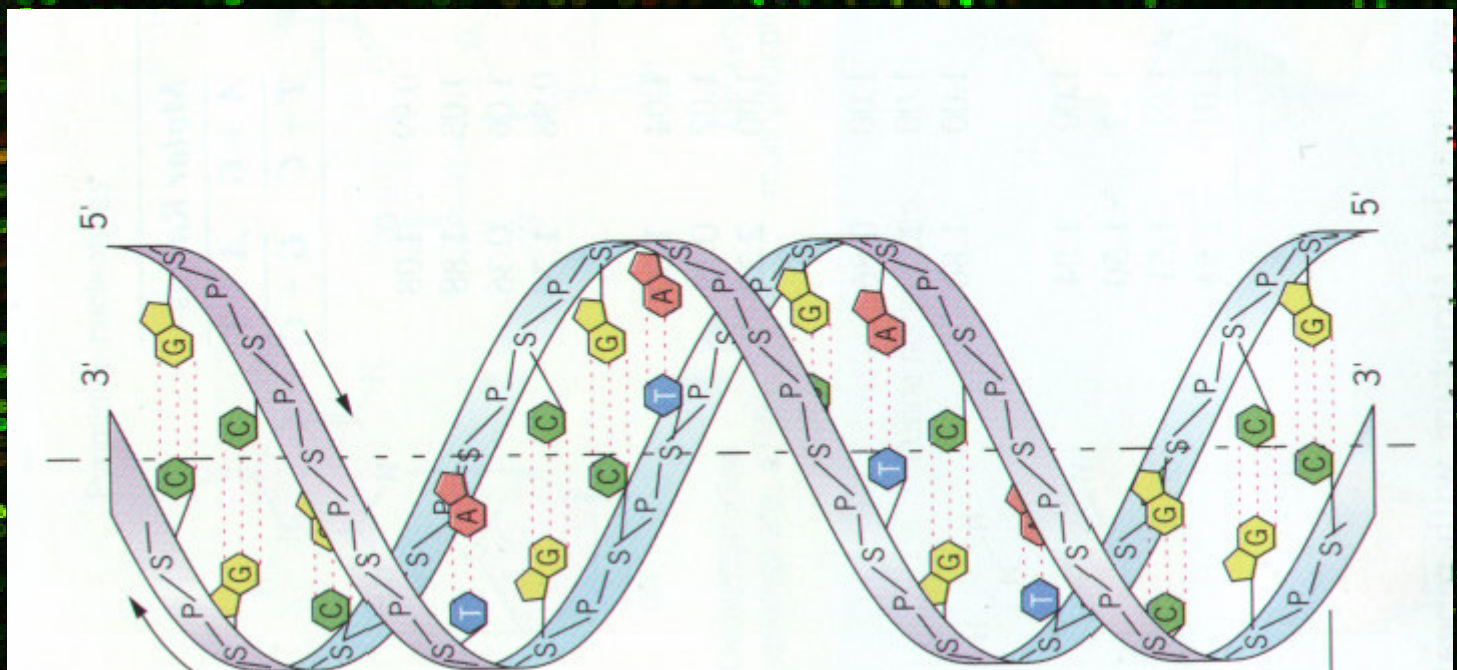
- Chromosome Theory - Morgan (1910)
- Genes were situated in chromosomes



Introduction

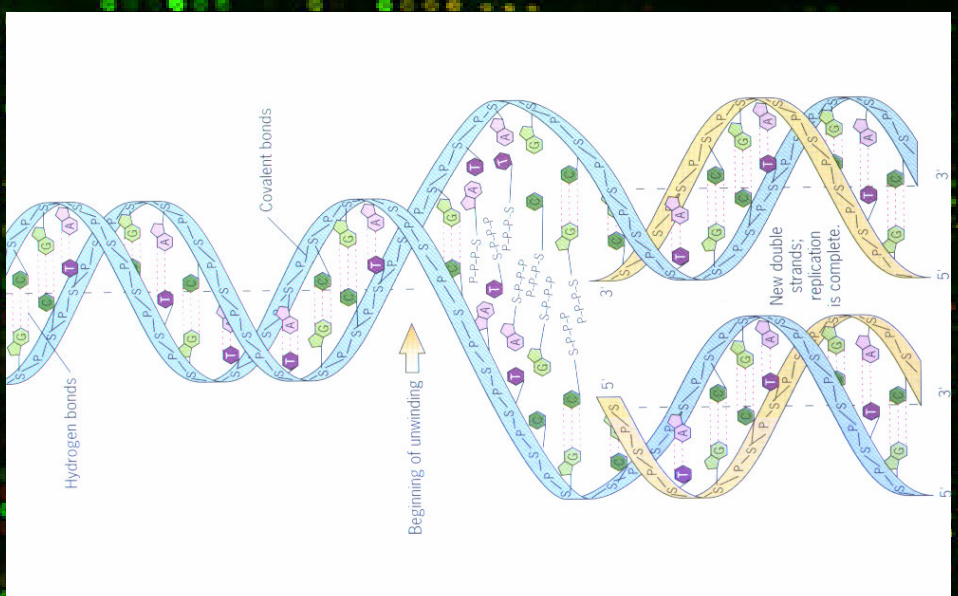
- The molecular structure of chromosomes (Watson and Crick - 1953)
- DNA structure: the double helix
- Four basis: adenine(A), guanine(G), thymine(T), cytosine(C)
- genes are sequences of nucleotides

Introduction



Introduction

- DNA manipulation
- cut, replication and decoding



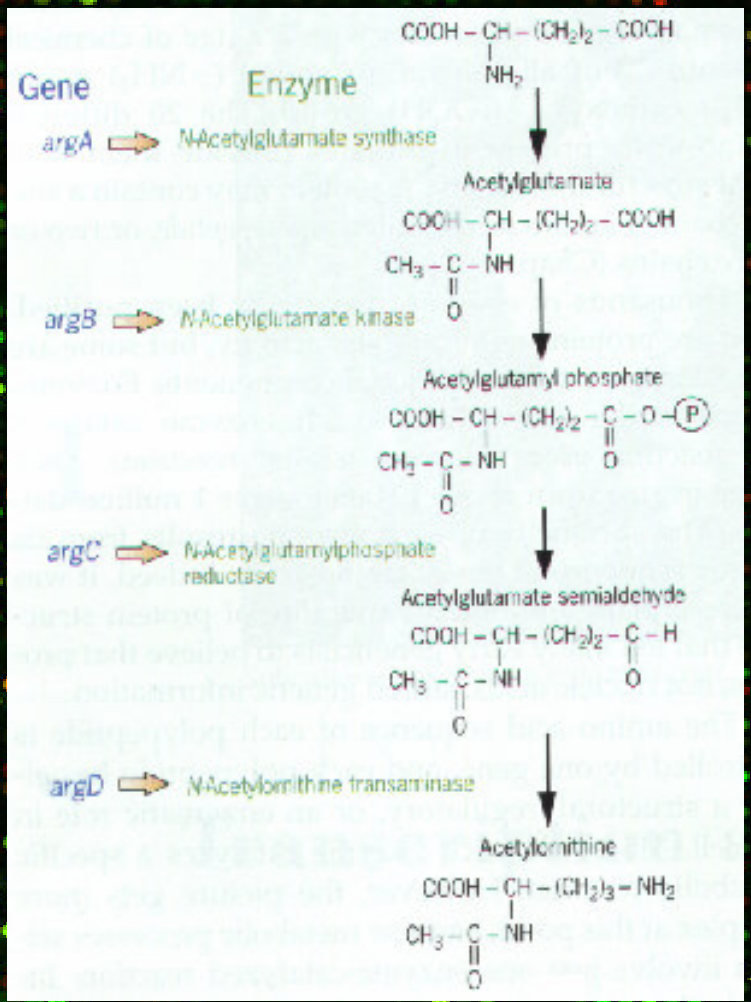
Introduction

- Genetic engineering
- species modification, drug production

Introduction

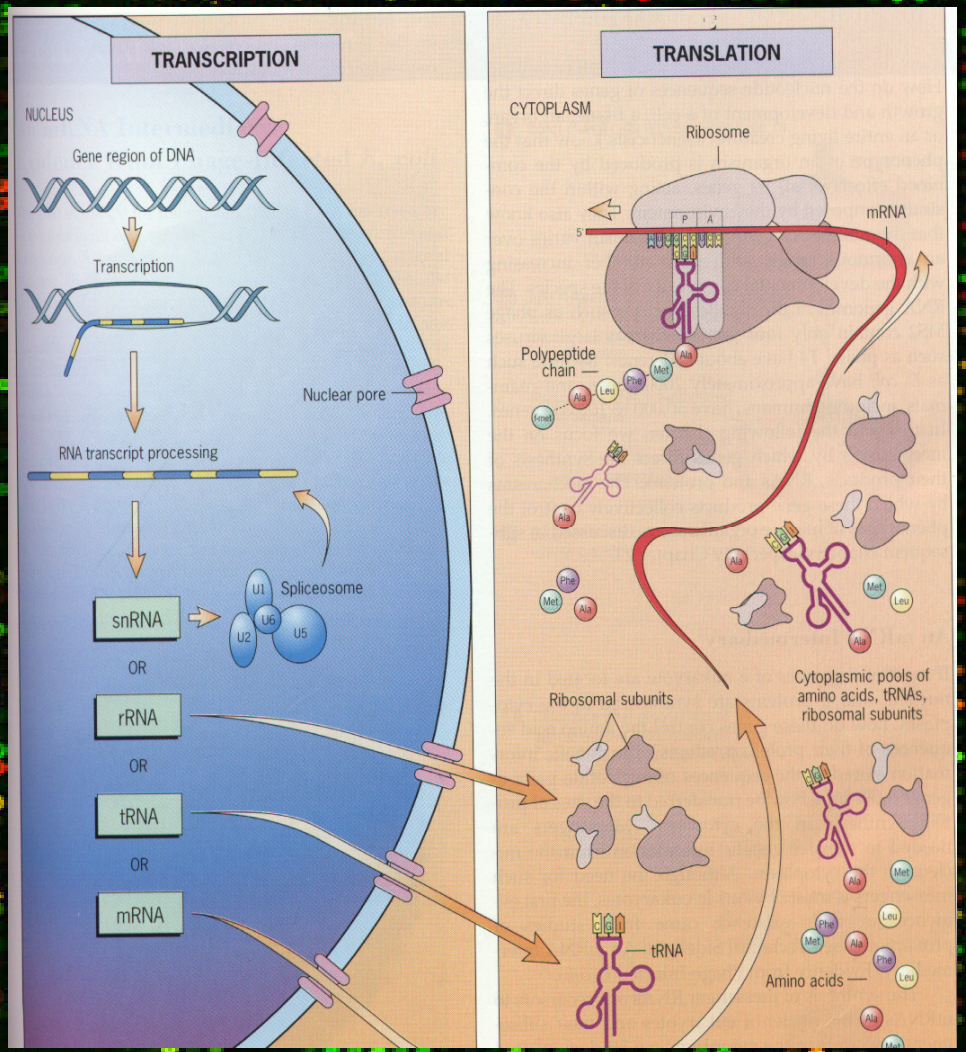
- Genes control the metabolism
- Metabolism occurs by sequences of enzyme-catalyzed reactions.
- Enzymes are specified by one or more genes

