IDENTIFICATION OF GENE NETWORKS

Junior Barrera

Layout

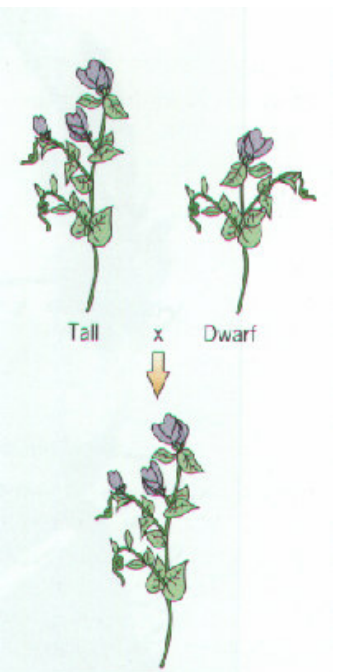
- Introduction
- Network model
- Microarray: the genetic signal
- Network identification
- Computational learning
- Genetic signal x digital images
- Image operators designed by learning

Layout

- Introduction
- Network model
- Microarray: the genetic signal
- Network identification

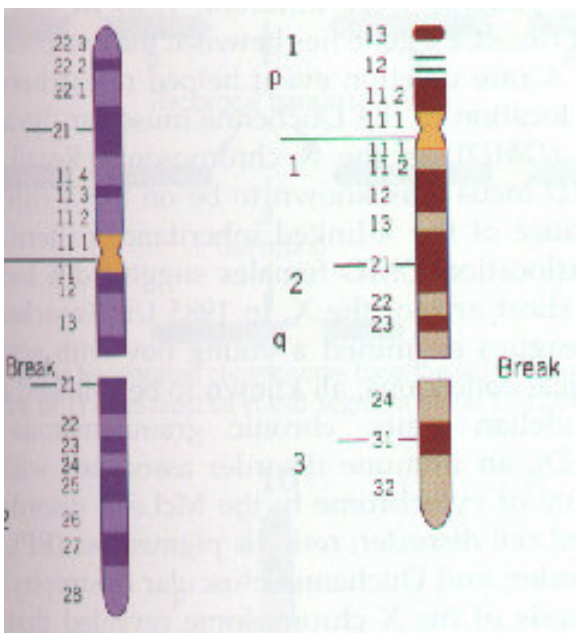
Introduction

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



Introduction

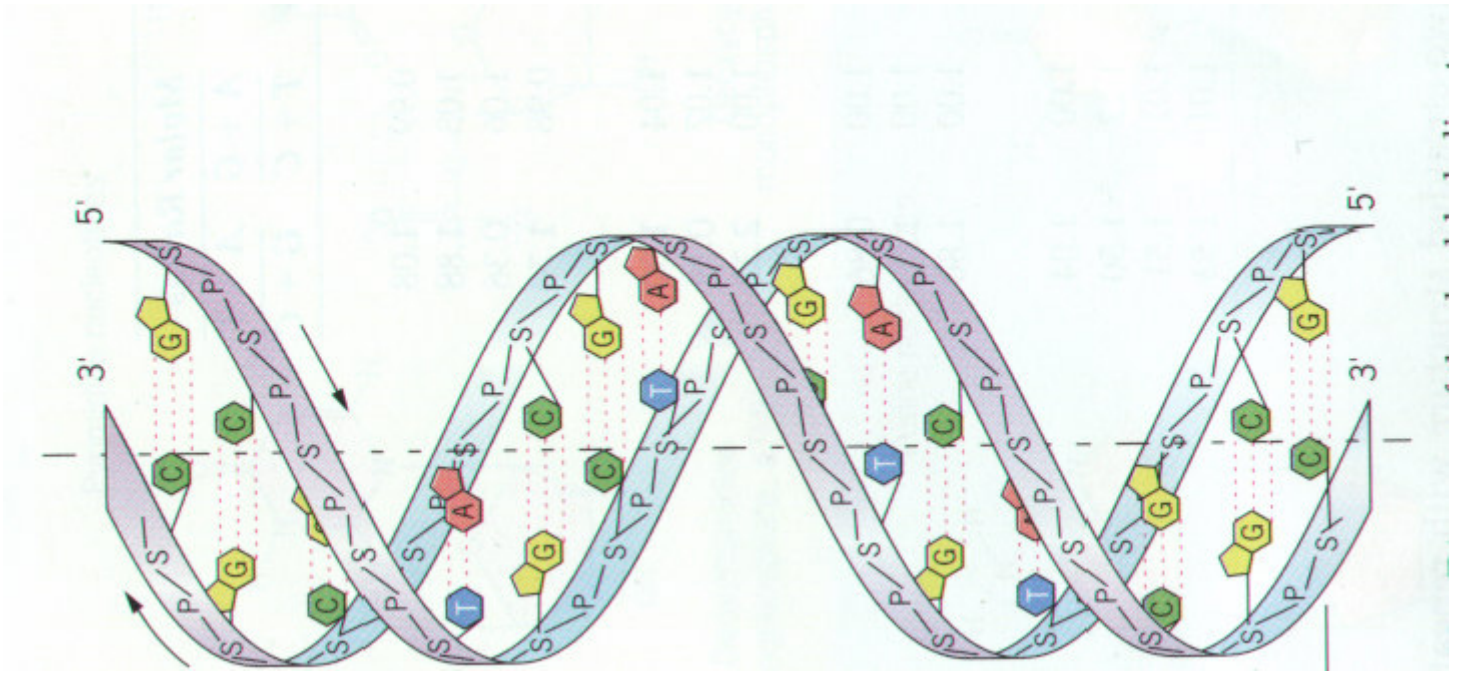
- Chromosome Theory - Morgan (1910)
- Genes were situated in chromosomes



Introduction

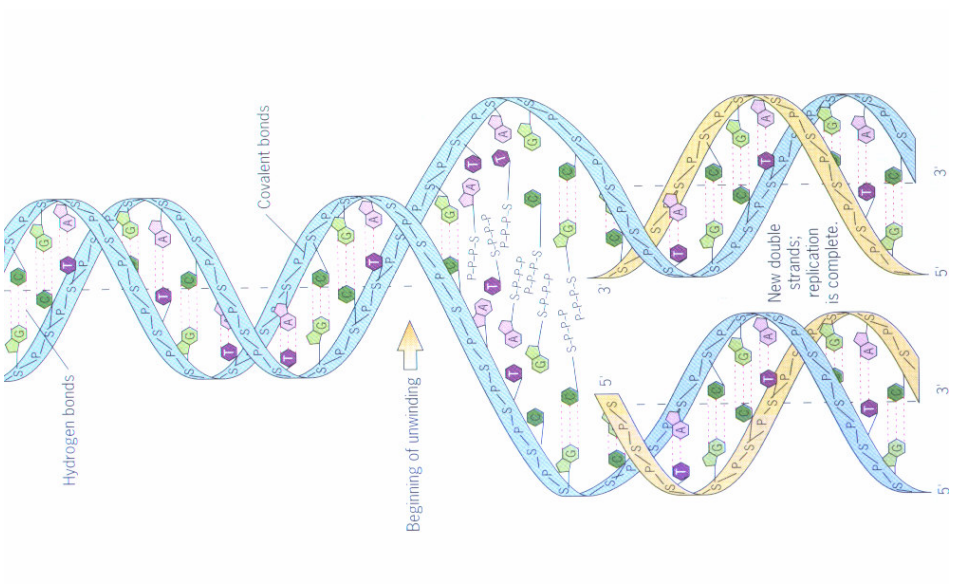
- The molecular structure of chromosomes (Watson and Crick - 1953)
- DNA structure: the double helix
- Four basis: adenine(A), guanine(G), thymine(T), cytosine(C)
- genes are sequences of nucleotides

Introduction



Introduction

- DNA manipulation
- cut, replication and decoding



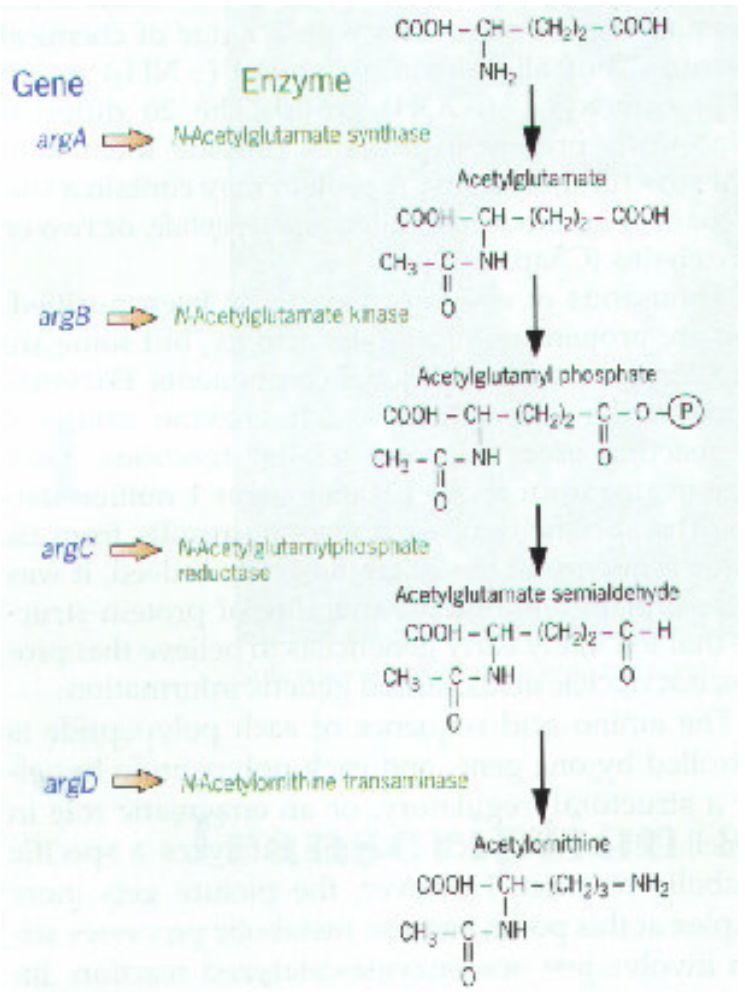
Introduction

- Genetic engineering
- species modification, drug production

Introduction

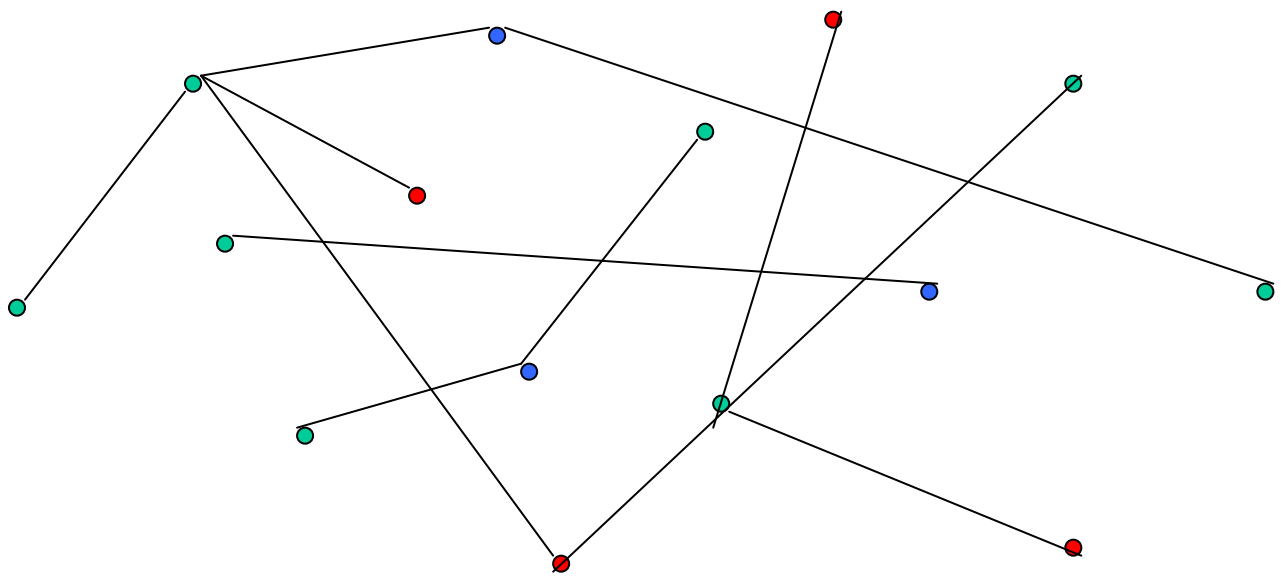
- Genes control the metabolism
- Metabolism occurs by sequences of enzyme-catalyzed reactions.
- Enzymes are specified by one or more genes

Introduction



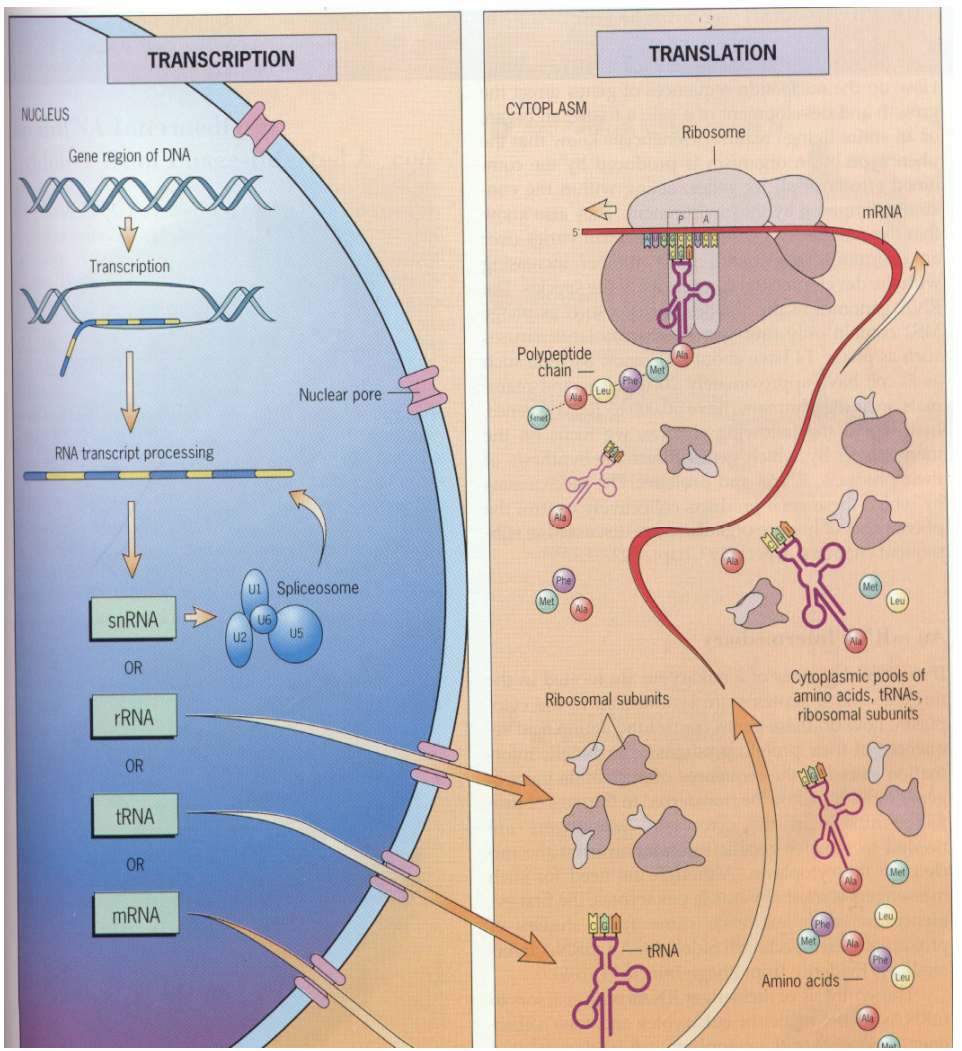
Introduction

Genes may activate or inactivate other genes



Introduction

- Gene expression

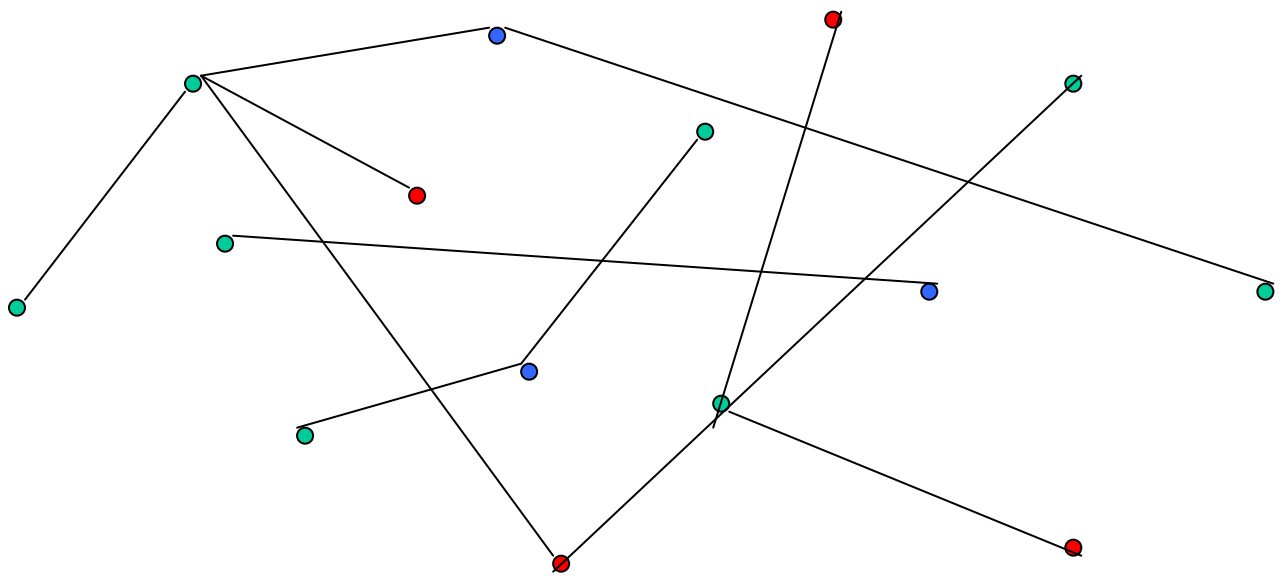


Introduction

- Nets of genes control cellular phenomena
- Their architectures and dynamics are unknown
- Microarrays permit to observe some network states
- Computational learning permits to estimate some net architectures and dynamics