Introduction

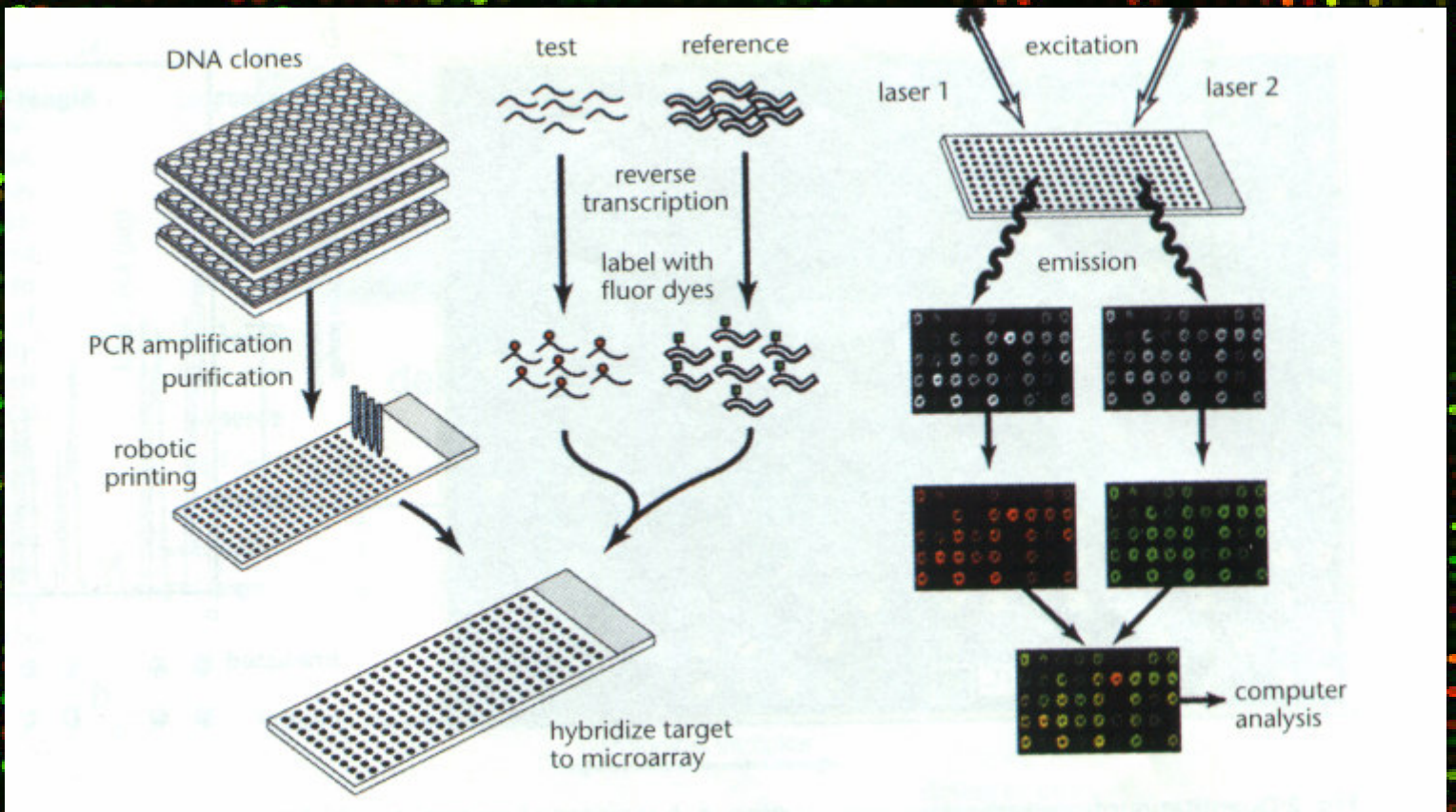


Introduction

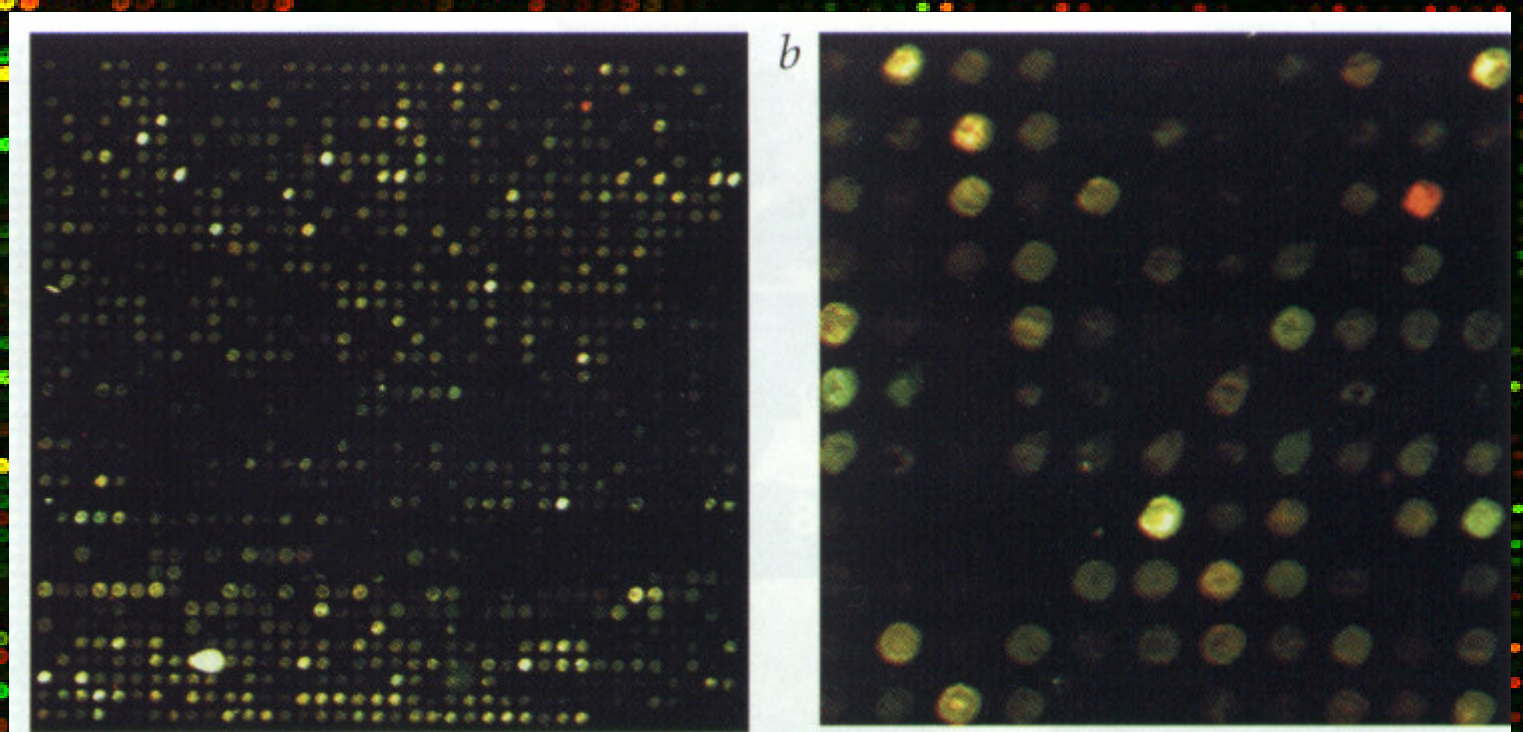
- Gene expression



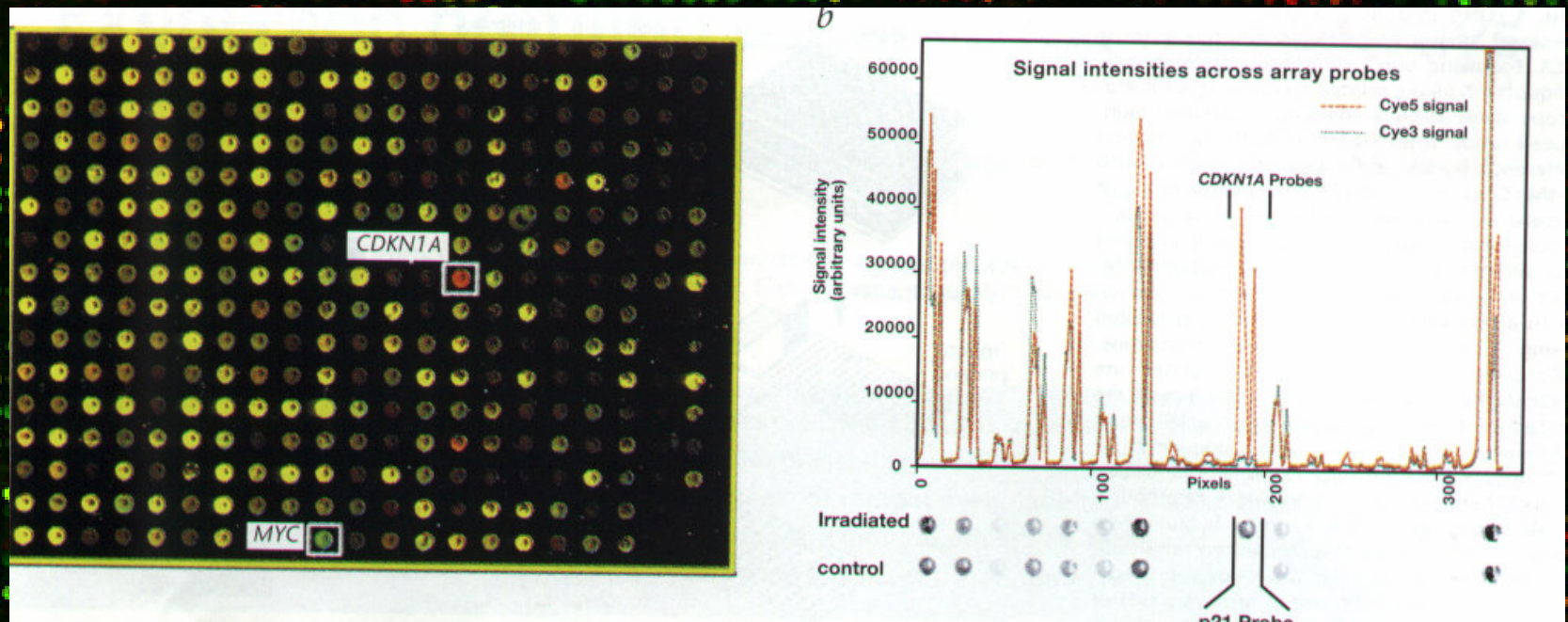
Data acquisition



Data acquisition



Data acquisition

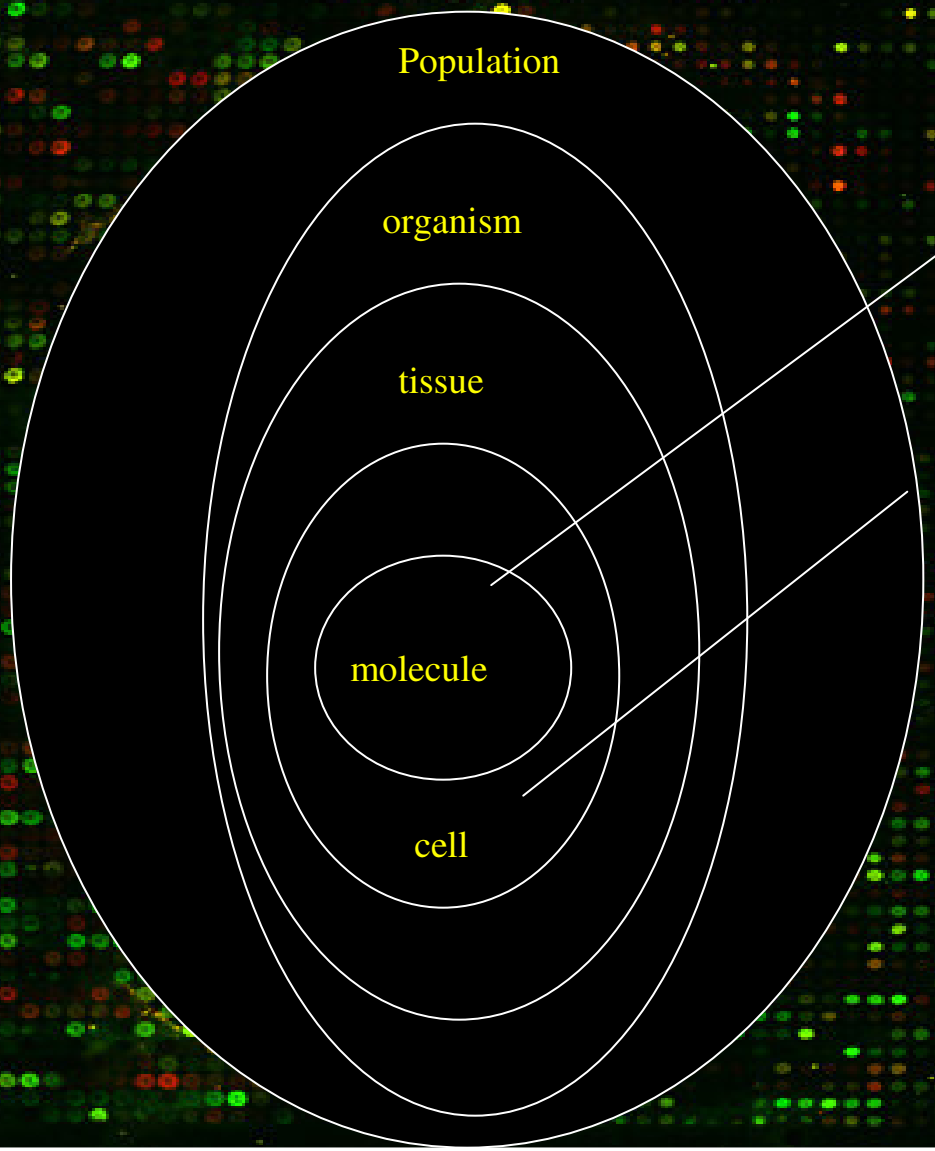


Quantization - {-1,0,1}

Data Analysis

- Data classes definition
- Relational search
- Data transformation
- Mining
- Integration of information
- Interpretations

Data classes definition



Protein structure
and dynamics

DNA, Protein,
Gene Expression,
Gene Networks

Relational Search

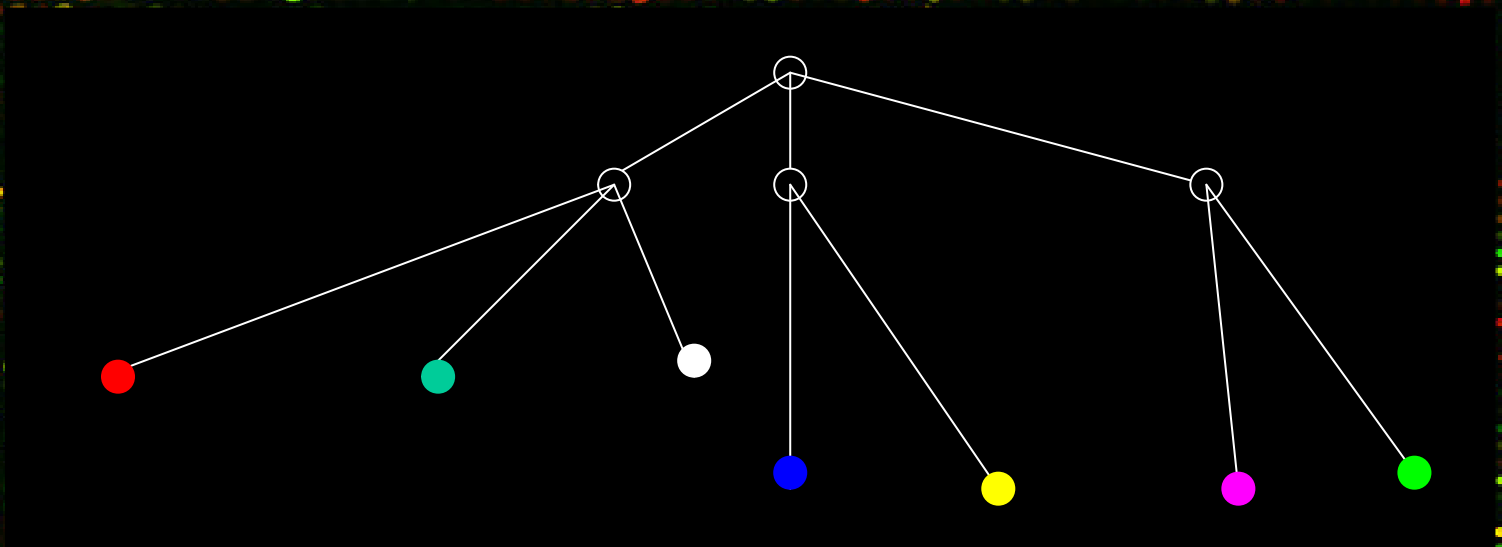
- Get a subset of the available data
- Define relations between categories of data
- Select by logical operations on relations

Data Transformations

- Image analysis
- Measures on DNA sequences
- Measure on Protein sequences
- …

Mining

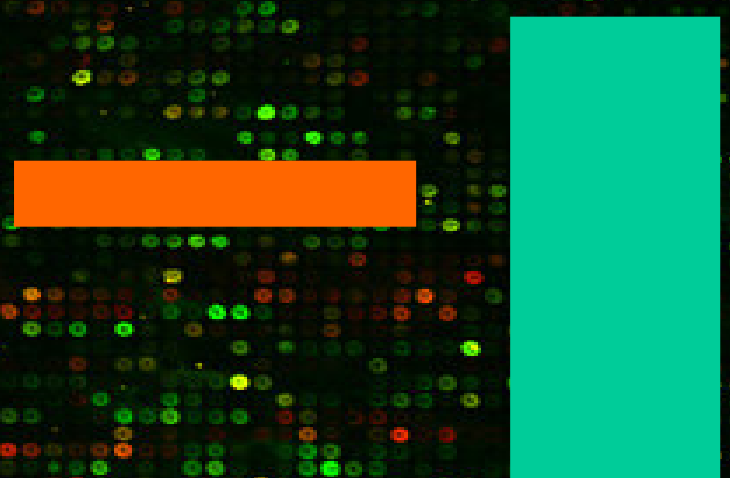
Classifier

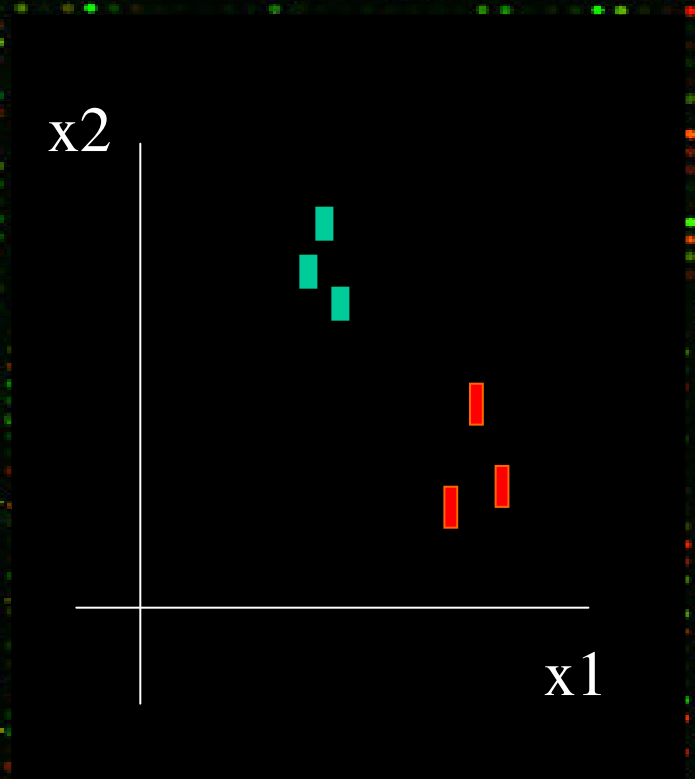


Examples : DNA assembling, Protein and DNA homology, DNA philogeny, genes characterizations of tissue, time pattern similarity

Mining

- Attribute Space

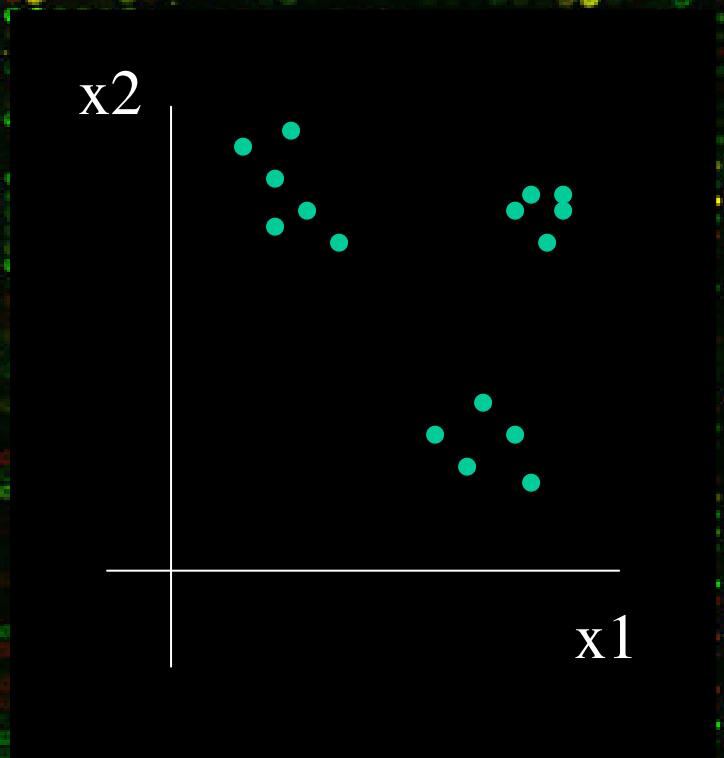


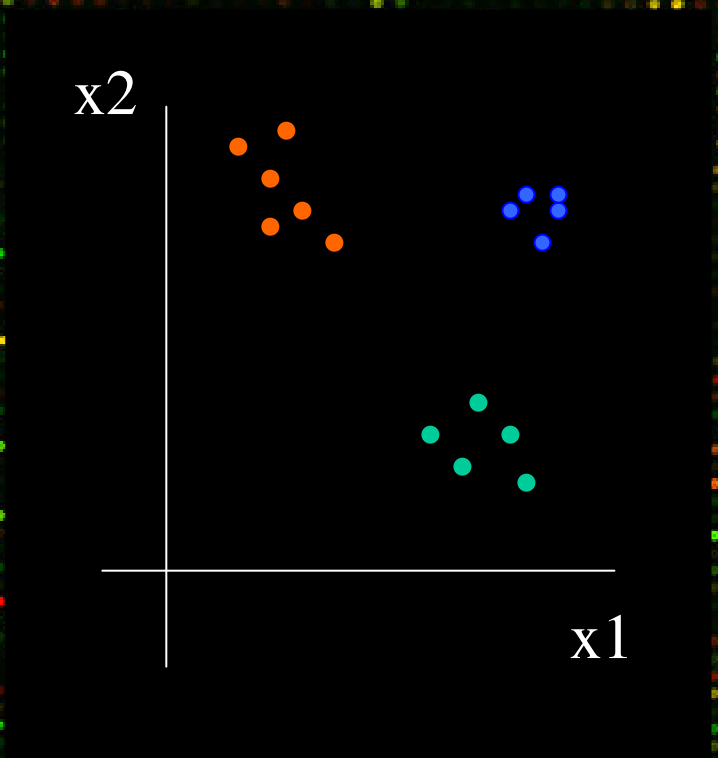


x=(x1,x2)

Mining

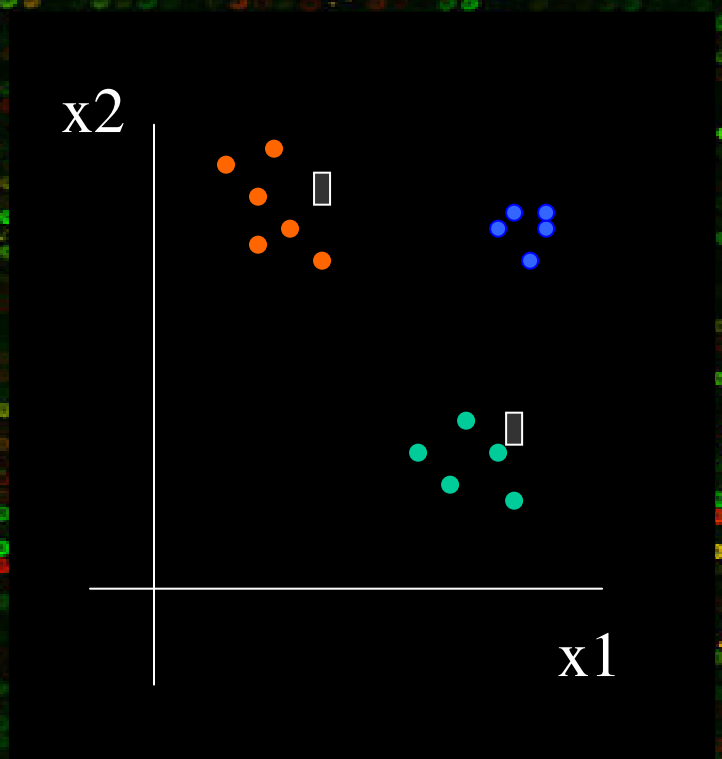
- Clustering

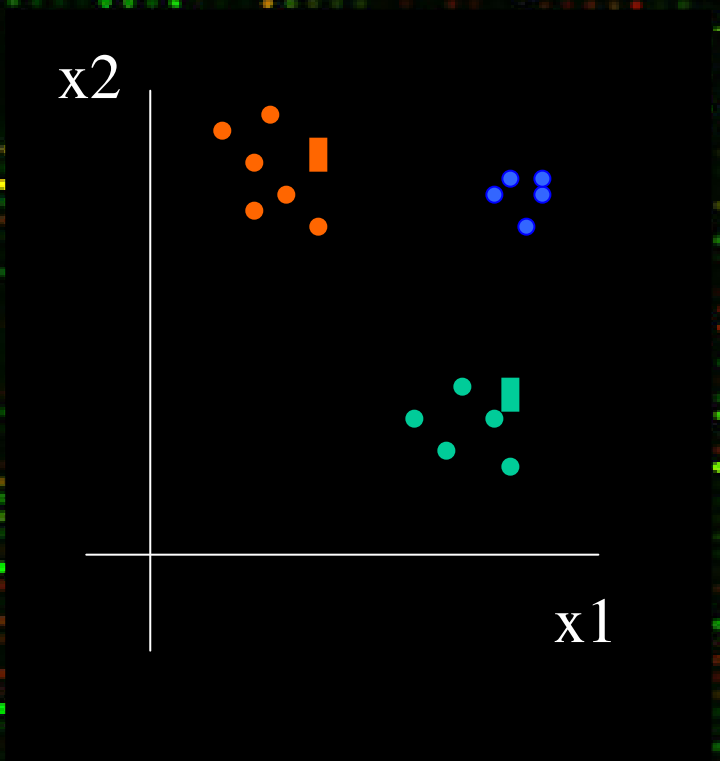




Mining

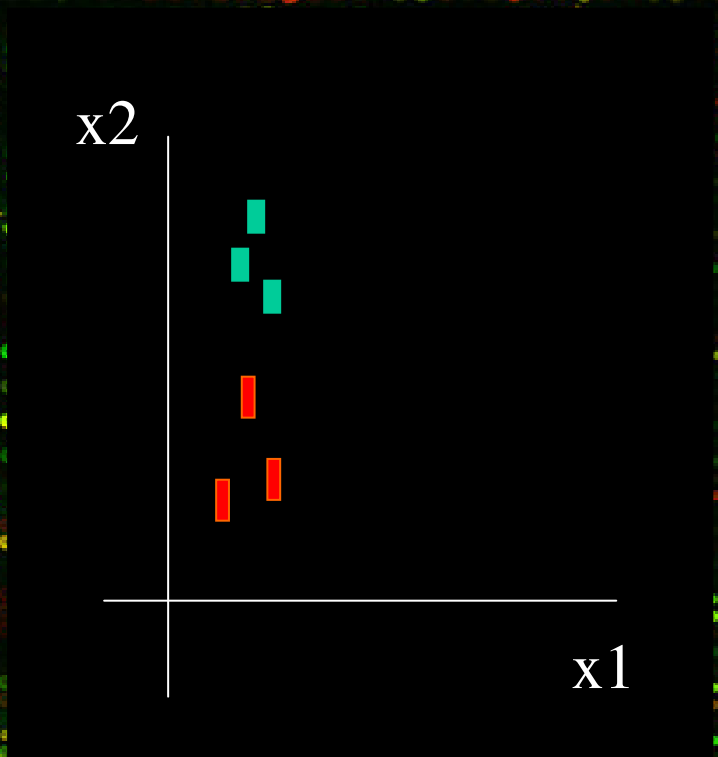
- Classification





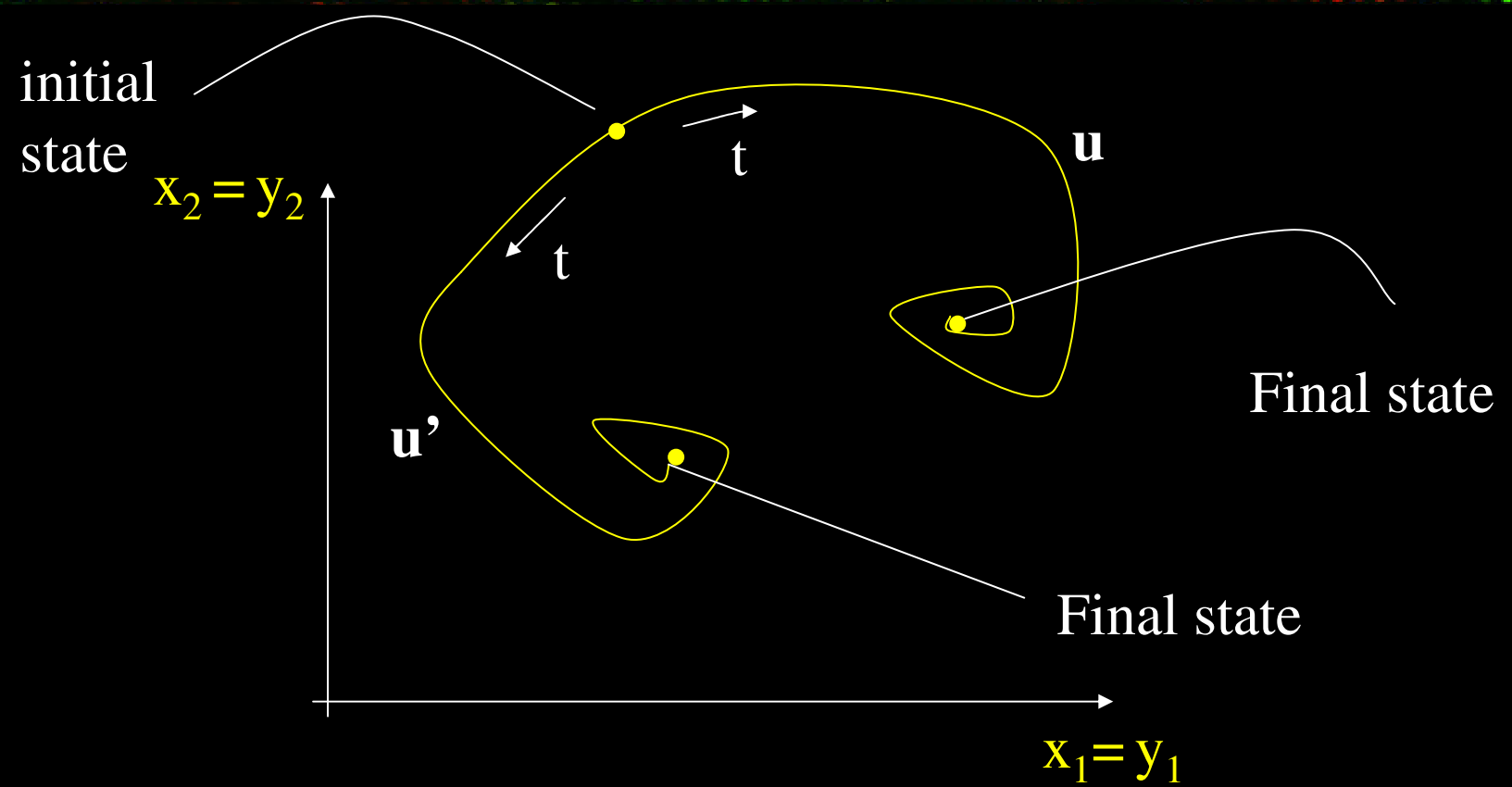
Mining

- Attribute Space Dimension



Mining

Dynamical System



Examples : gene networks, protein structure, cells, organisms, populations, drug reaction

Mining

$$\mathbf{x} : T \to \mathcal{L}^n \qquad \mathbf{y} : T \to \mathcal{L}^m \qquad \mathbf{x}[t] \in \mathcal{L}^n$$

$$\mathbf{u} : T \to \mathcal{L}^n$$

$$\mathbf{x}[t+1] = \Phi_t(\mathbf{x}[t-N], ..., \mathbf{x}[t], ..., \mathbf{x}[t+N], \mathbf{u}[t-N], ..., \mathbf{u}[t], ..., \mathbf{u}[t+N])$$

$$\mathbf{y}[t] = \Psi_t(\mathbf{x}[t-N], ..., \mathbf{x}[t], ..., \mathbf{x}[t+N], \mathbf{u}[t-N], ..., \mathbf{u}[t], ..., \mathbf{u}[t+N])$$

$$S(\Phi_t, \Psi_t)$$

$$\Phi_t : \mathcal{L}^{2(2N+1)n} \to \mathcal{L}^n$$

$$\Psi_t : \mathcal{L}^{2(2N+1)n} \to \mathcal{L}^m$$

Integration of Information

- different resolutions data classes
- transformed data
- selected data
- mined data

Interpretation

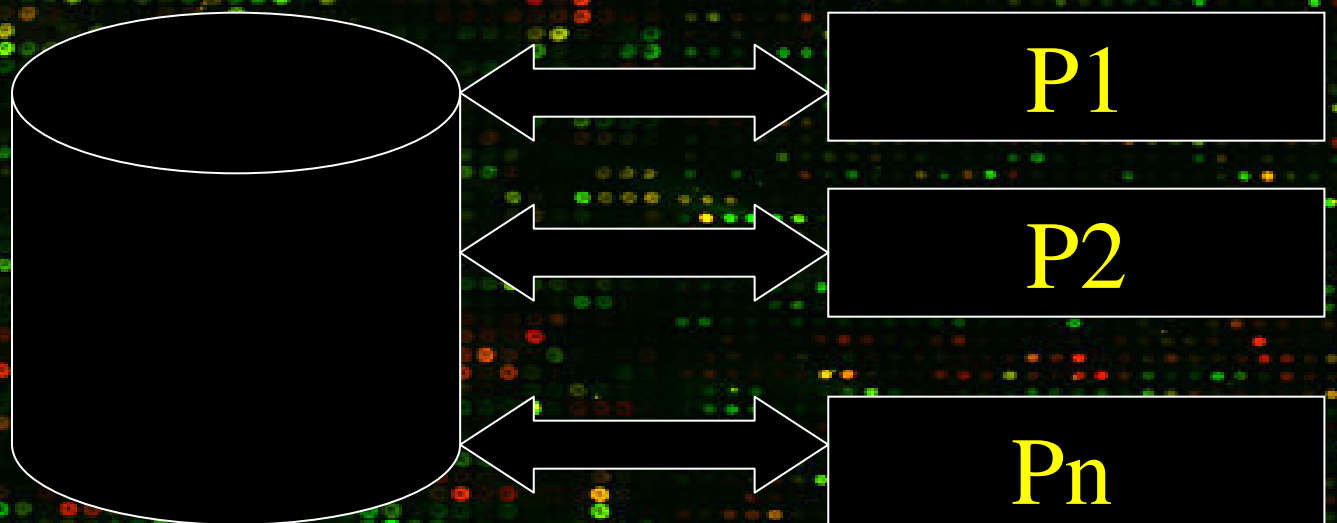
- Integrated information
- Known concepts
- Propose hypothesis
- confirm or negate hypothesis

A system for genetic data analysis

- Database
- Analytical procedures
- Data mining
- High performance computing

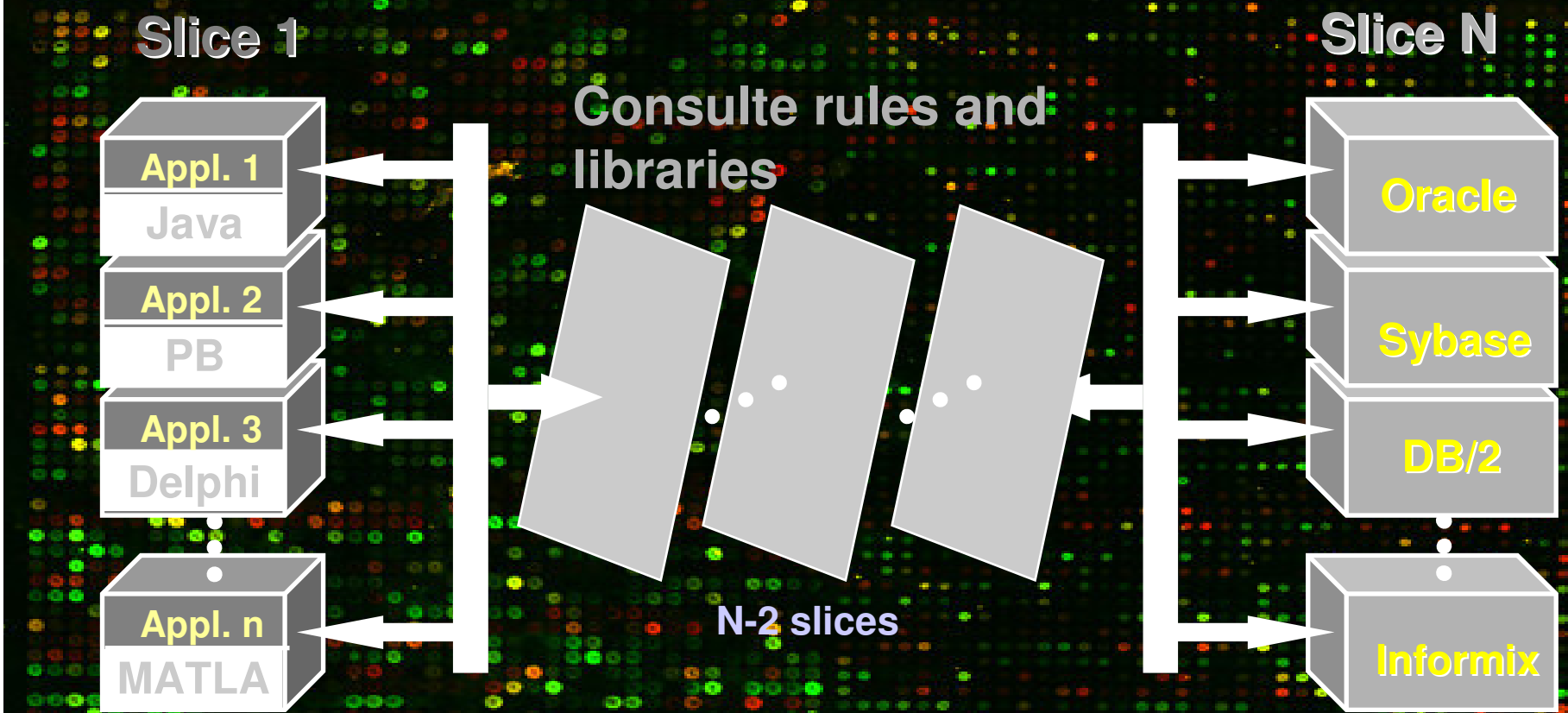
System

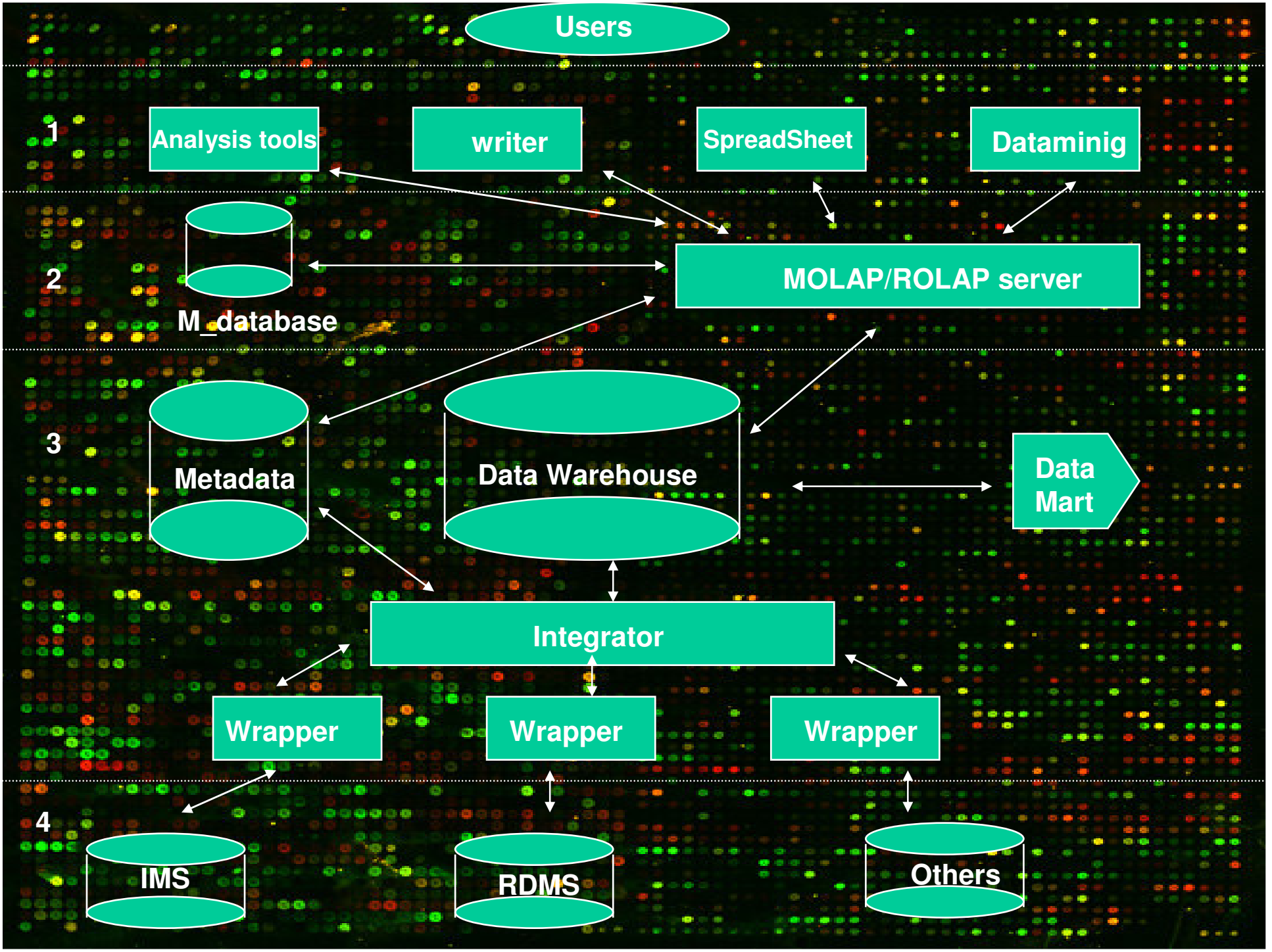
Objected oriented database



Pi : analytical and mining procedures (kernel parallel)