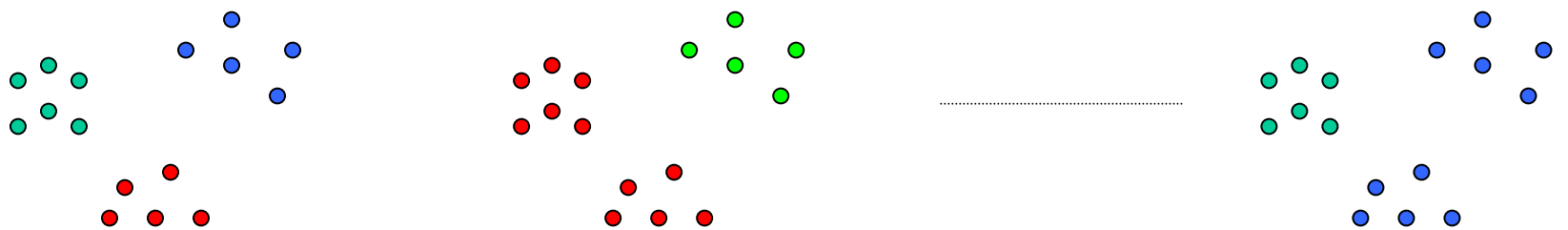
Network model

- gene network is a communication network



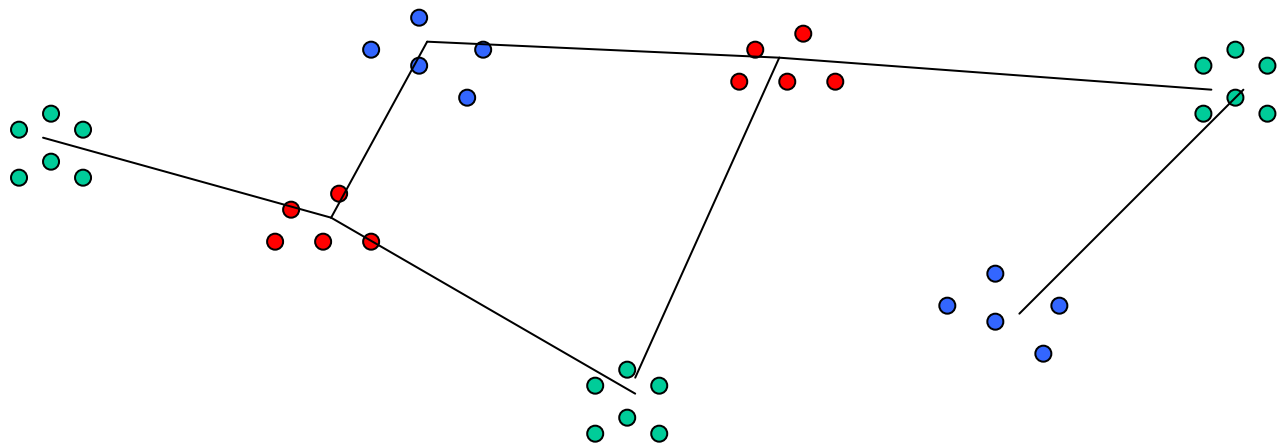
Network model

- Study system evolution in time



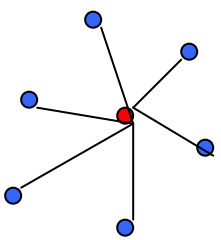
Network model

- Redundant paths



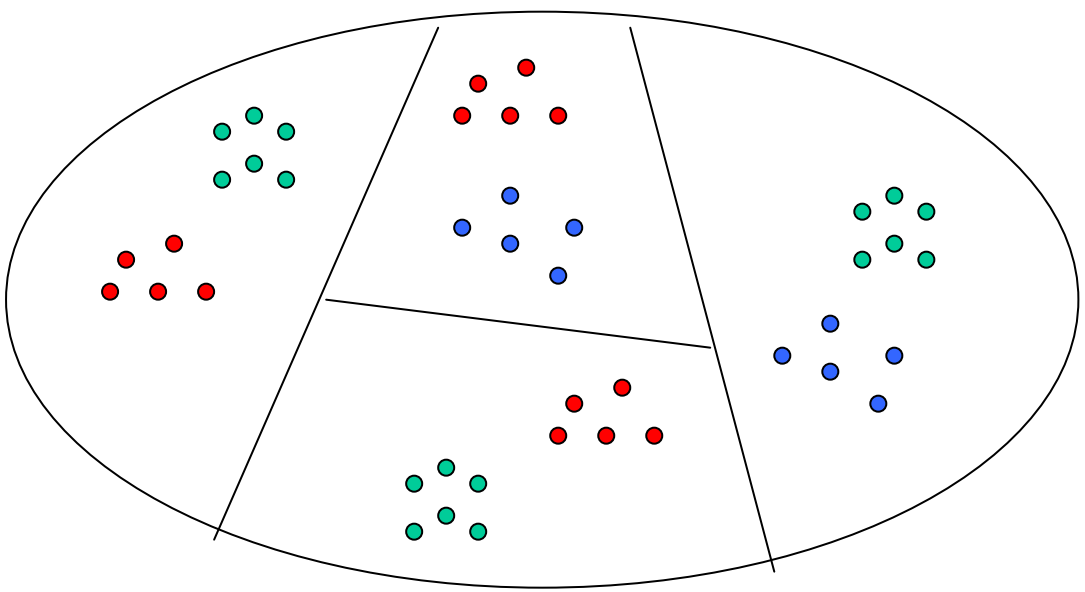
Network model

- Limited number of neighbors



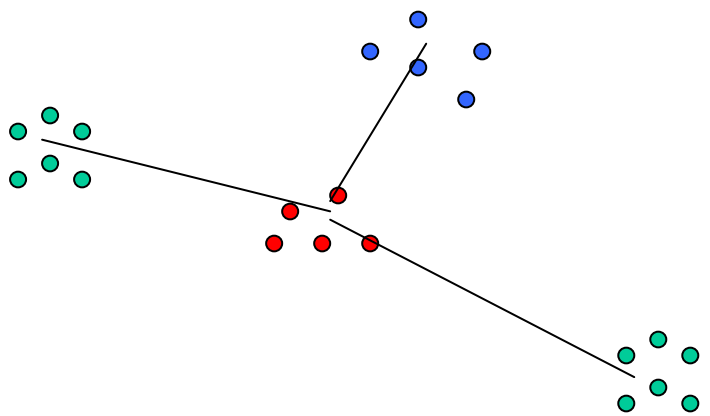
Network model

- Decentralized processing



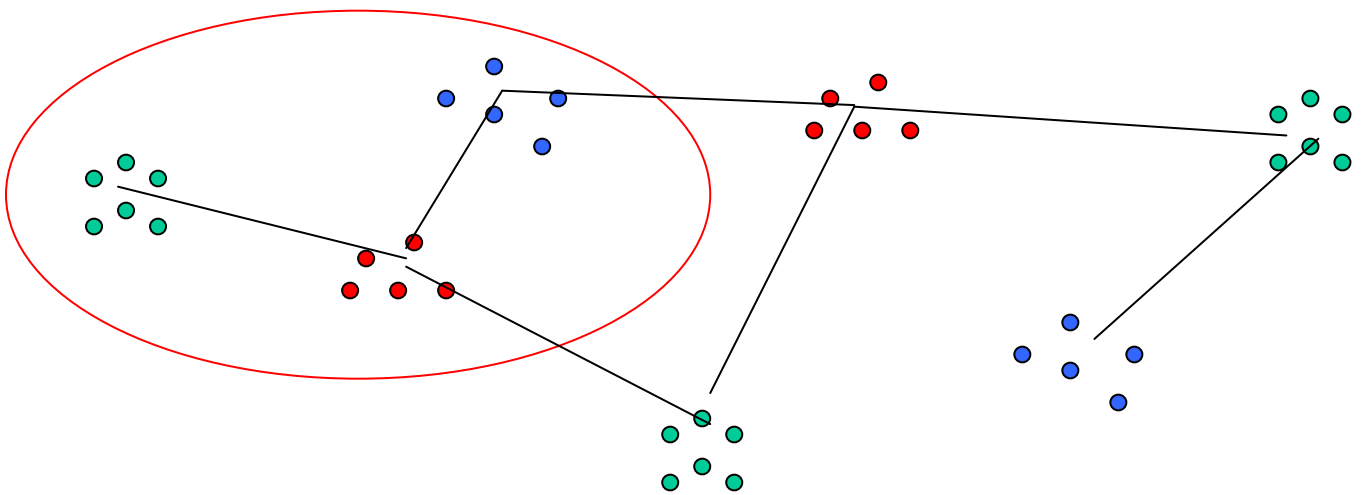
Network model

- Small systems