System Architecture



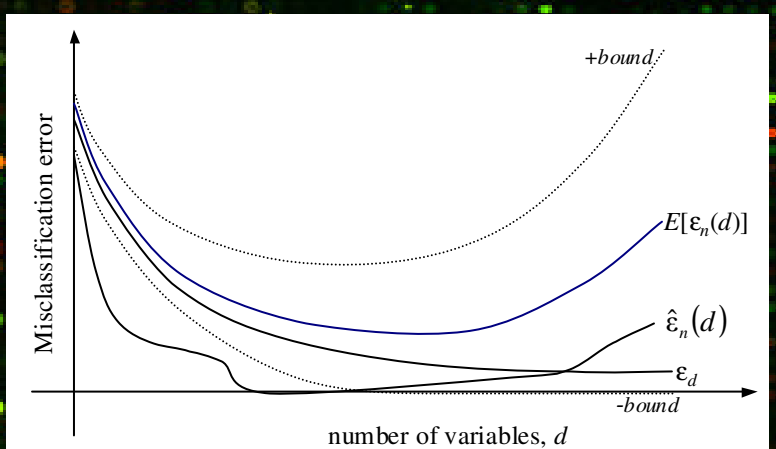

Users
1
Analysis tools
writer
SpreadSheet
Dataminig
2
M_database
MOLAP/ROLAP server
3
Metadata
Data Warehouse
Data Mart
Integrator
Wrapper
Wrapper
Wrapper
4
IMS
RDMS
Others

Applications

- Cancer tissue characterization
- Cell cycle simulation
- Inference from clustering
- Gene regulation

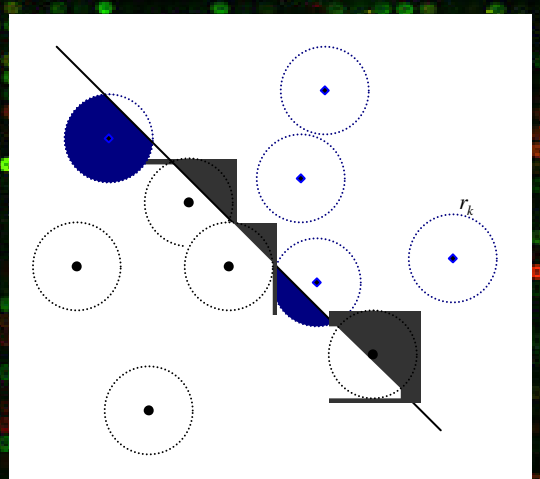
Cancer tissue characterization

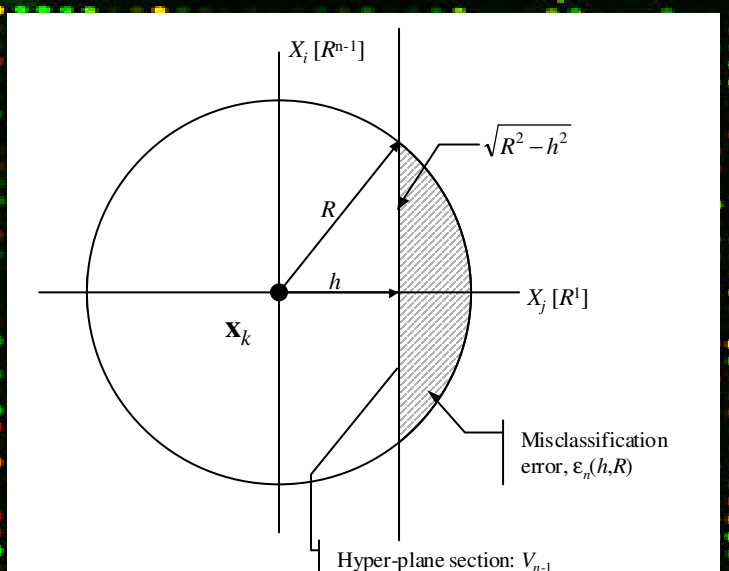
- Problem: from a small set (20) of microarrays, find a minimum number of genes that are enough to separate cancer A and B.



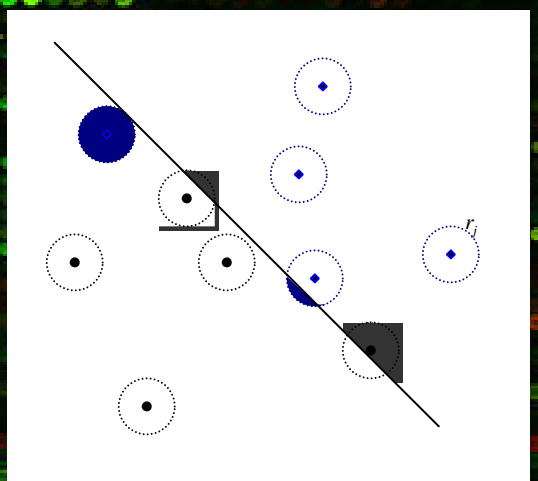
- Approach: randomize data, compute classifier using genes subsets, measure error for different dispersions, choose the subset that balance small error and high dispersion. A supercomputer is required.

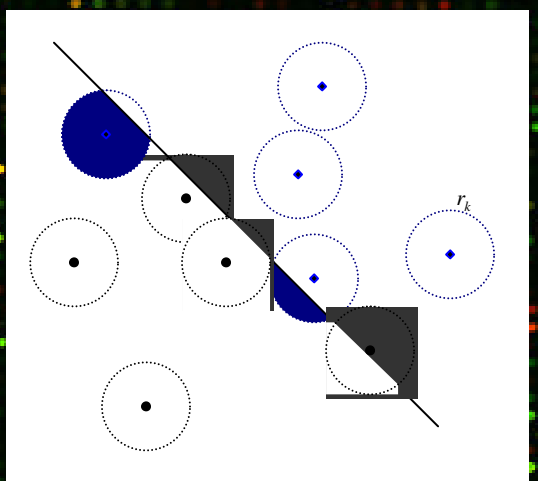
- Linear classifier
- Dispersion centered in the sample
- Flat round dispersion model
- Error computed analytically (faster)

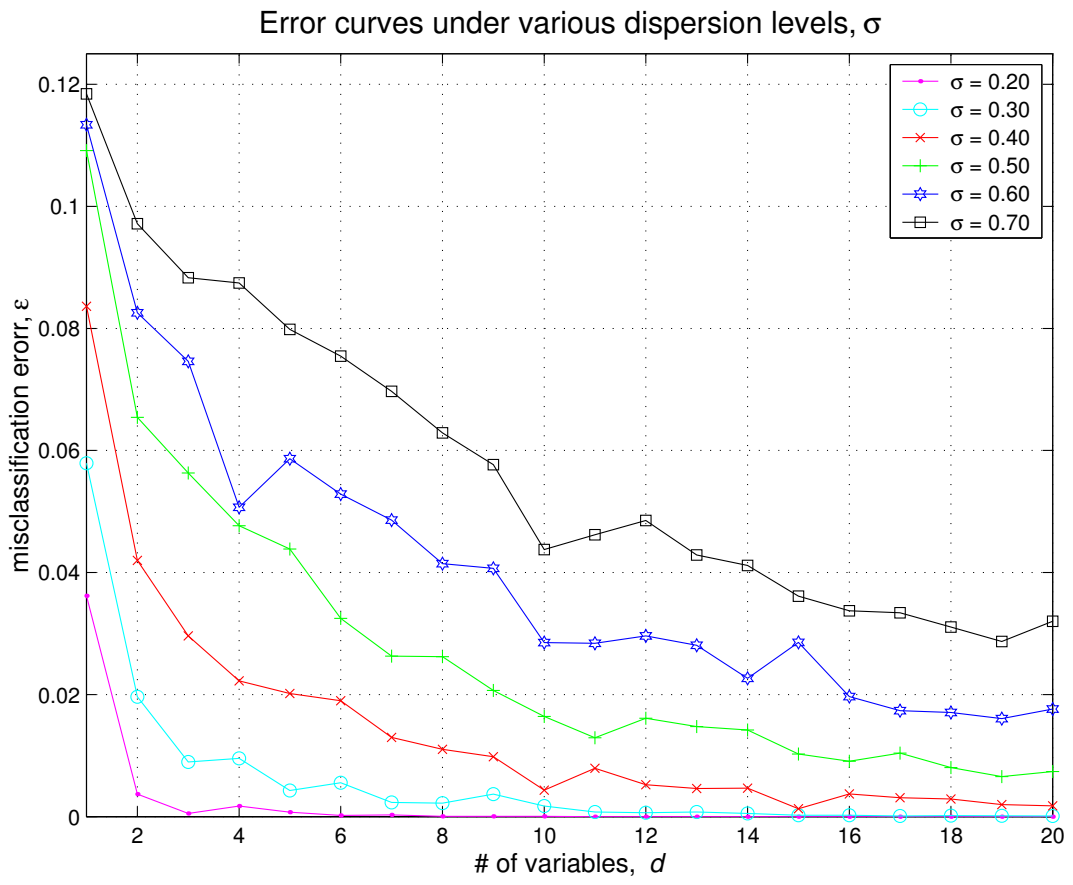


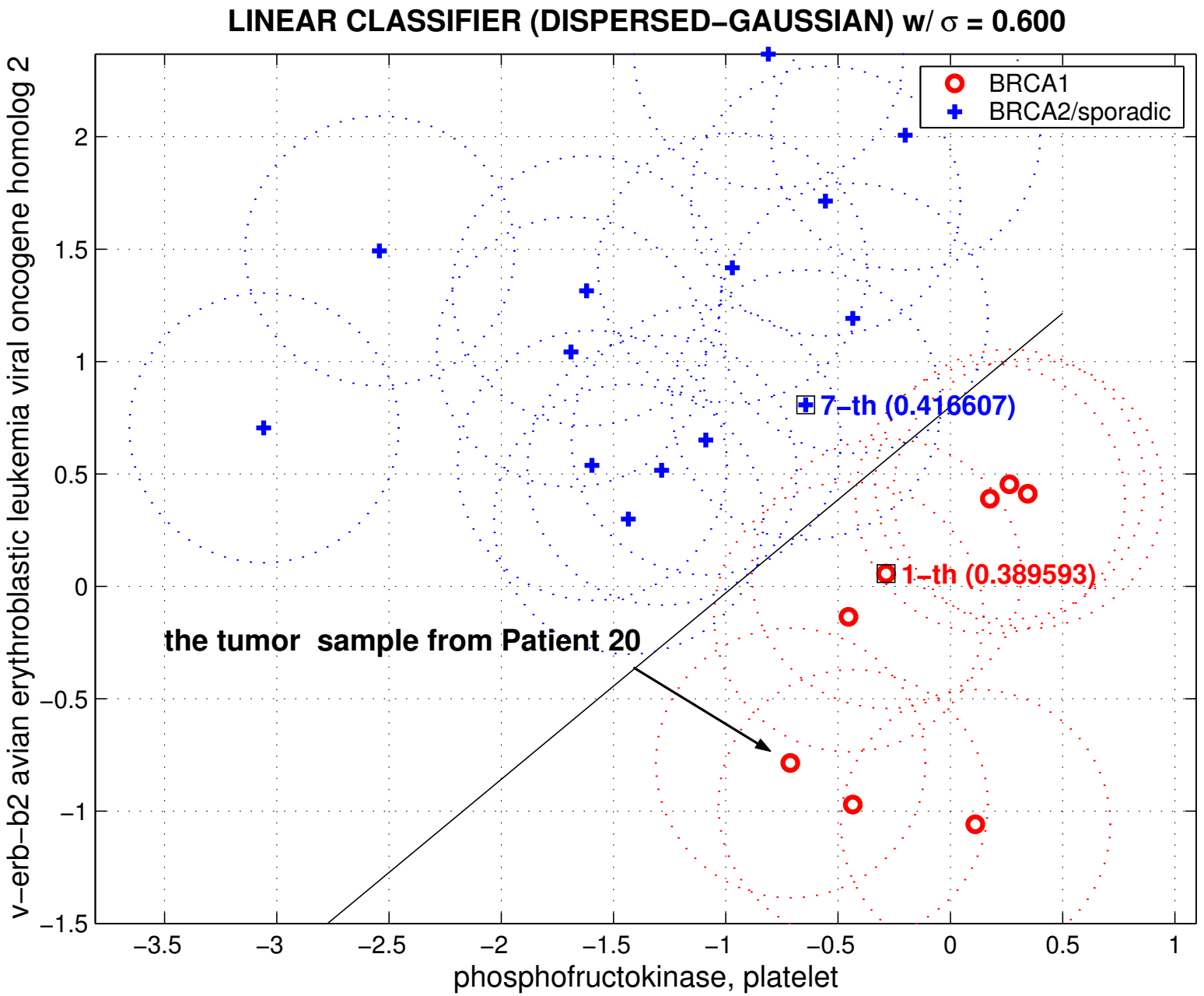


- Robustness analysis

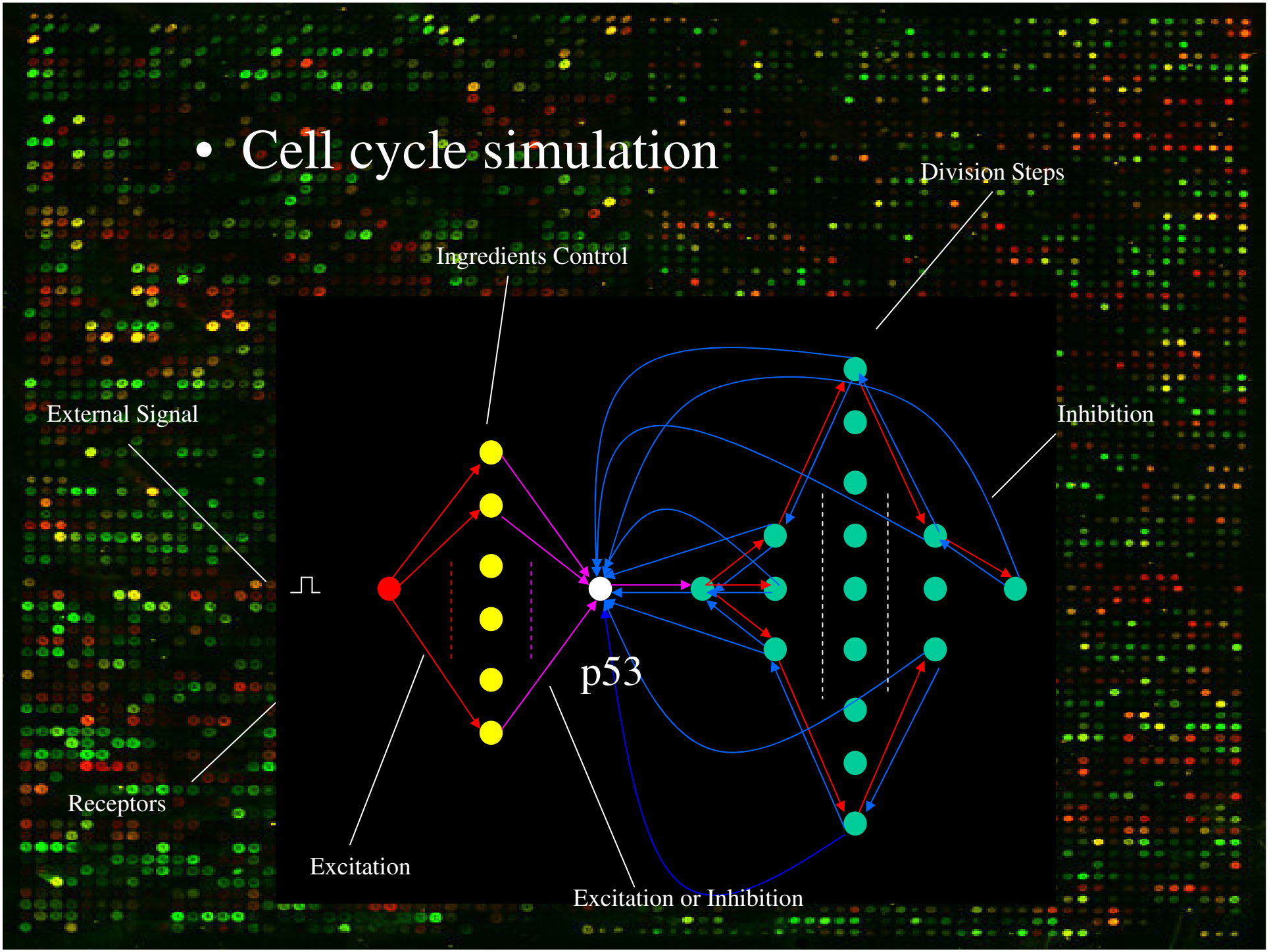





Error curves under various dispersion levels, σ
σ = 0.20
σ = 0.30
σ = 0.40
σ = 0.50
σ = 0.60
σ = 0.70
misclassification erorr, ε
# of variables, d
0
0.02
0.04
0.06
0.08
0.1
0.12
2
4
6
8
10
12
14
16
18
20

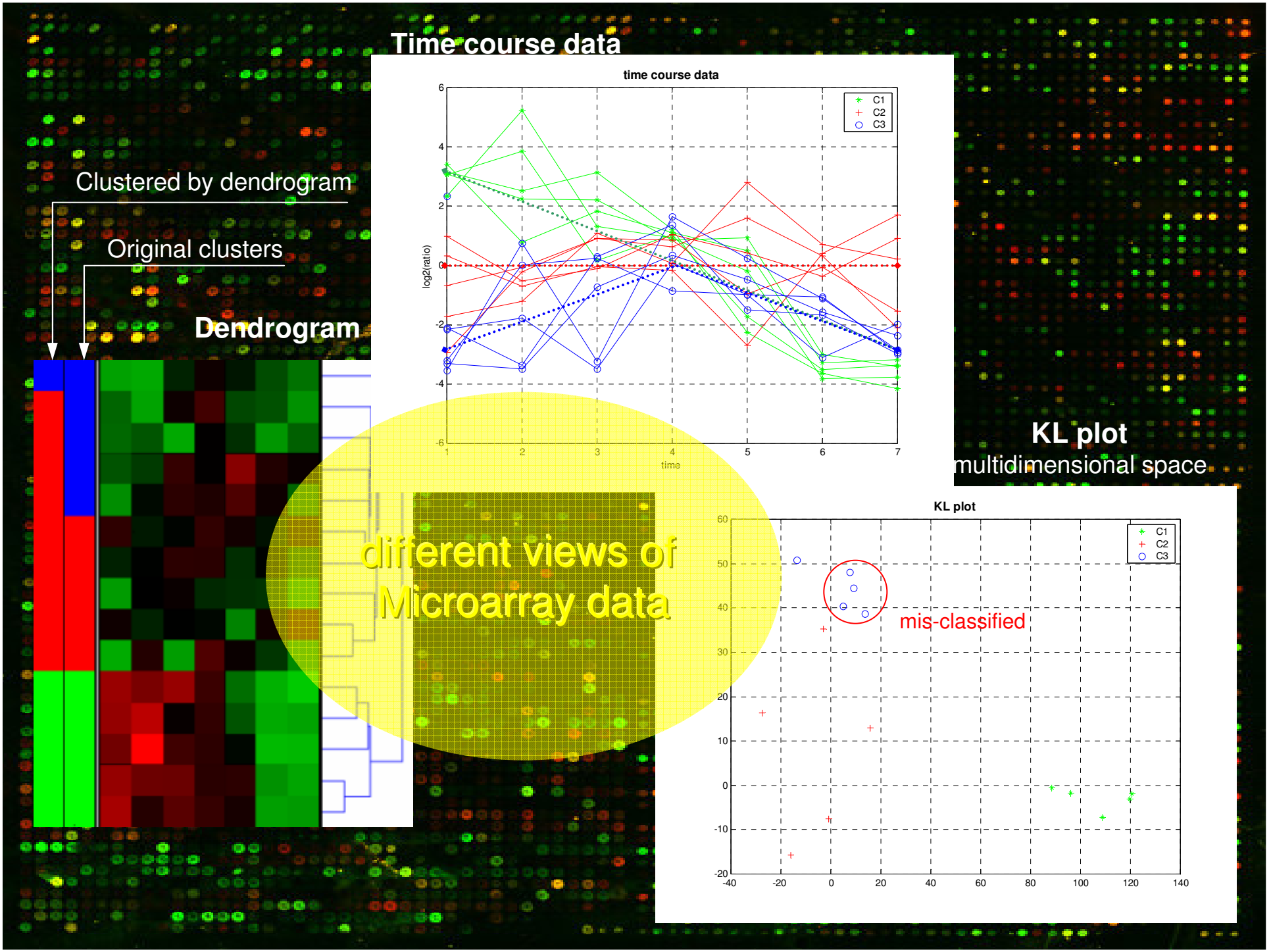

LINEAR CLASSIFIER (DISPERSED−GAUSSIAN) w/ σ = 0.600
BRCA1
BRCA2/sporadic
7−th (0.416607)
1−th (0.389593)
the tumor sample from Patient 20
v−erb−b2 avian erythroblastic leukemia viral oncogene homolog 2
phosphofructokinase, platelet
2
1.5
1
0.5
0
−0.5
−1
−1.5
−3.5
−3
−2.5
−2
−1.5
−1
−0.5
0
0.5
1

- Cell cycle simulation

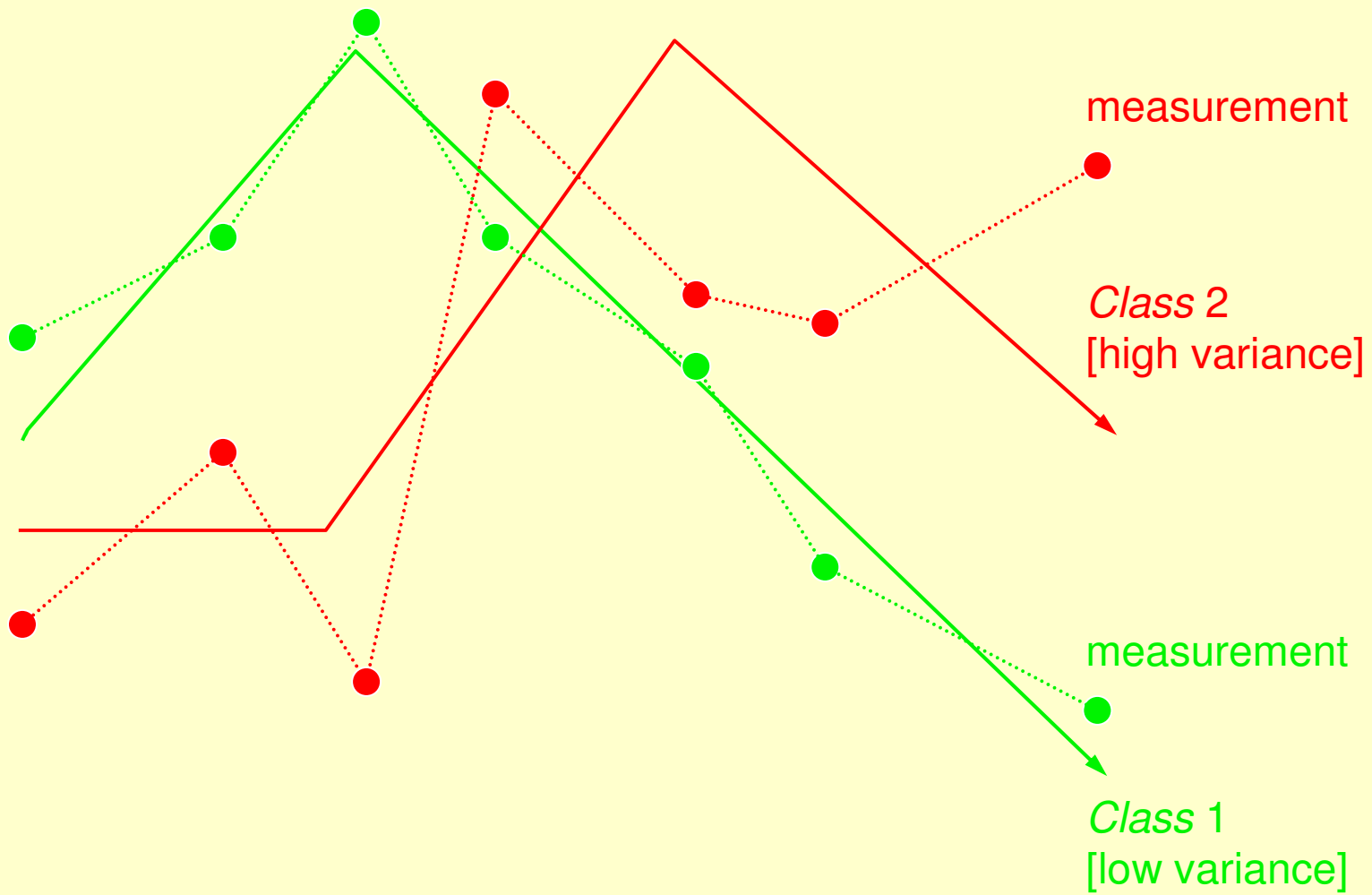


Inference from clustering

- Examine the precision of sample-based clustering relative to population inference
- Compare the number of replicates of microarray experiments
- Compare the various clustering methods


Time course data
time course data
C1
C2
C3
log2(ratio)
time
Clustered by dendrogram
Original clusters
Dendrogram
KL plot
multidimensional space
different views of
Microarray data
KL plot
mis-classified

Time Course Model



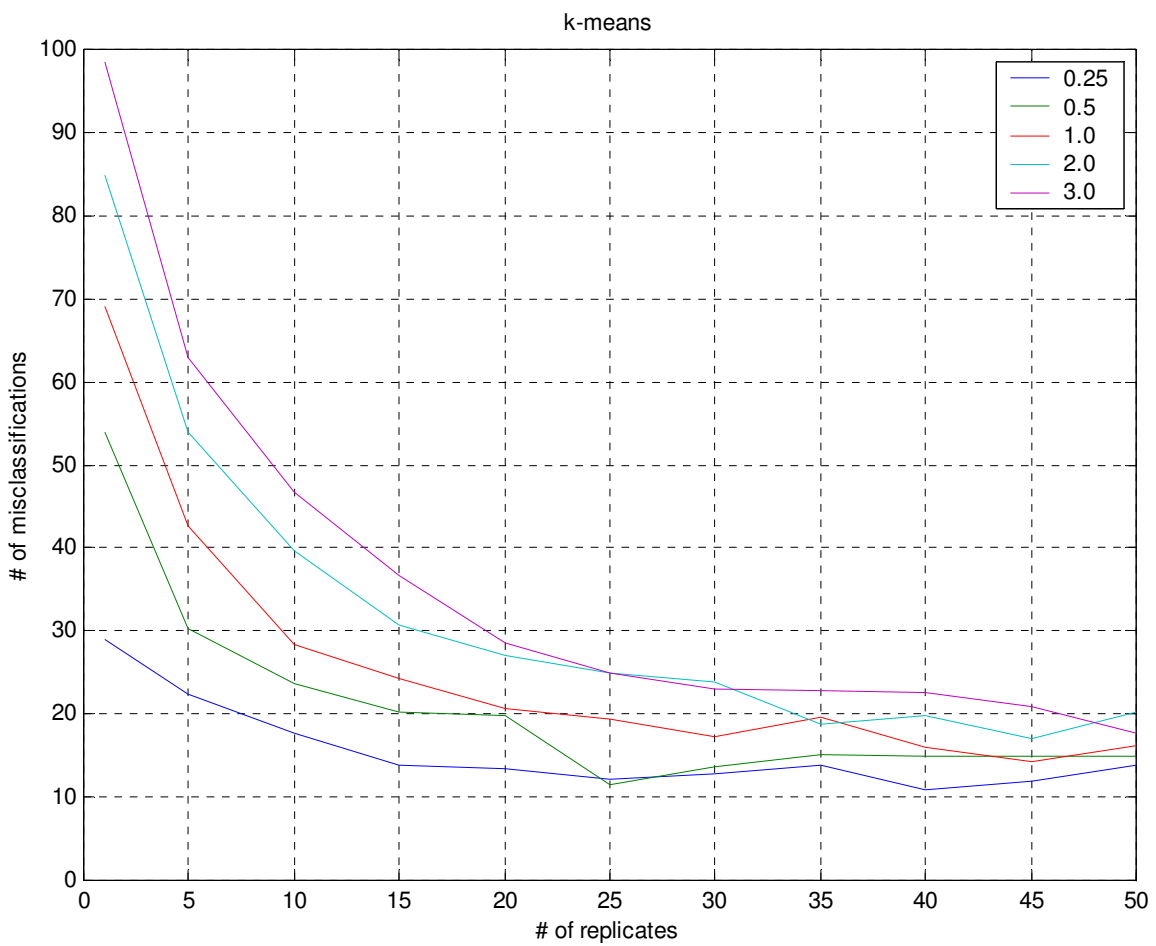
Clustering algorithms

- K-means
- Fuzzy c-means
- Self Organizing Map
- Hierarchical (dendrogram)

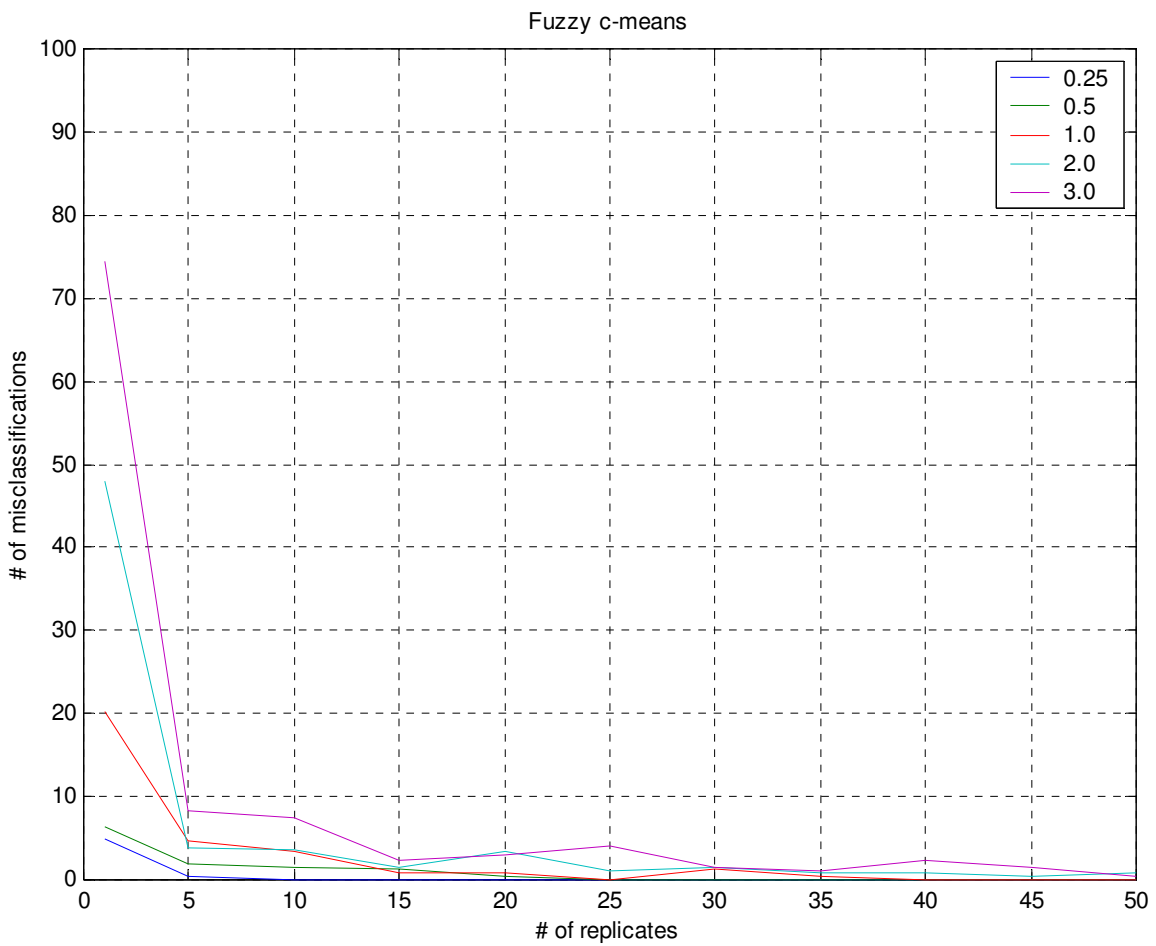
Clustering errors

- How clustering errors change as the number of replicates increases?
- How differently each clustering algorithm perform?

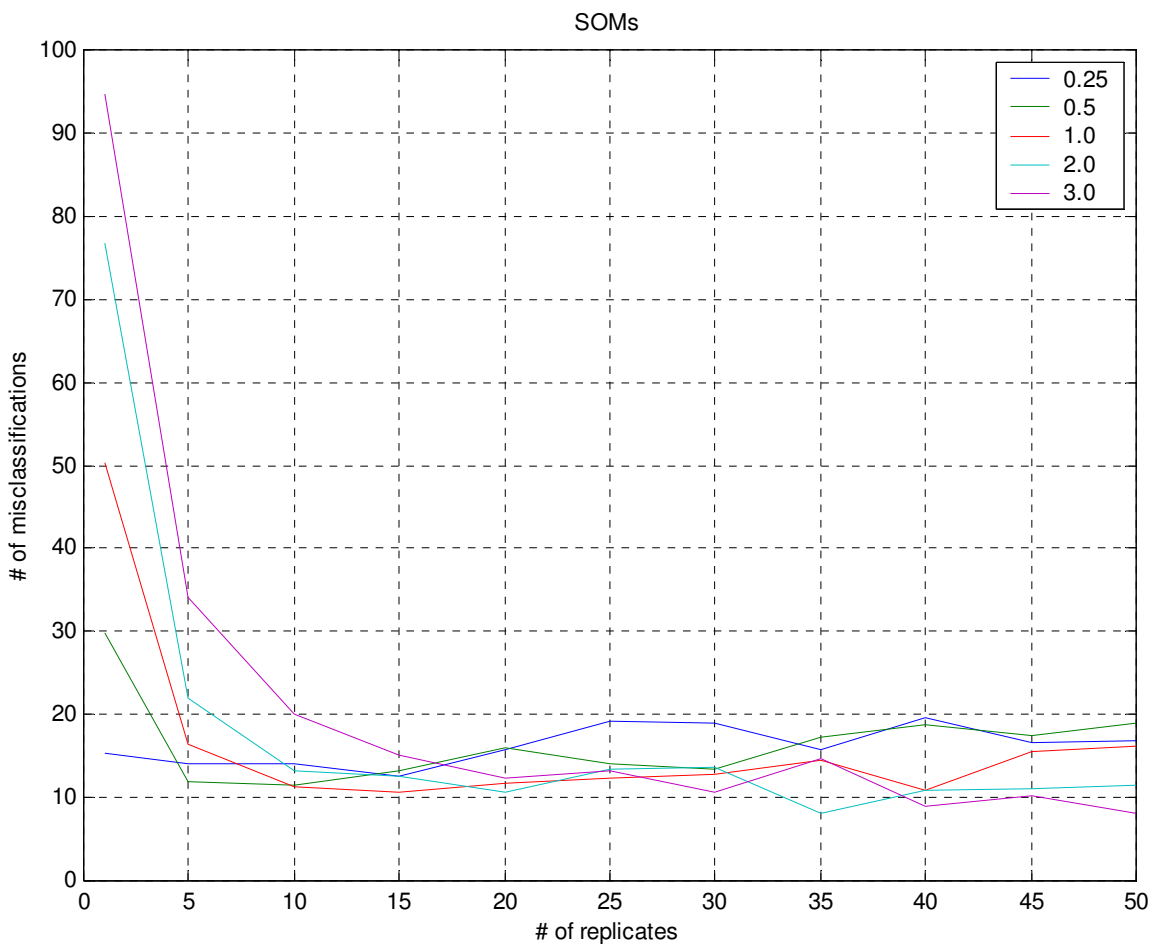
K-means


k-means
# of misclassifications
# of replicates
0
10
20
30
40
50
60
70
80
90
100
0
5
10
15
20
25
30
35
40
45
50
0.25
0.5
1.0
2.0
3.0

Fuzzy c-means



Self Organizing Maps



Variances of Data x Replicates

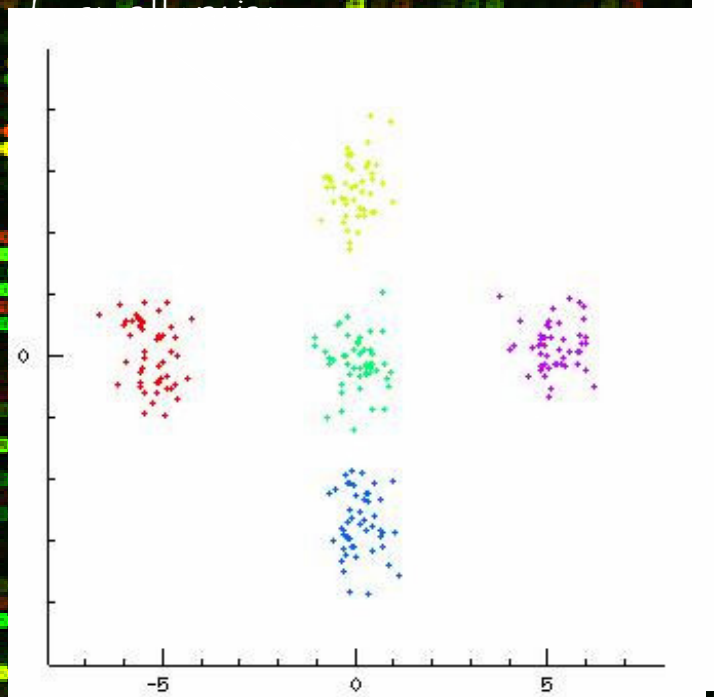
- The number of replicates required to get a reasonable clustering result varies, depending on the variance of gene expression levels
- Clustering algorithm must also be chosen correspondingly to get the best clustering algorithm. No universal clustering algorithm!