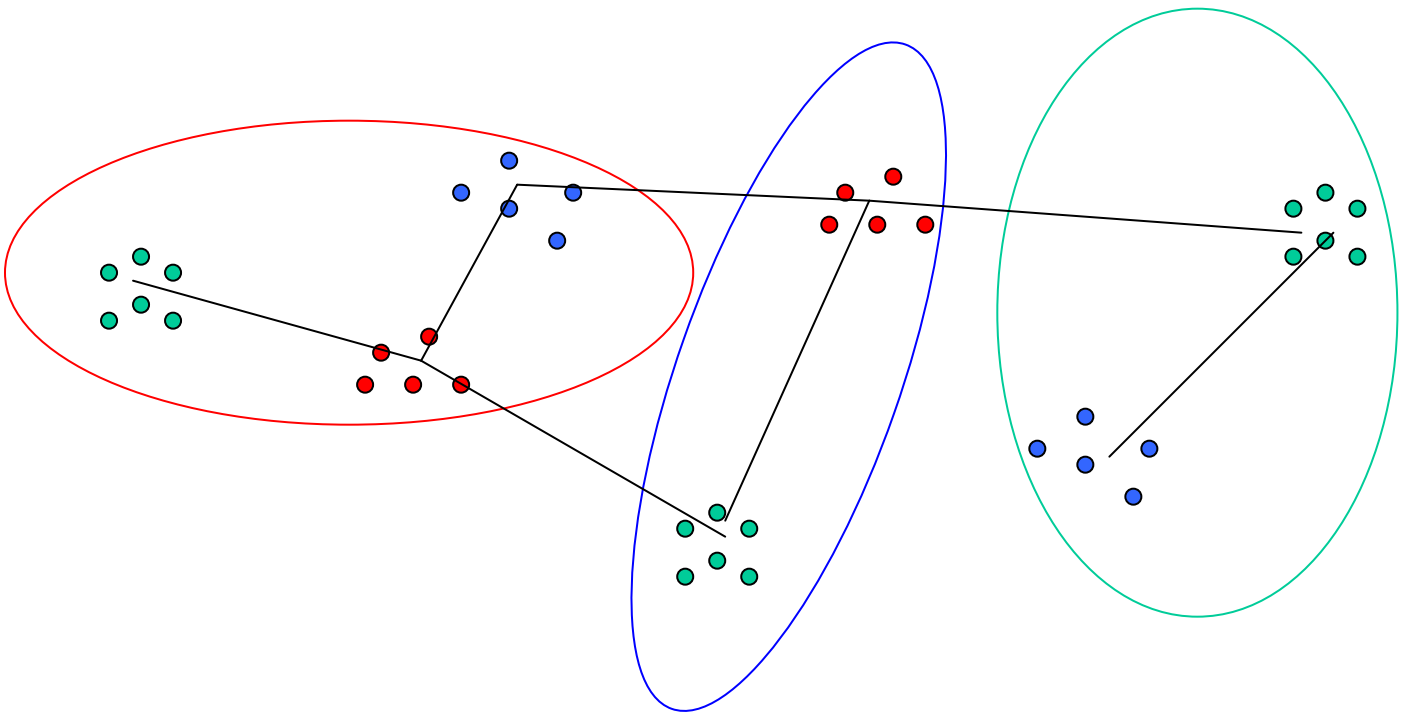
Network model

- Isolated small subsystem

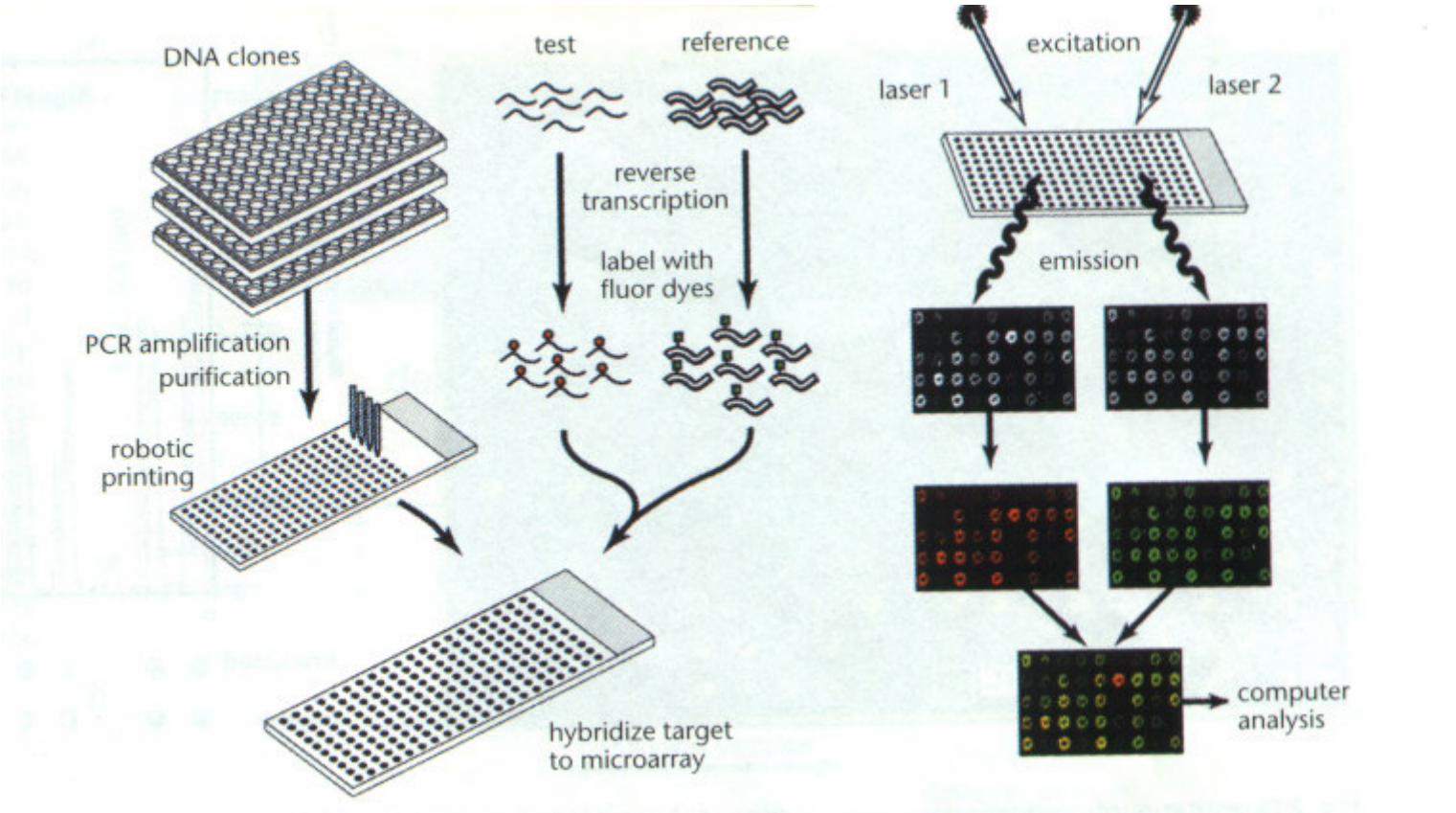


Network model

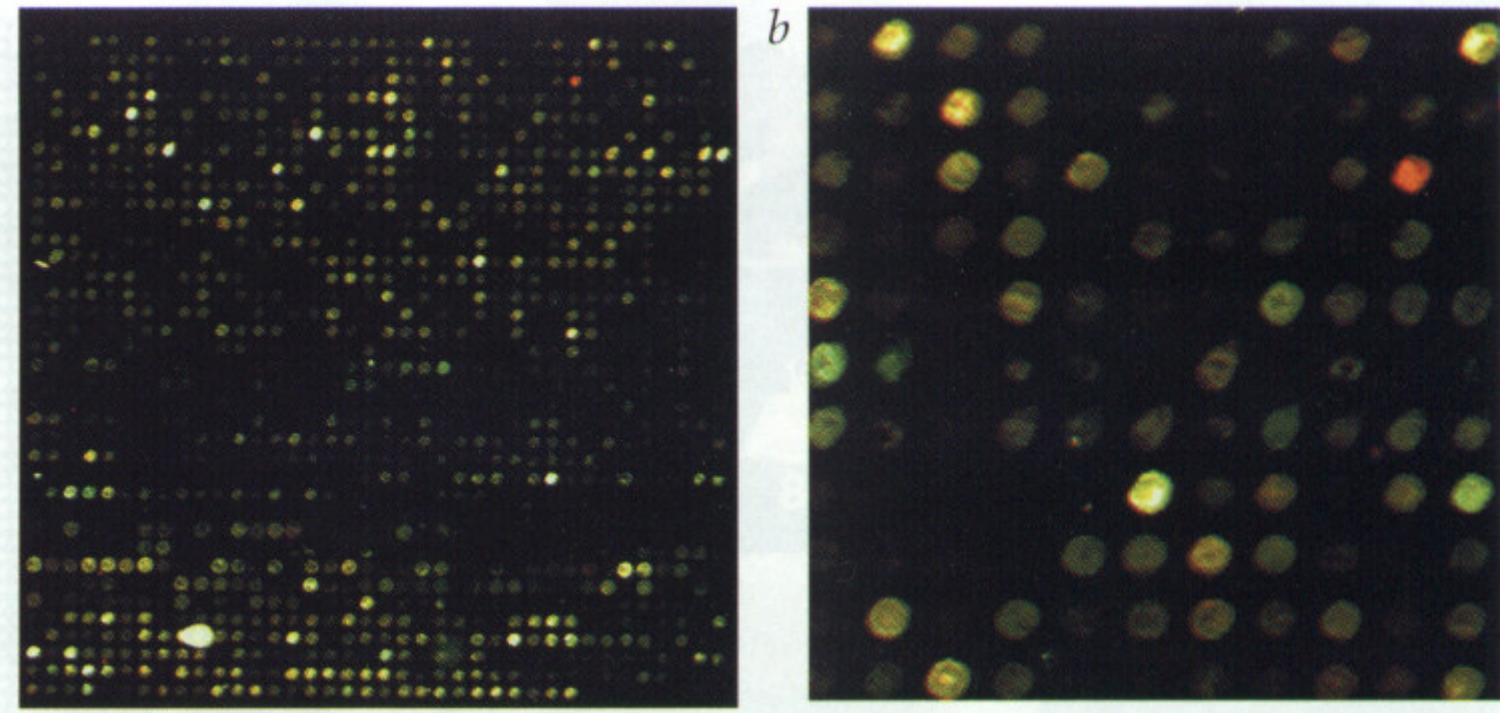
- Systems that have local behavior



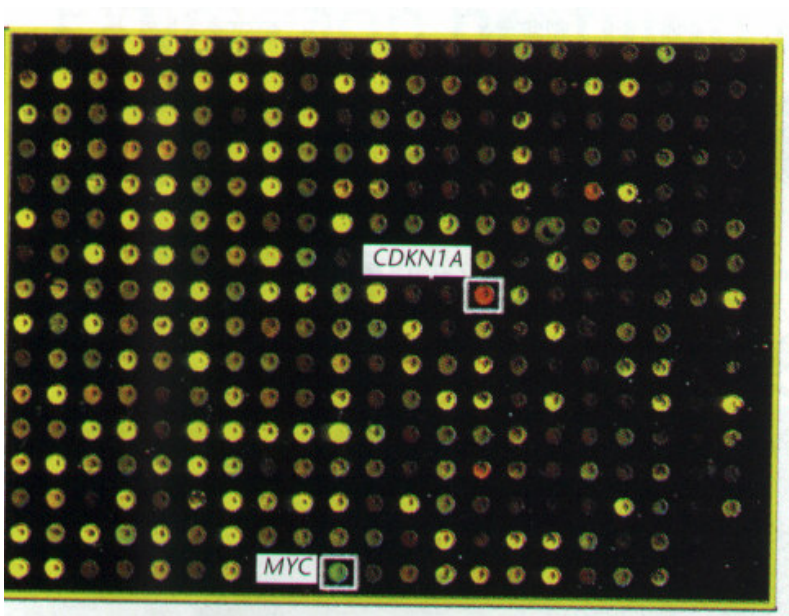
Microarray: the genetic signal

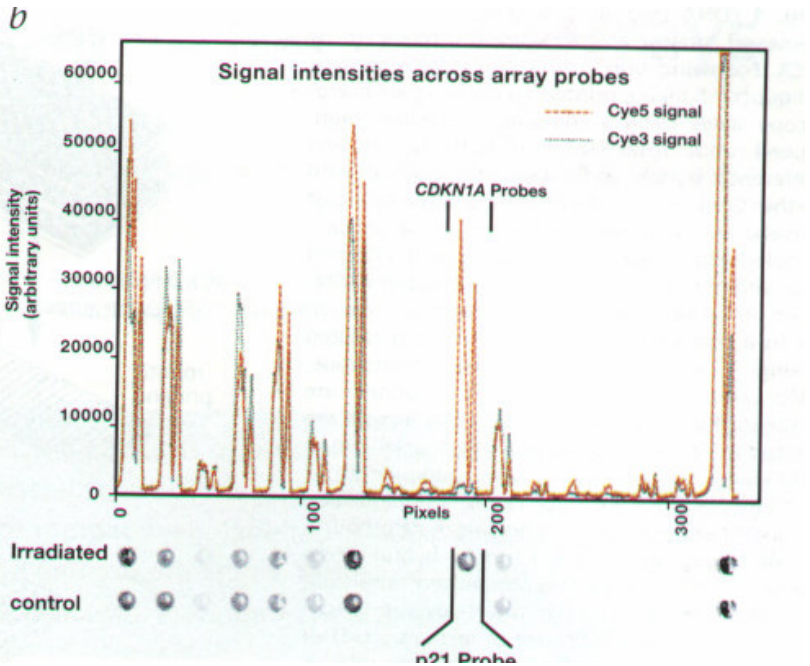


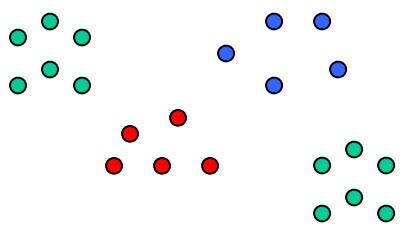
Microarray: the genetic signal



Microarray: the genetic signal







Quantization - {-1,0,1}

Microarray: the genetic signal

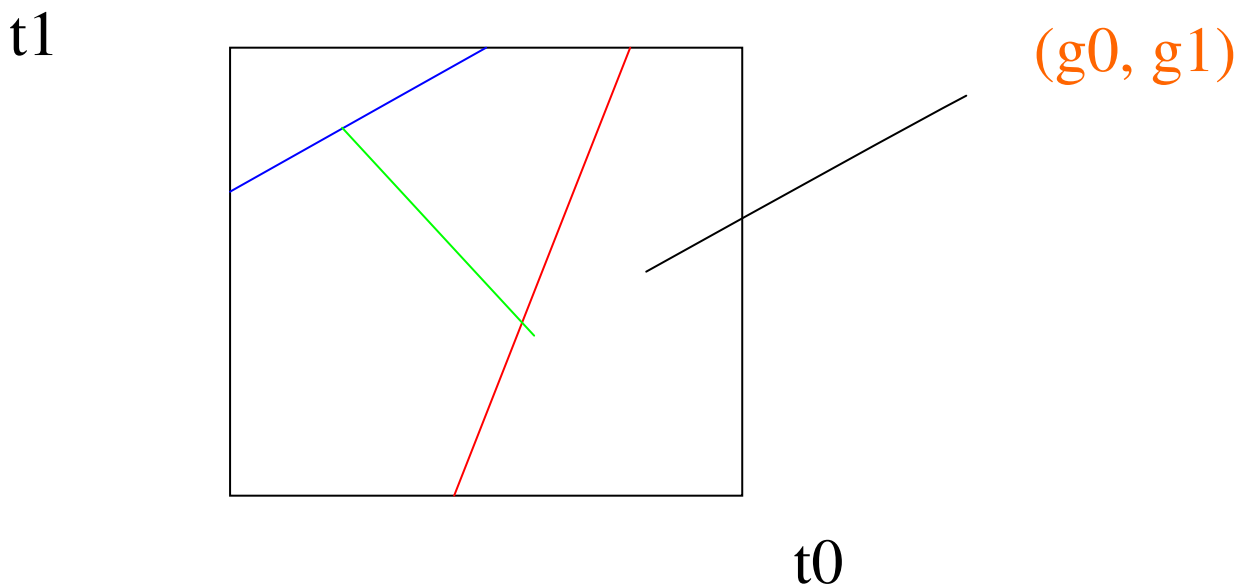
- Resolution: variations of 30% in the hybridizations are perceived
- -1: less than 30%
- 0: between 30% and 60%
- +1: more than 60%
- constraint resolution: limitations in hybridization and visual inspection

Microarray: the genetic signal

- To work with microarrays of the complete genomic system studied
- Nowadays this is not possible for complex systems
- For developing the methodology, it is desired to work with simple known systems
- In some years, data for complex systems should also be complete