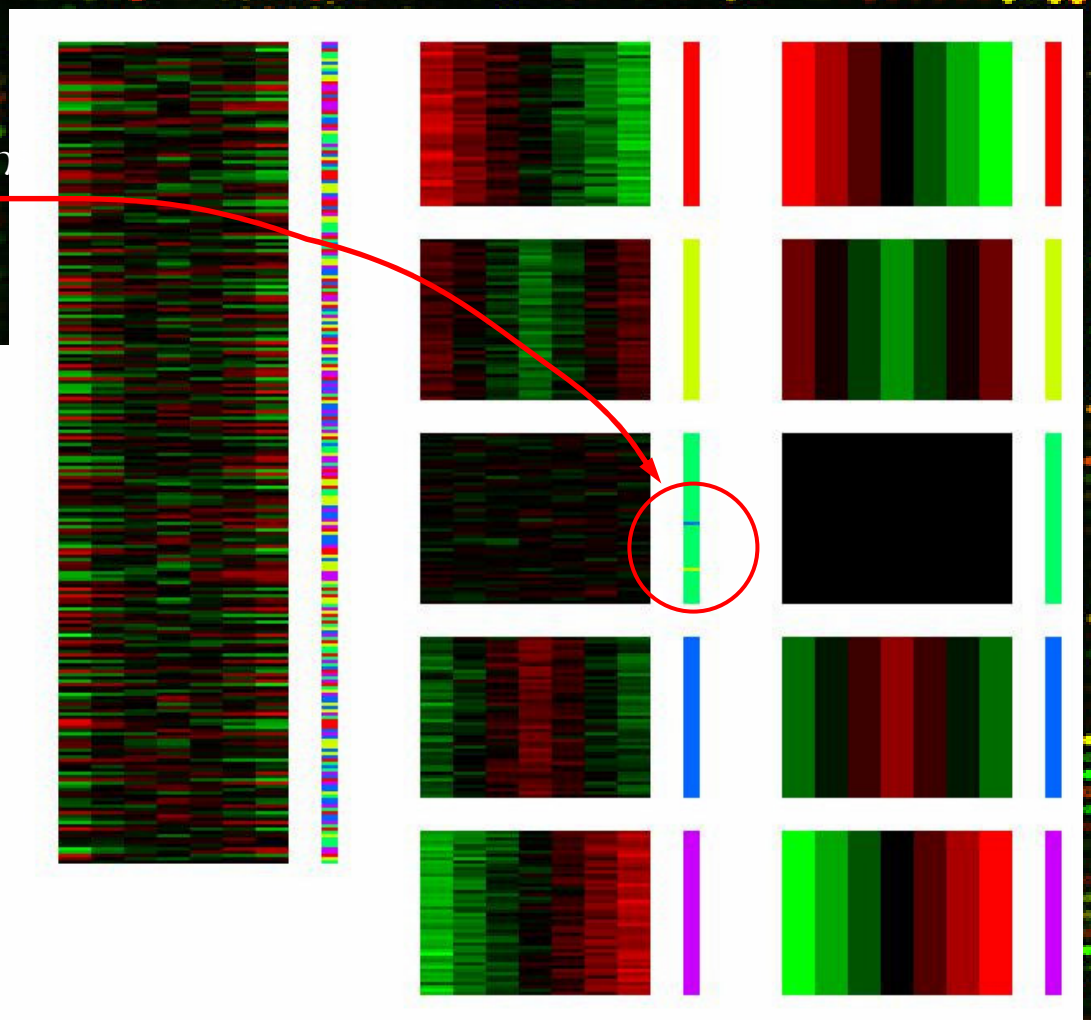
No replicate

variance = 0.25

misclassification

Tighter clusters due





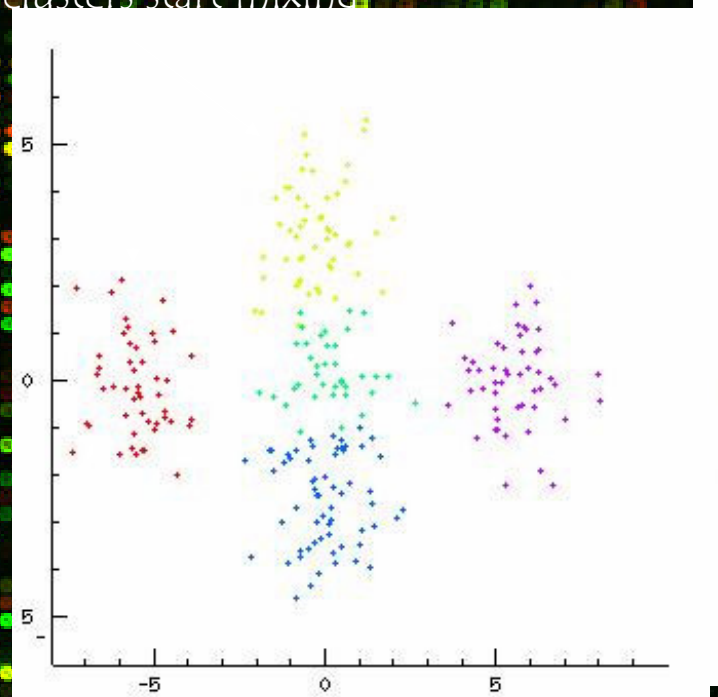
Results from Fuzzy c-means

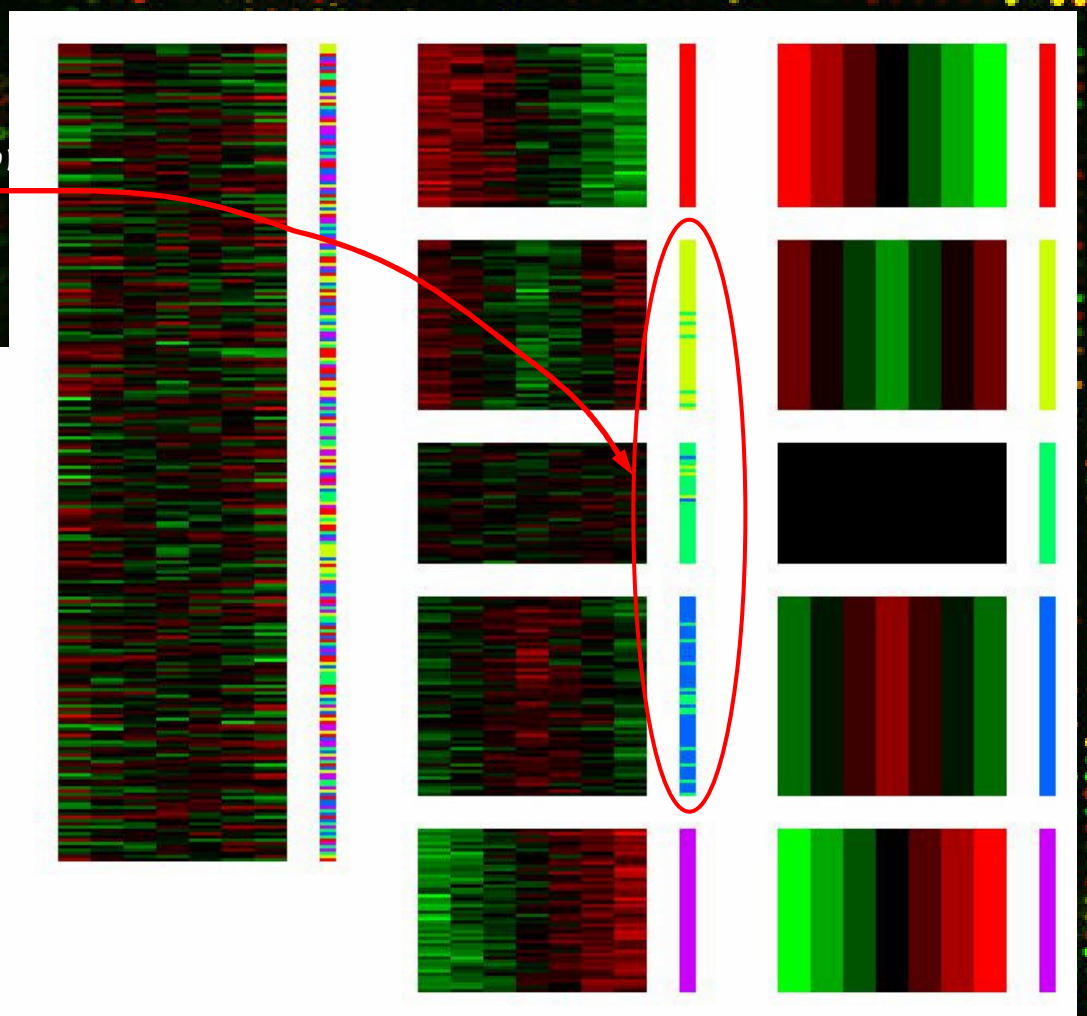
No replicate

variance = 1.0

many
misclassificatio

clusters start mixing



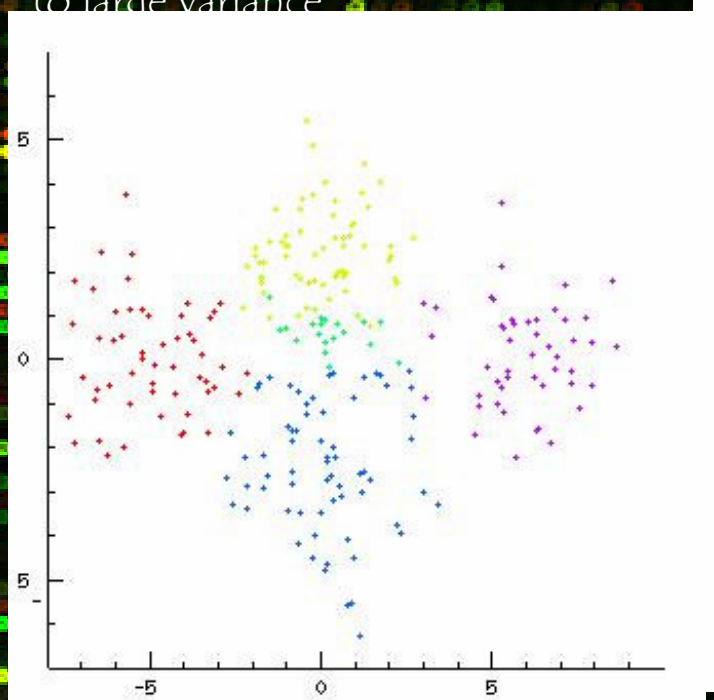


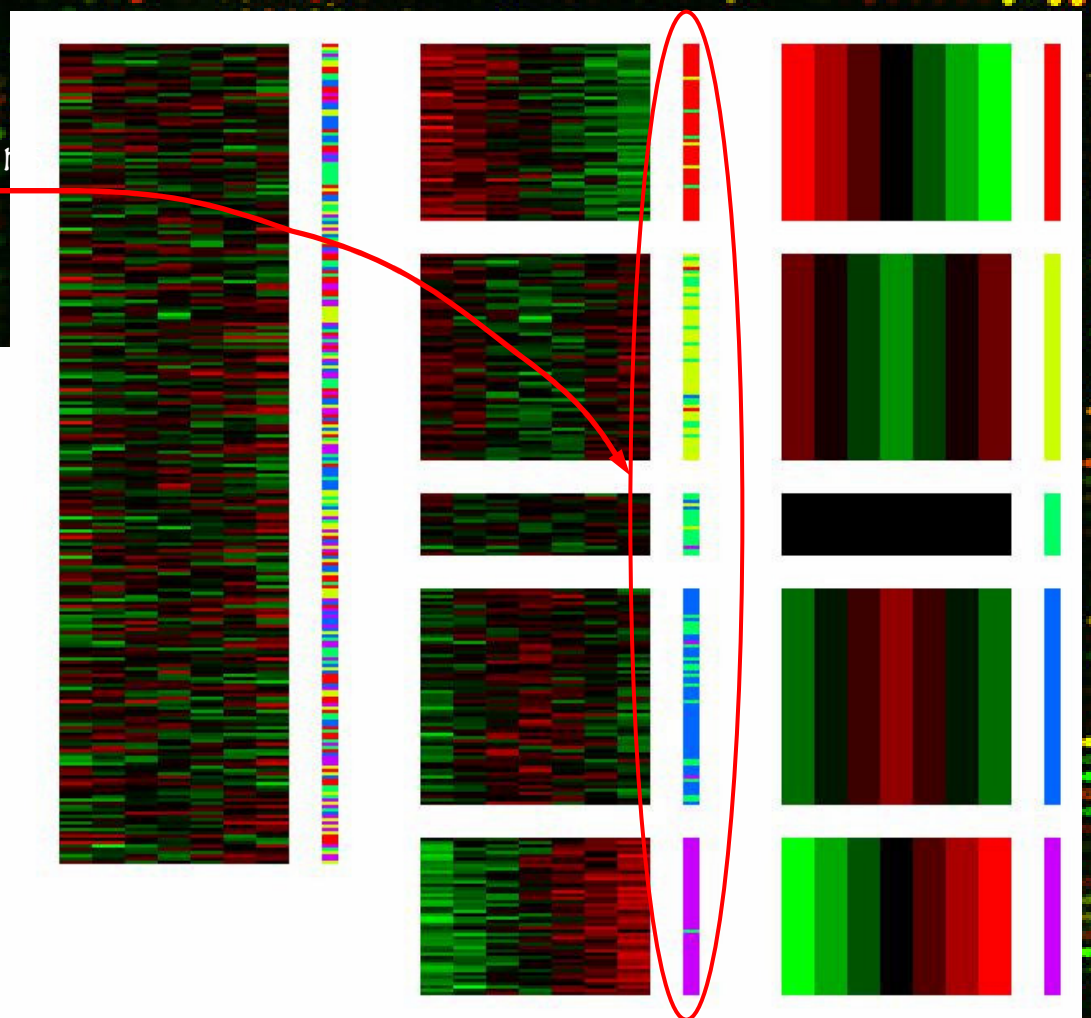
No replicate

variance = 2.0

LOTS OF
misclassificatio

Looser clusters due
to large variance



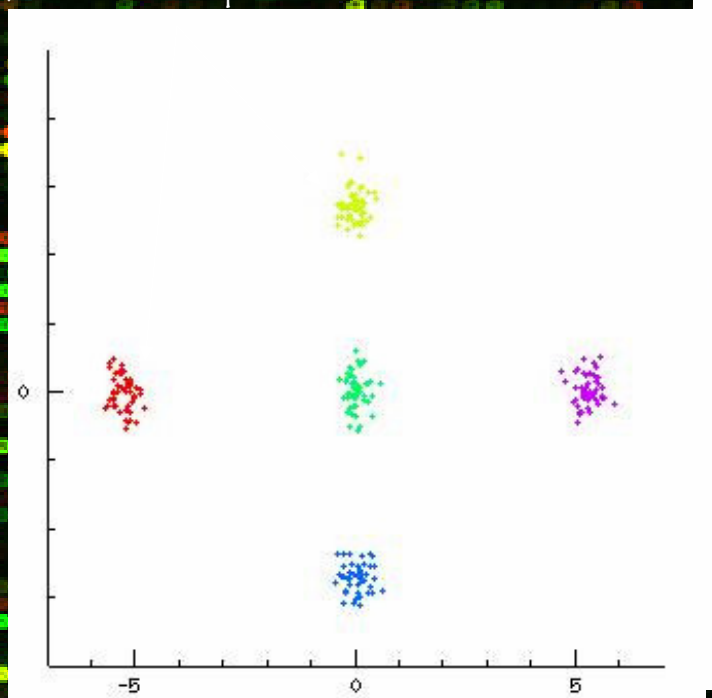


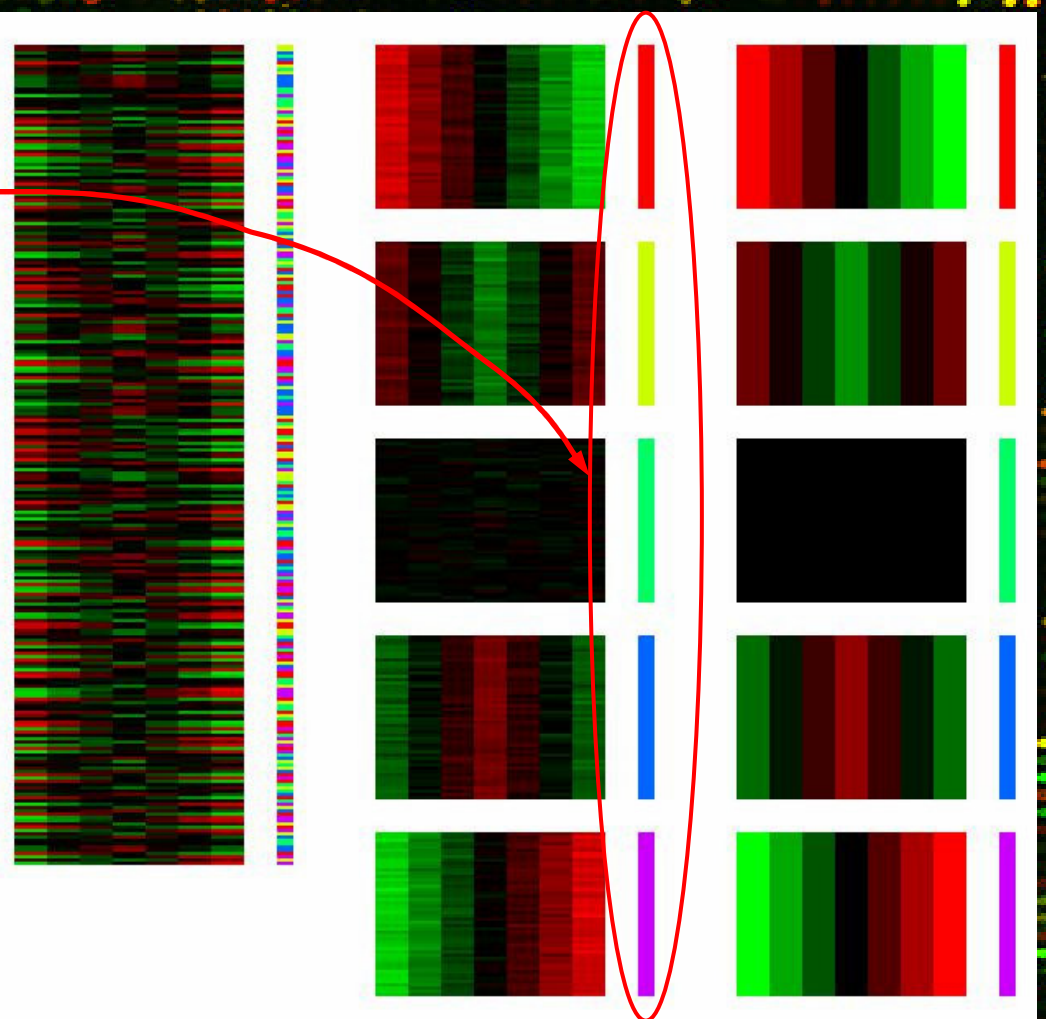
5 replicates

variance = 0.25

NO misclassificatio

Much tighter clusters due to the replications



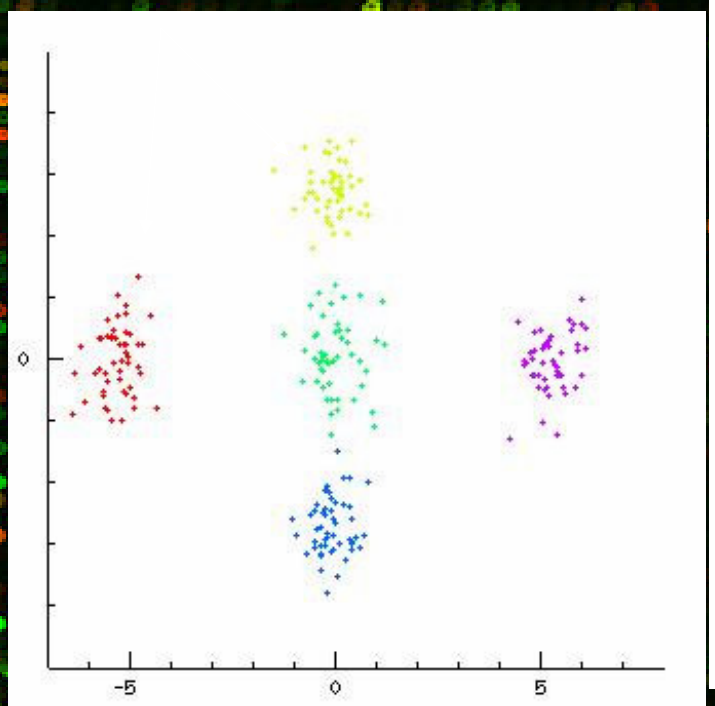


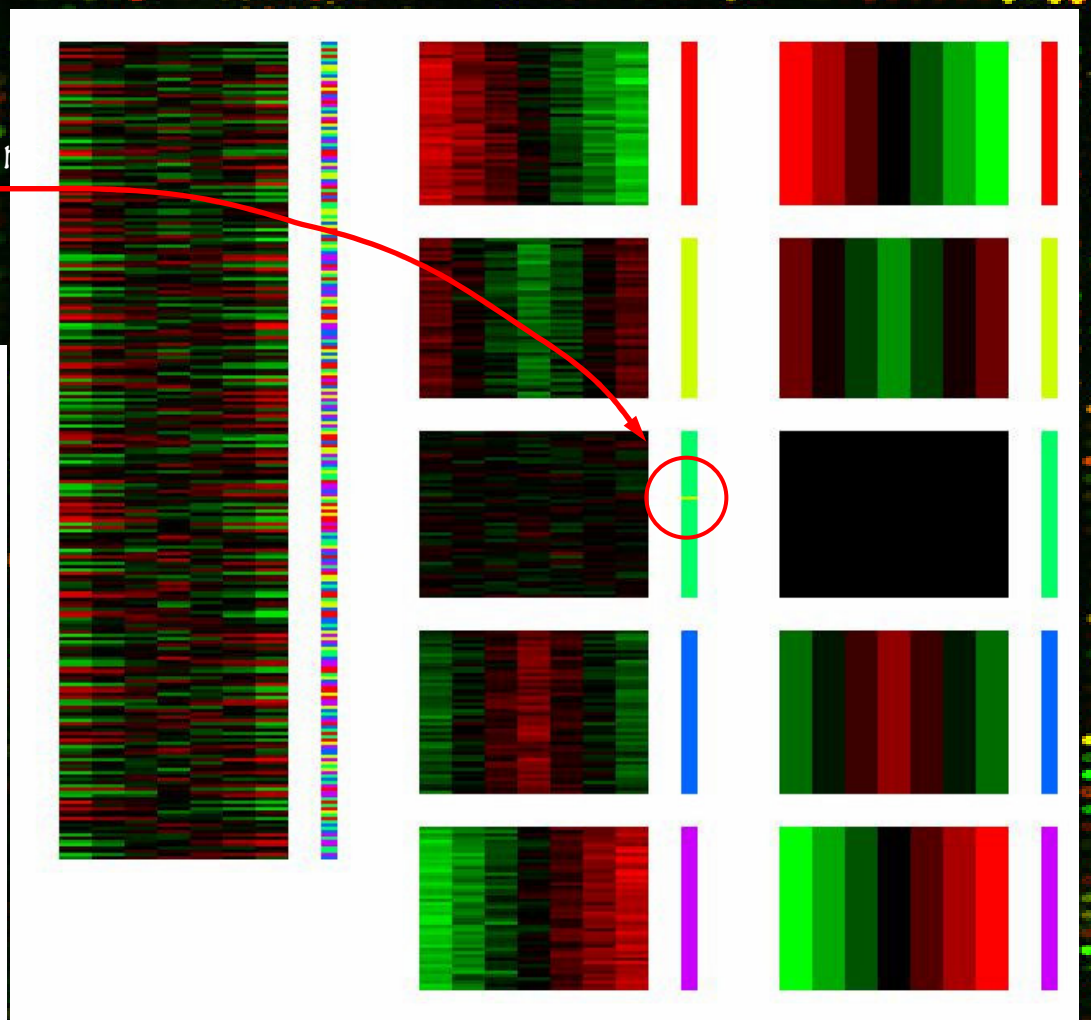
5 replicates

variance = 1.0

Clusters well separated due to the replications and relatively small variance

Very few misclassificatio



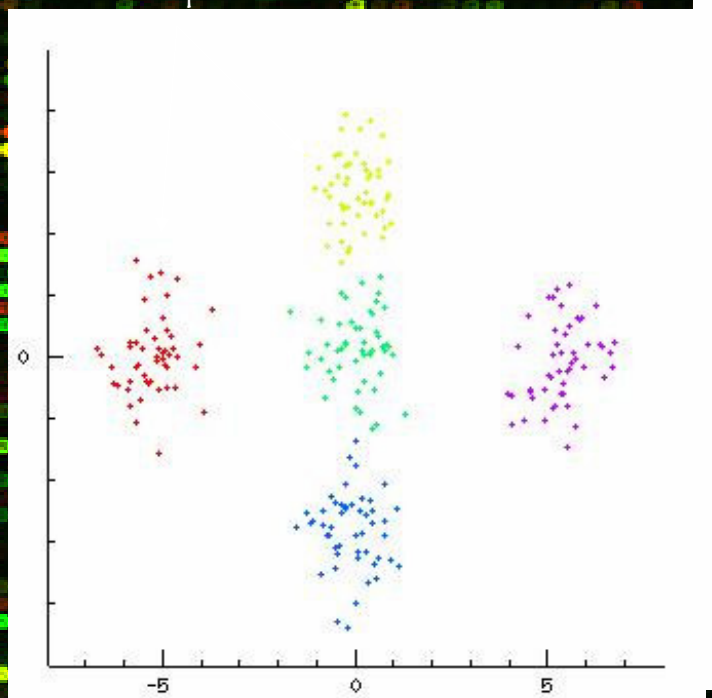


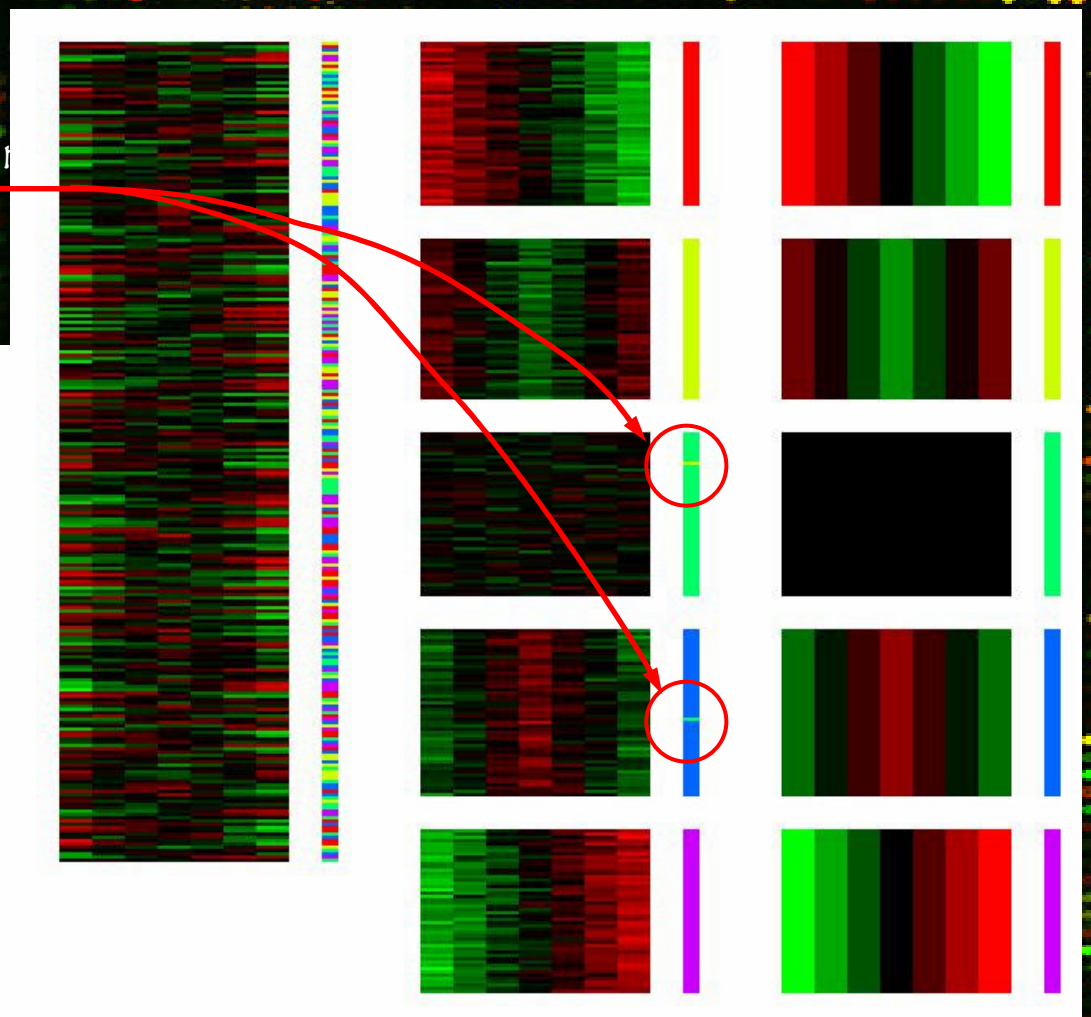
5 replicates

variance = 2.0

Very few misclassificatio

Clusters tightened due to the replications

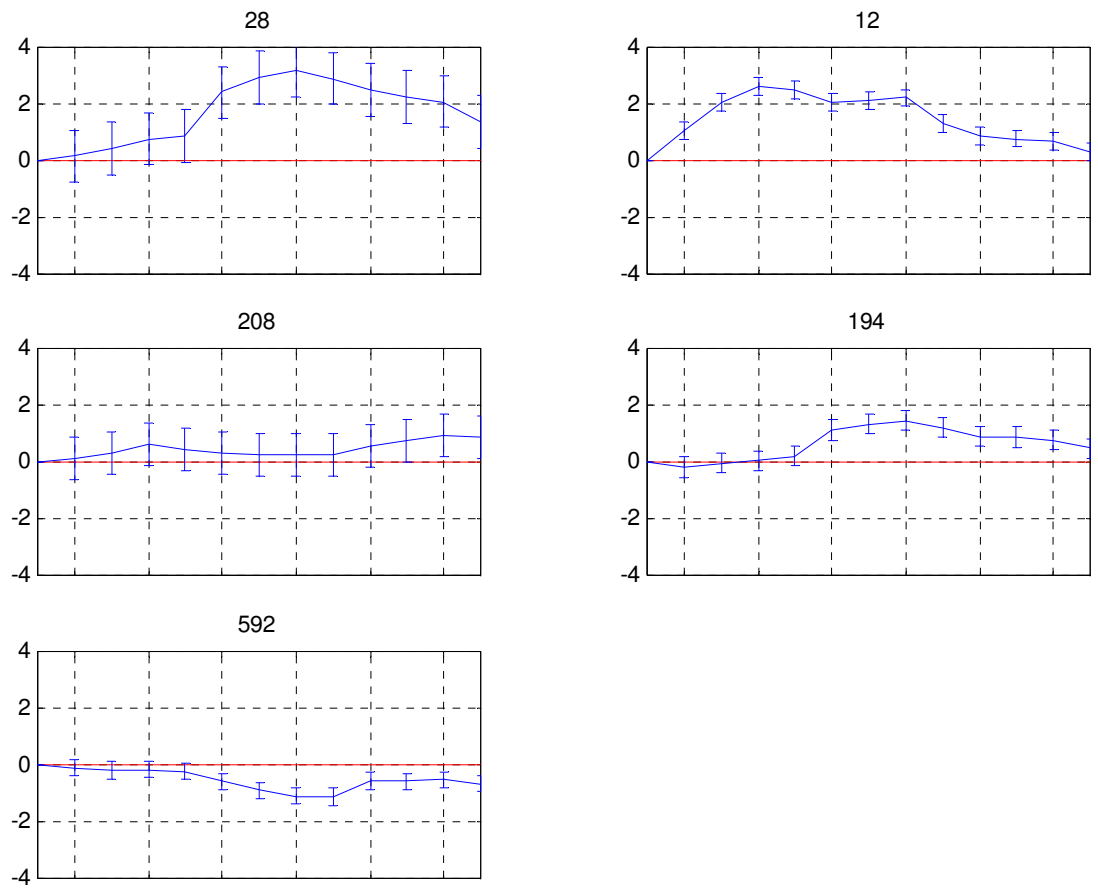




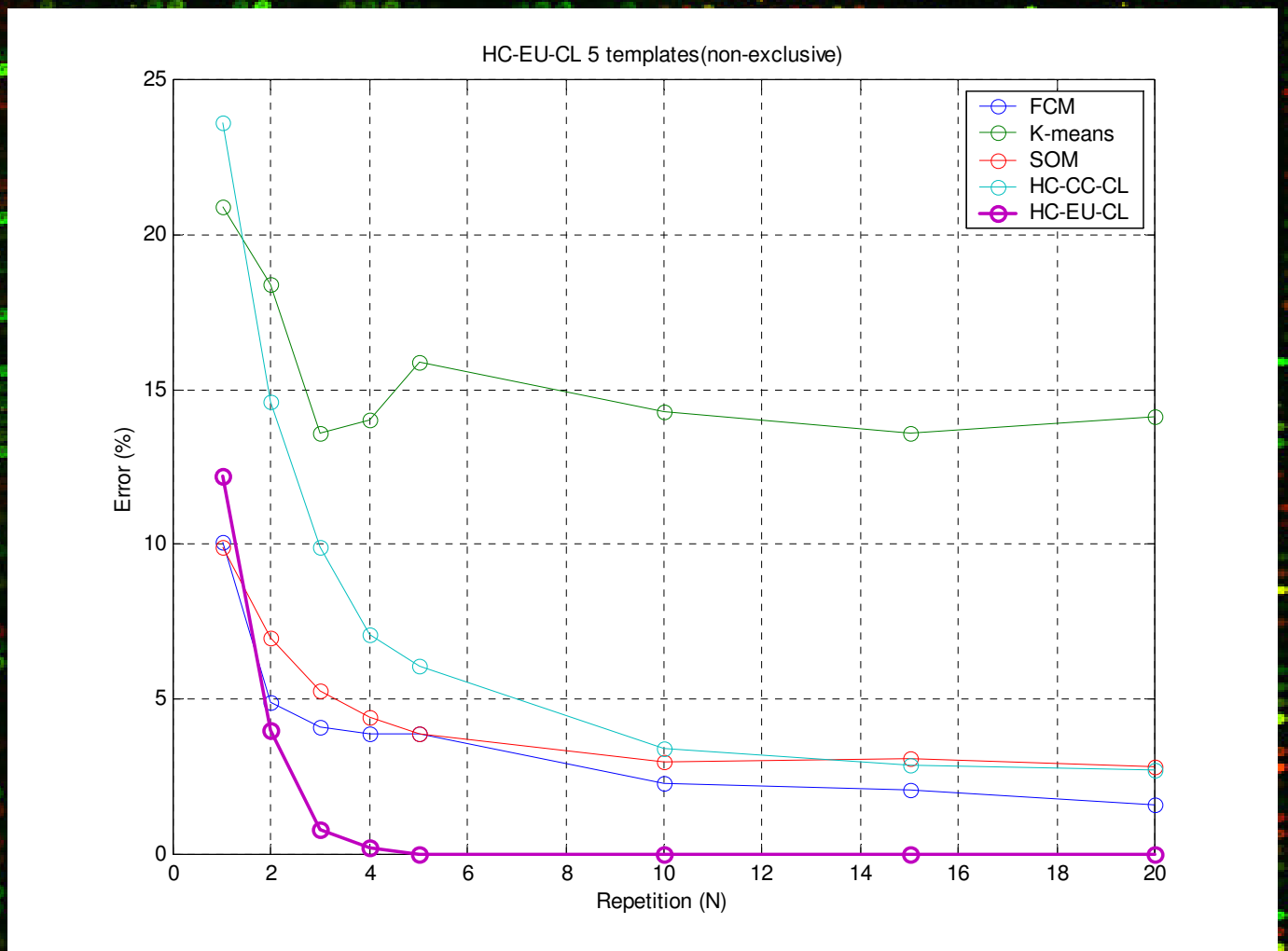
Real Data Application

- Initial clustering to generate templates
  - means
  - variance
- Simulate time course data based on the templates generated by initial clustering
  - different number of replicates
- Apply various clustering methods
  - expected clustering error for each method

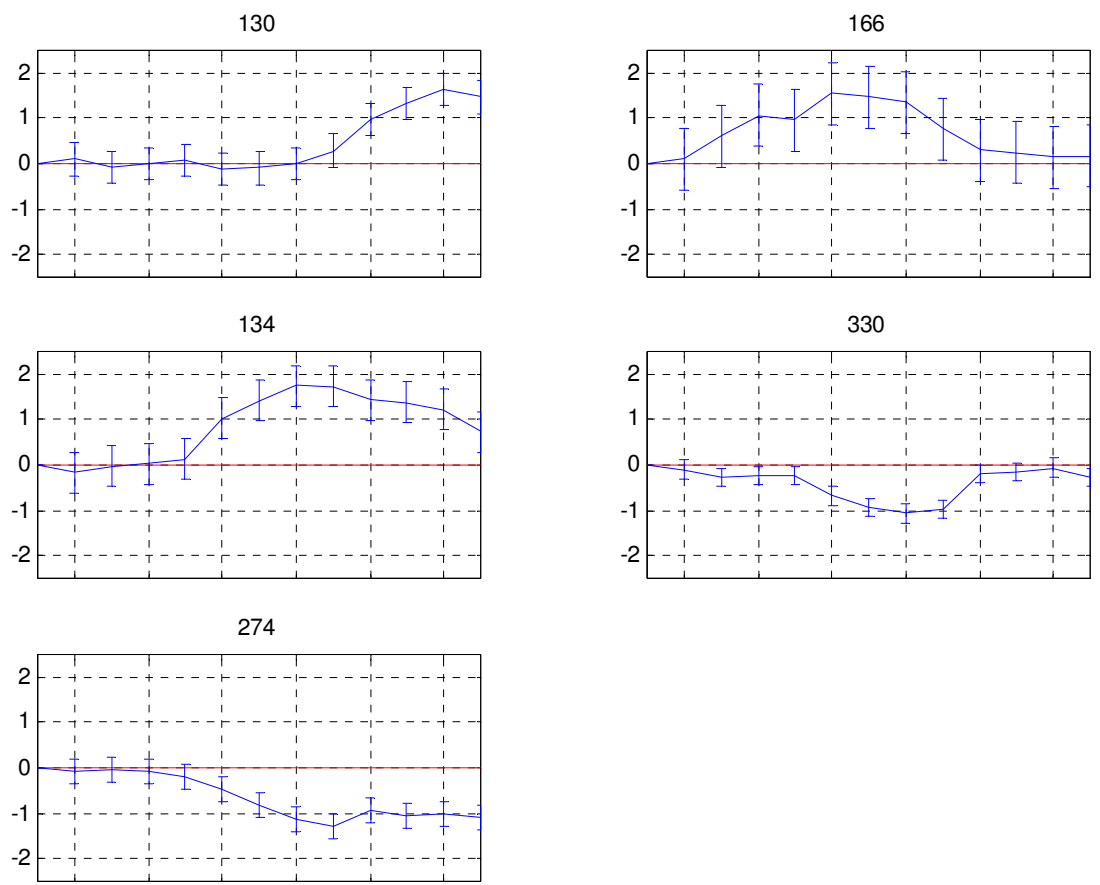
5 templates by HC



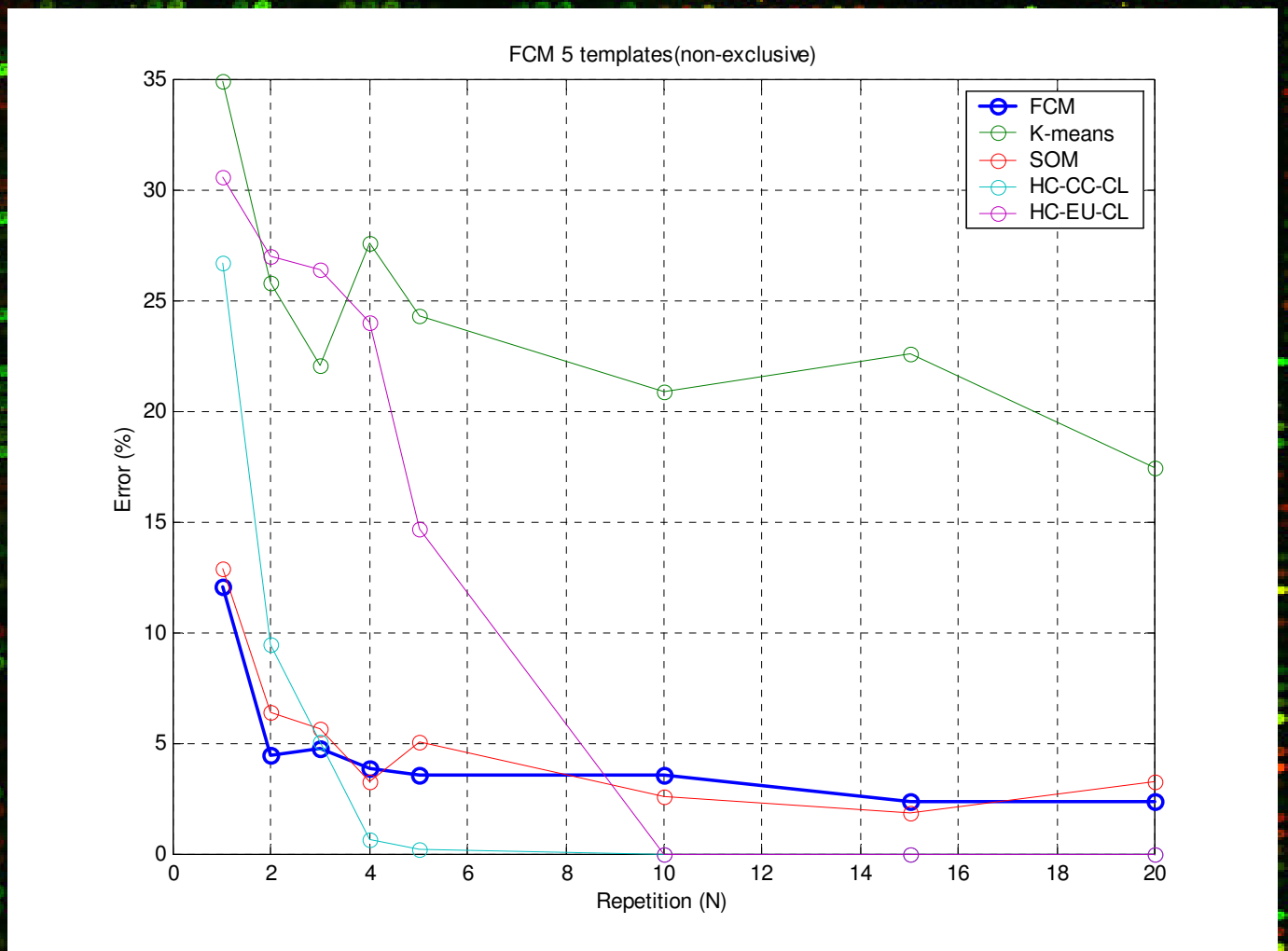
Clustering errors on HC



5 templates by FCM



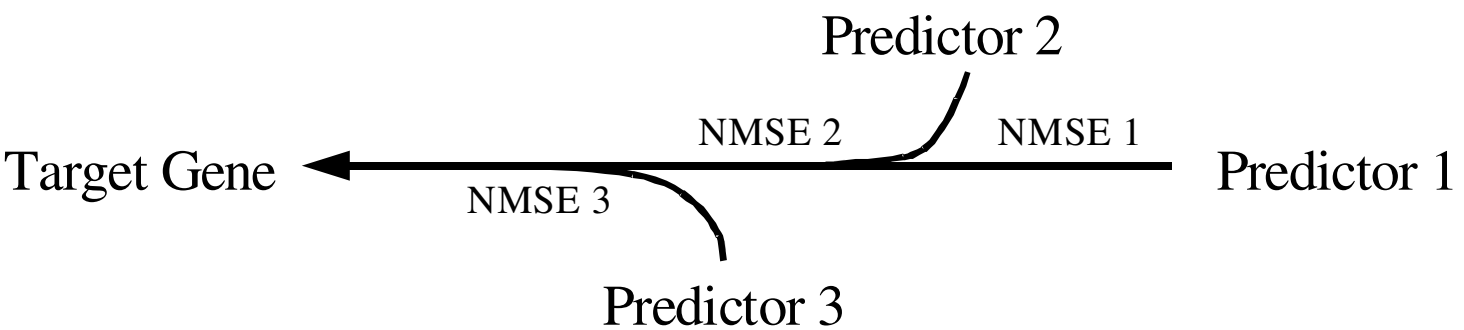
Clustering errors on FCM



Gene Regulation

- Problem: find the architecture of a gene regulation network from microarray data.
- Approach: choose small subsets of genes (2, 3 or 4), design classifier, compute the empirical error, choose the minimum error classifier. A supercomputer is required.

Gene Regulation



Gene Regulation

x1	x2	p(-1,x1,x2)	p(0,x1,x2)	p(1,x1,x2)	p(x1,x2)	y	Error
-1	-1	0.05	0.1	0.05	0.2	0	0.1
-1	0	0.03	0.03	0.04	0.1	1	0.06
-1	1	0.02	0.01	0.07	0.1	1	0.03
0	-1	0.01	0.01	0.03	0.05	1	0.02
0	0	0.03	0.01	0.01	0.05	-1	0.02
0	1	0.07	0.1	0.03	0.2	0	0.1
1	-1	0.04	0.06	0.1	0.2	1	0.1
1	0	0.03	0.01	0.01	0.05	-1	0.02
1	1	0.02	0.02	0.01	0.05	-1	0.03
							0.48

Gene regulation

		Genes														Condition		
Cell line	Condition	*RCH1*	*BCL3*	*FRA1*	*REL-B*	*ATF3*	*IAP-1*	*PC-1*	*MBP-1*	*SSAT*	*MDM2*	*p21*	*p53*	*AHA*	*OHO*	*IR*	*MMS*	*UV*
ML-1	IR	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
ML-1	MMS	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0
Molt4	IR	-1	0	0	1	1	0	1	0	0	1	1	1	1	1	1	0	0
Molt4	MMS	0	0	1	0	1	0	0	0	0	0	1	1	1	0	0	1	0
SR	IR	-1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0
SR	MMS	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0
A549	IR	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0
A549	MMS	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	1	0
A549	UV	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1
MCF7	IR	-1	0	1	1	0	0	0	0	0	1	1	1	0	1	1	0	0
MCF7	MMS	0	0	1	0	1	0	0	0	0	1	1	1	1	0	0	1	0
MCF7	UV	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1
RKO	IR	0	1	0	1	1	1	1	0	0	1	1	1	1	0	1	0	0