Microarray: the genetic signal

- We should have adequate techniques to choose a minimum meaningful set of genes

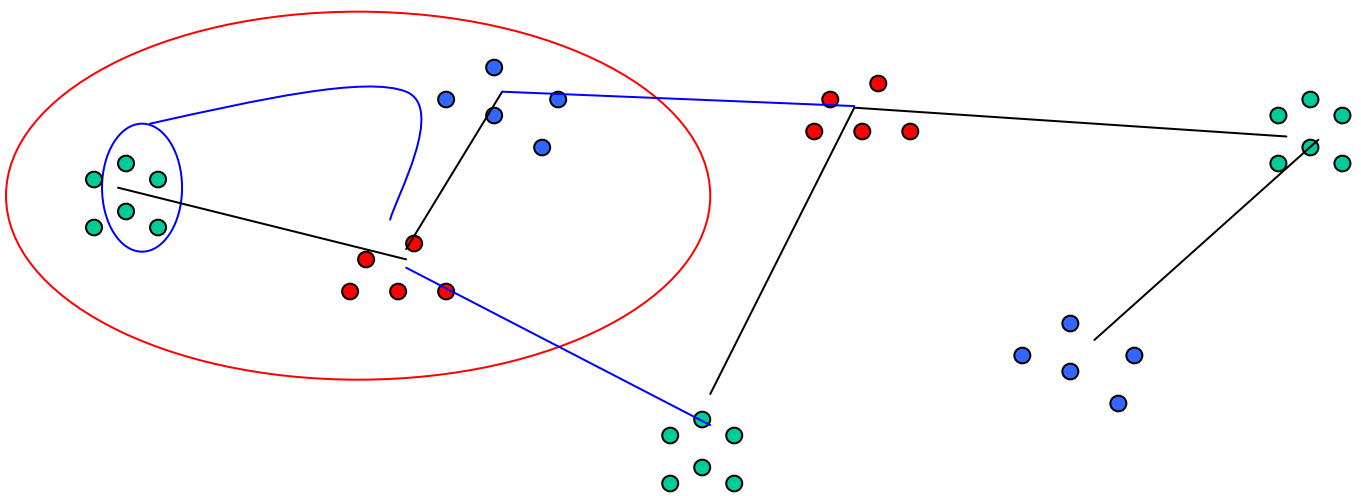


Microarray: the genetic signal

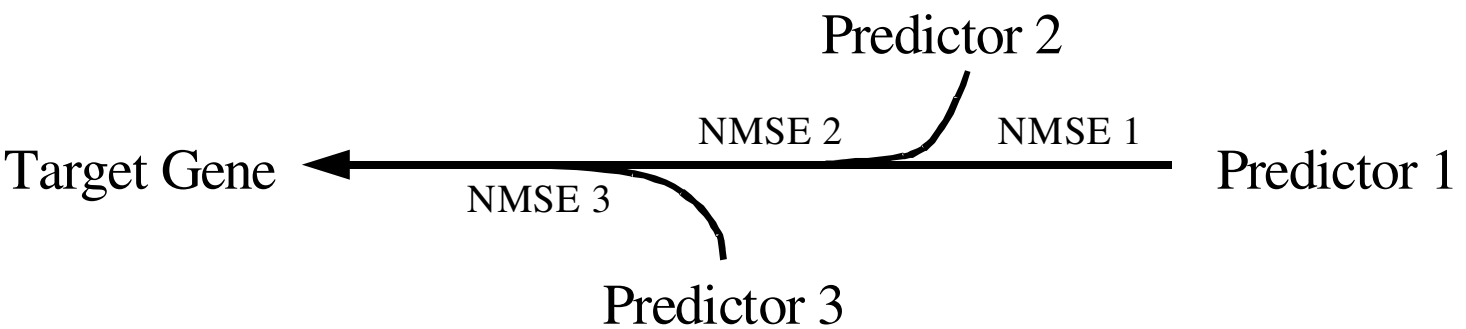
- From microarray data and the knockout biochemical technique the net topology can be identified
- However, a combinatorial number of knockouts would be necessary, what is not feasible
- The design of predictors may put this number in a more feasible value.

Network identification

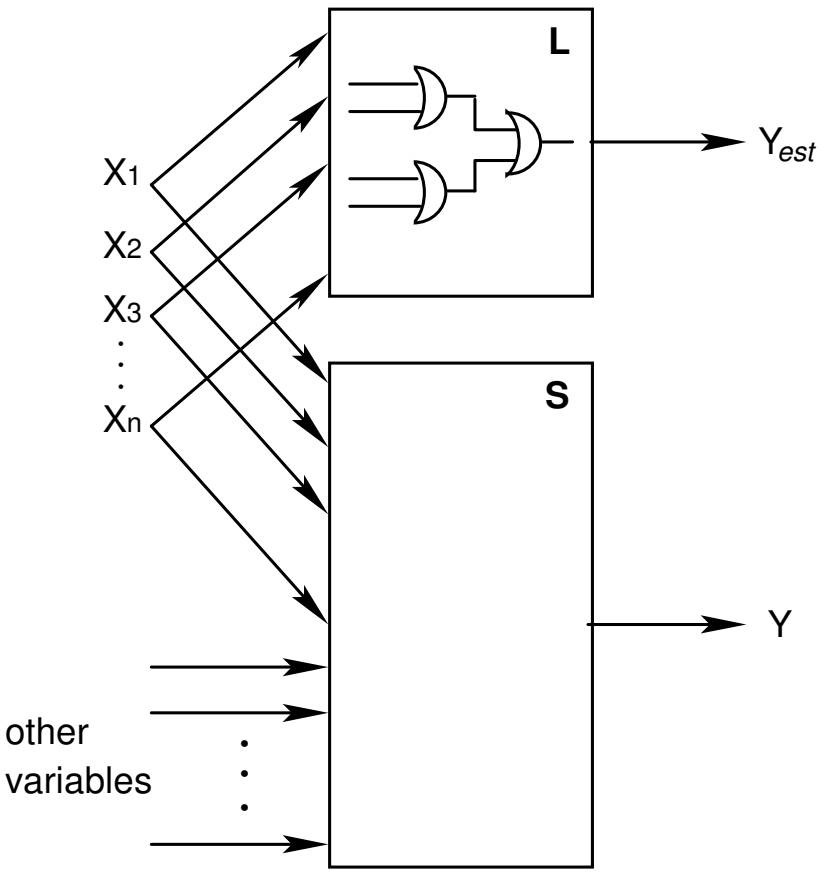
- Non isolated small subsystem
- Predict the states in t_{i+1} from the states in t_i



Network identification



Network identification



Network identification

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

Network identification

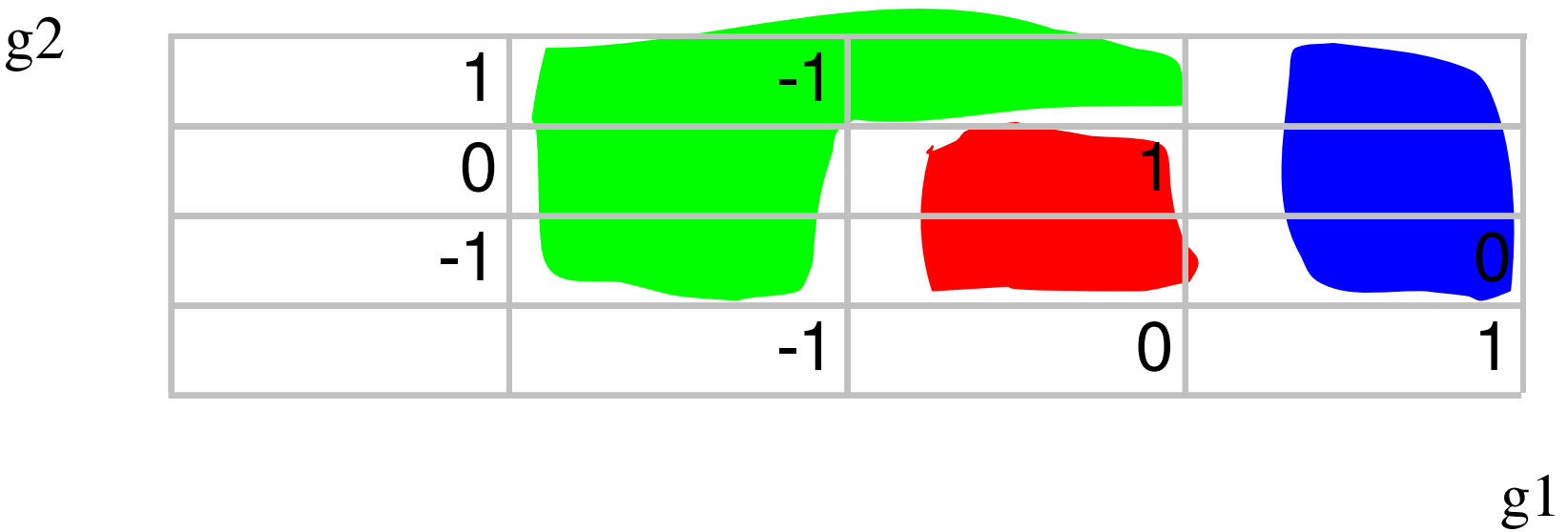
		Genes															Condition		
Cell line	Condition	*RCH1*	*BCL3*	*FRA1*	*REL-B*	*ATF3*	*IAP-1*	*PC-1*	*MBP-1*	*SSAT*	*MDM2*	*p21*	*p53*	*AHA*	*OHO*		*IR*	*MMS*	*UV*
ML-1	IR	-1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	0	0
ML-1	MMS	0	0	0	0	1	0	0	0	0	1	1	1	1	0		0	1	0
Molt4	IR	-1	0	0	1	1	0	1	0	0	1	1	1	1	1		1	0	0
Molt4	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.

Network identification

- 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

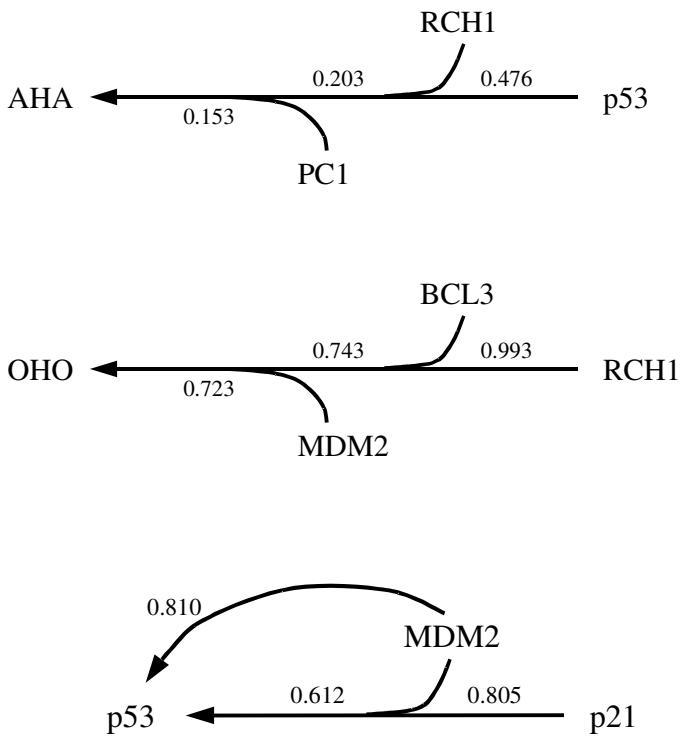
g3 = f(g1,g2)



Network identification

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

Network identification



Network identification

- 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

Network identification

- Preliminary results motivated new experiments:
- They want to scan about 100 genes
- 100 new microarrays are in preparation
- a team of software engineers was formed
- a supercomputer is available

Computational Learning

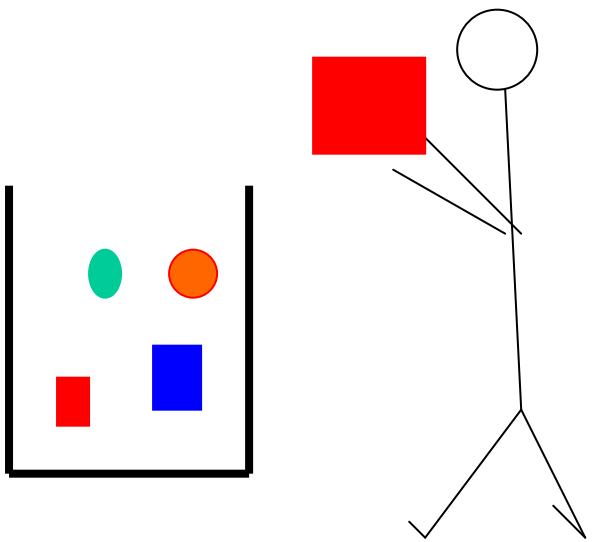
- domain: objects with a random distribution
- concept: a set of objects of a given domain
- teacher: he says if a generic object satisfies the concept, but he may make mistakes

Computational Learning

- example: an object classified by the teacher
- learning algorithm: gives an hypothesis for the concept from a collection of examples
- training data: examples used in the learning algorithm

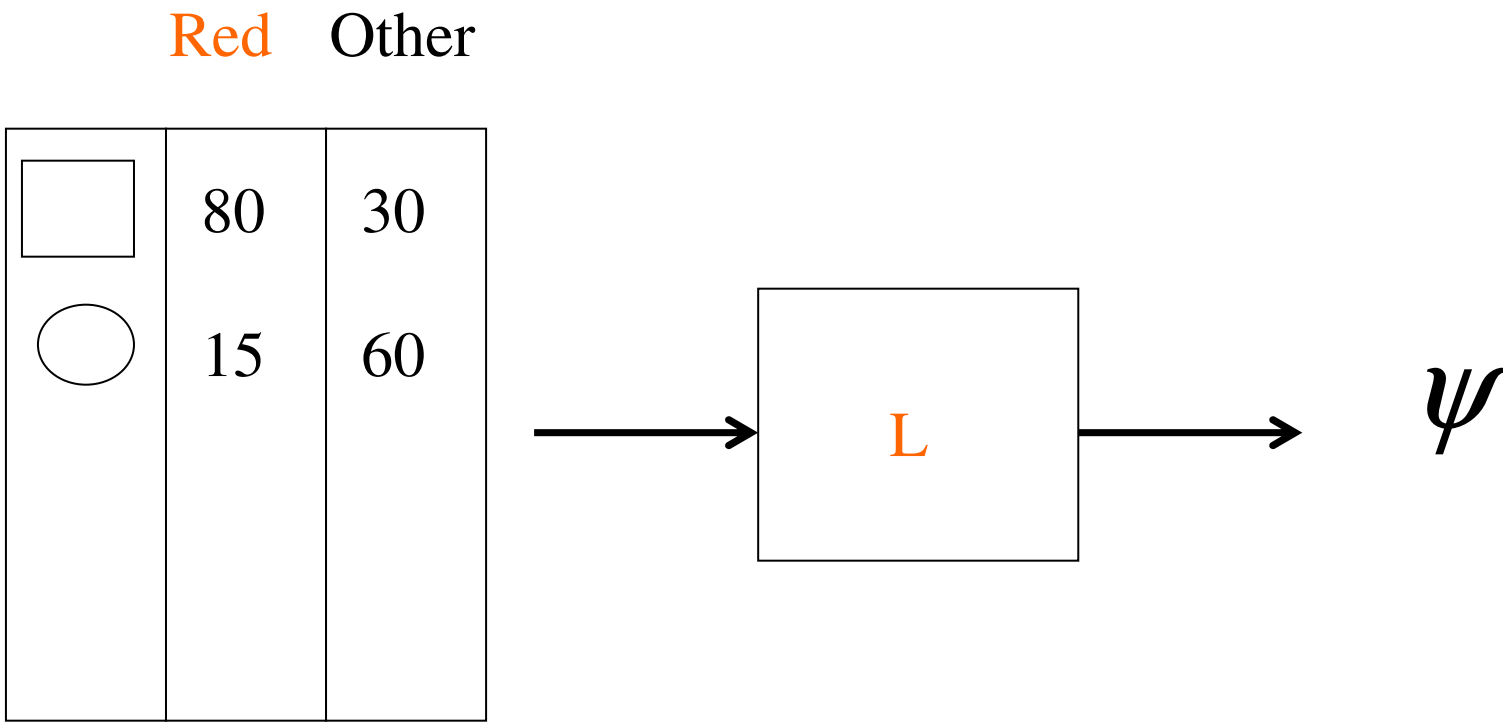
Computational Learning

- Domain: planar shapes
- concept: color RED



	Red	Other
□	80	30
○	15	60

Computational Learning



Computational Learning

- application: the designed function should be applied in the same statistical context of training
- precision: depends on the amount of available training data
- new experiments: repetitions of the design should give operators with similar quality

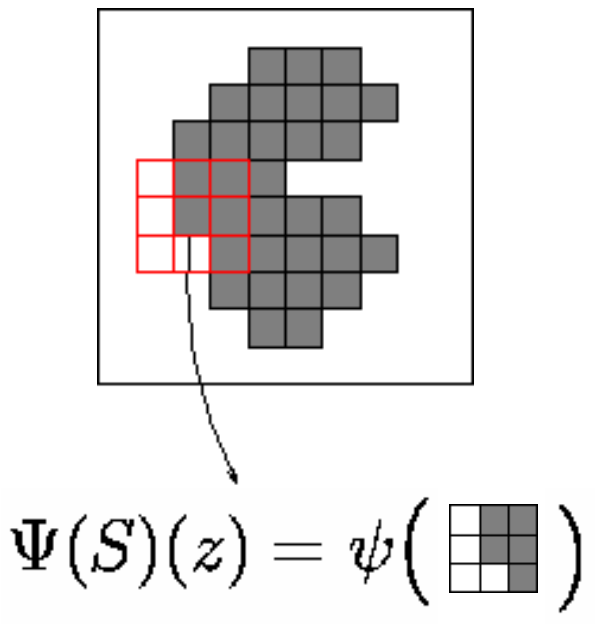
Computational Learning

- table: the number of lines in the table grows exponentially with the number of variables
- Generalization: non observed shapes should also be classified properly
- prior information: rules to generalize the results

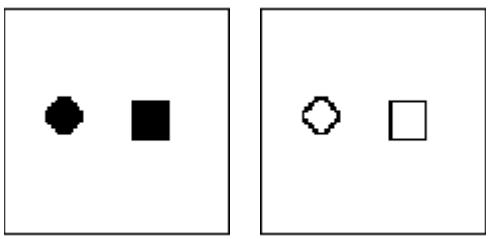
Genetic signal x digital images

- images have a well known topological structure
- images operators have interesting local and translation invariant properties