RKO	MMS	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	1	0
RKO	UV	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1
CCRF-CEM	IR	-1	1	1	1	1	0	1	0	0	0	0	-1	-1	0	1	0	0
CCRF-CEM	MMS	0	0	0	0	1	0	0	0	0	0	0	-1	0	0	0	1	0
HL60	IR	-1	1	0	1	1	0	1	0	1	0	1	-1	-1	-1	1	0	0
HL60	MMS	0	0	1	0	1	0	0	0	0	1	1	-1	0	1	0	1	0
K562	IR	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	1	0	0
K562	MMS	0	0	0	0	1	0	0	0	0	0	0	-1	0	0	0	1	0
H1299	IR	0	0	0	1	0	0	1	0	0	0	0	-1	0	0	1	0	0
H1299	MMS	0	0	0	0	1	0	0	0	0	0	1	-1	0	1	0	1	0
H1299	UV	0	0	0	0	1	0	1	0	0	0	1	-1	0	1	0	0	1
RKO/E6	IR	-1	1	0	1	0	1	1	0	0	0	0	-1	-1	0	1	0	0
RKO/E6	MMS	-1	0	0	0	1	0	0	0	0	0	1	-1	-1	1	0	1	0
RKO/E6	UV	-1	0	0	0	1	0	0	0	0	0	1	-1	-1	1	0	0	1
T47D	IR	0	0	0	1	0	0	0	0	0	0	1	-1	0	-1	1	0	0
T47D	MMS	0	0	0	0	1	0	0	0	0	0	1	-1	0	1	0	1	0
T47D	UV	0	0	0	0	1	0	0	0	0	0	1	-1	0	1	0	0	1

Rows are cell lines subjected to different experimental conditions.
Comparisons are to the same cell line not exposed to the experimental treatment.

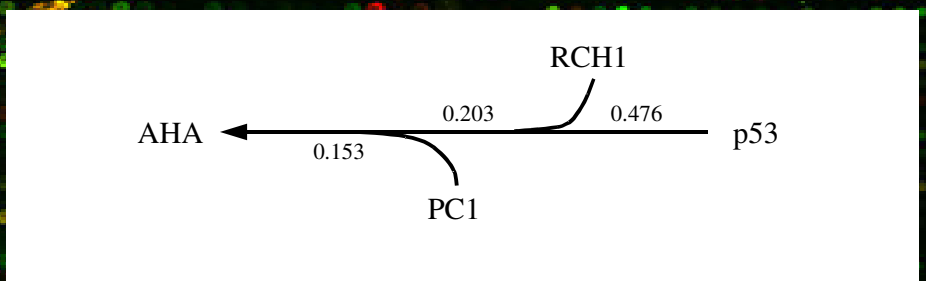
Gene regulation

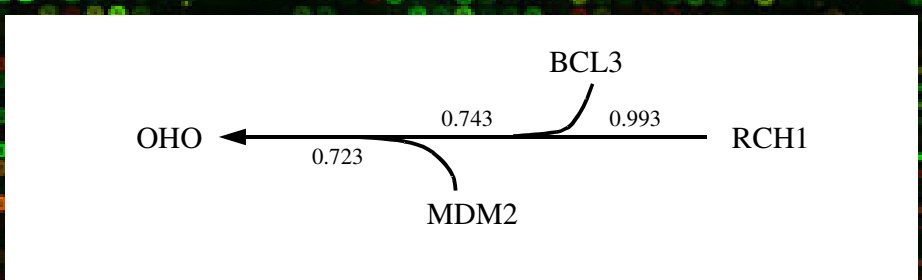
- split data in two parts: 2/3 and 1/3
- 2/3: training the predictor
- 1/3: empirical error measure
- create all predictors with less than 4 genes and measure their empirical error

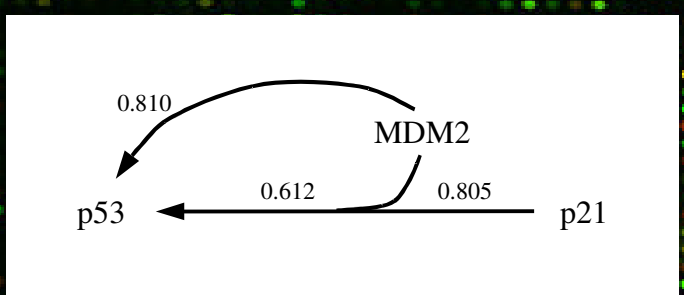
Gene regulation

- repeat for 256 random splitting and take their mean empirical error
- choose the predictors with error less than *75%*

Gene Regulation







Gene Regulation

- some well known paths of the graph were verified
- several unknown ones were suggested
- The possible new paths should be tested by specific biochemical experiments