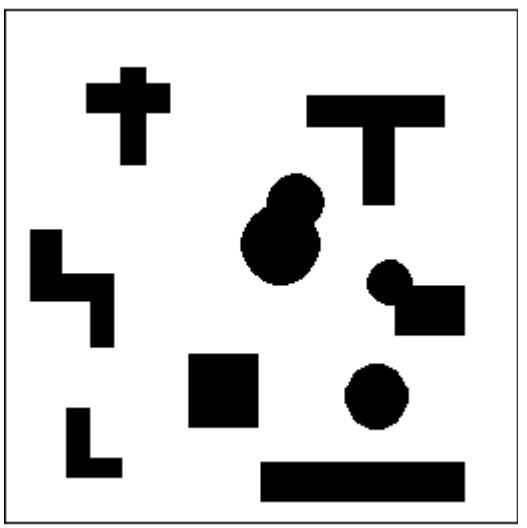
Image operators designed by learning

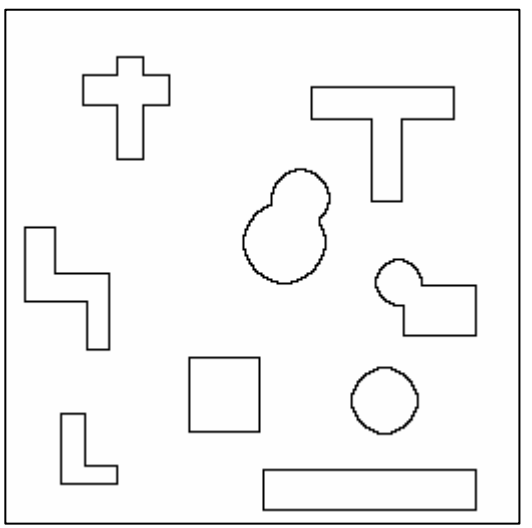


Edge Detection



Training images

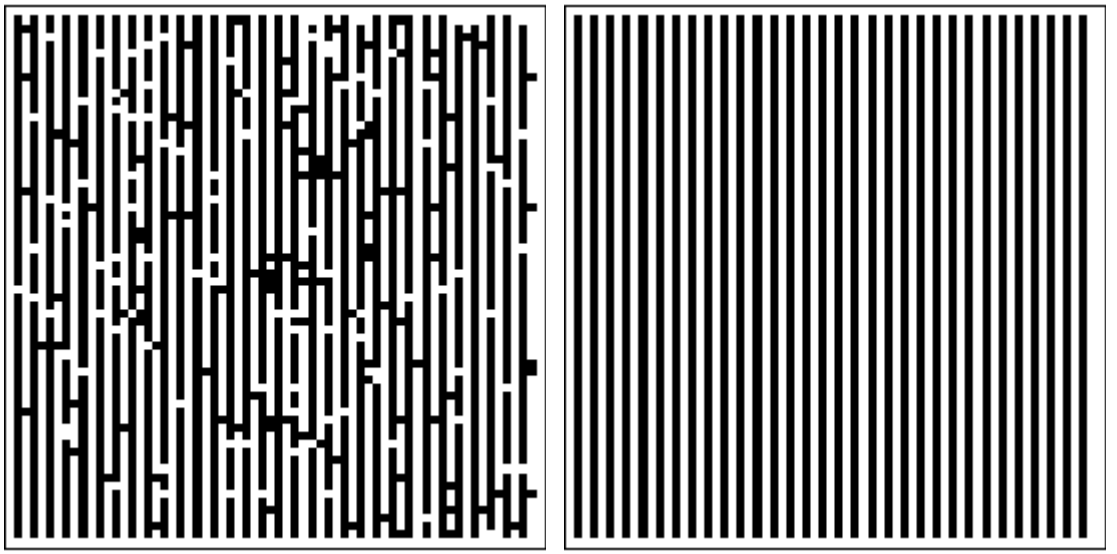




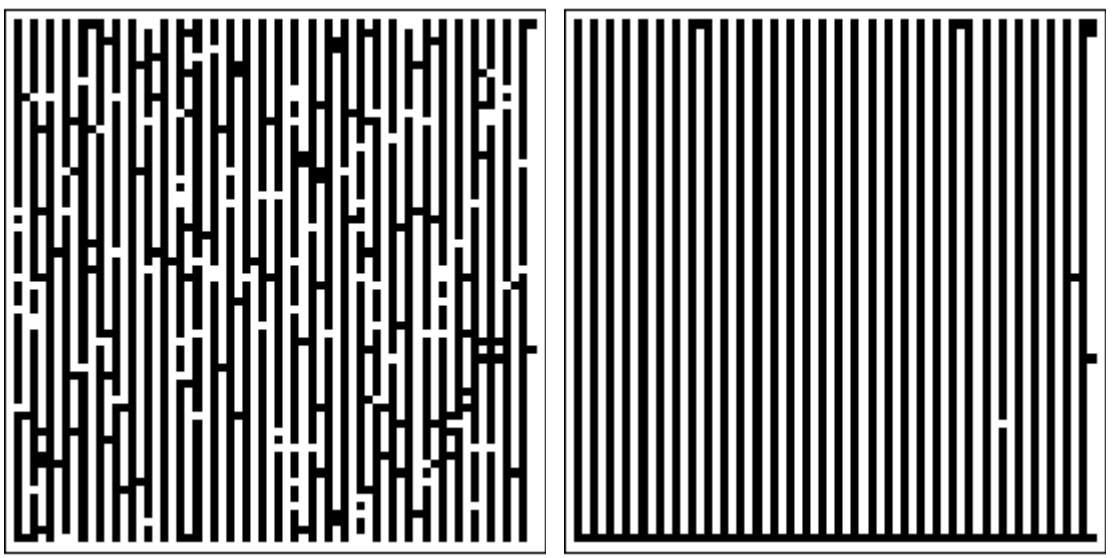
Test images

Noise filtering

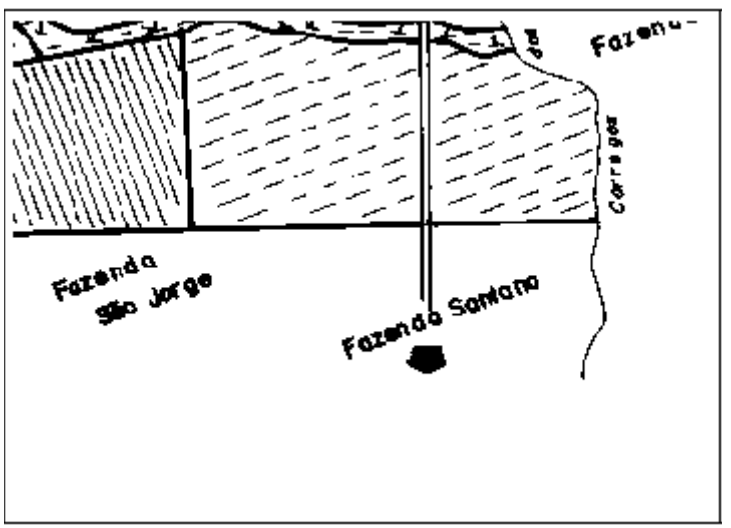
Training images

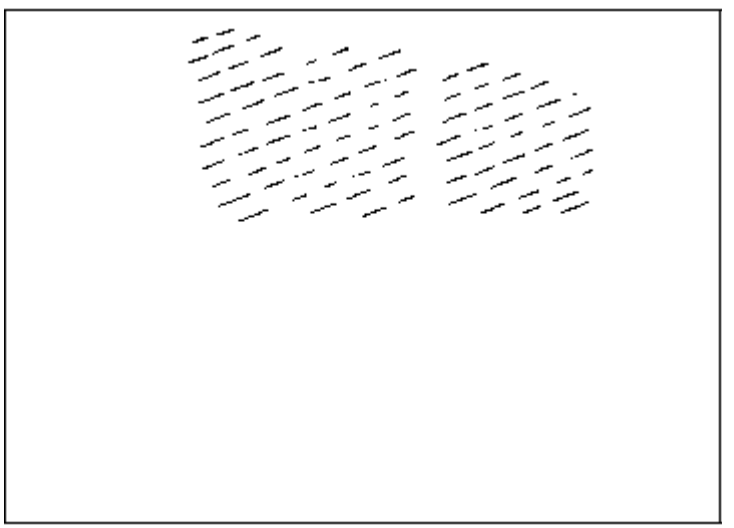


Test images

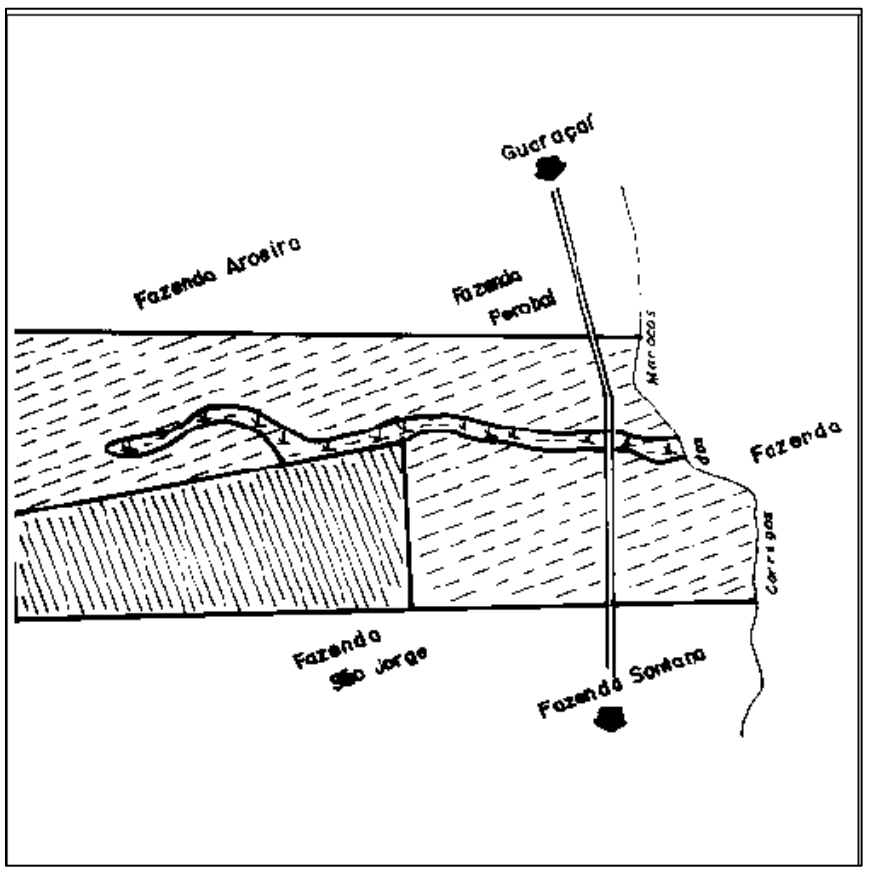


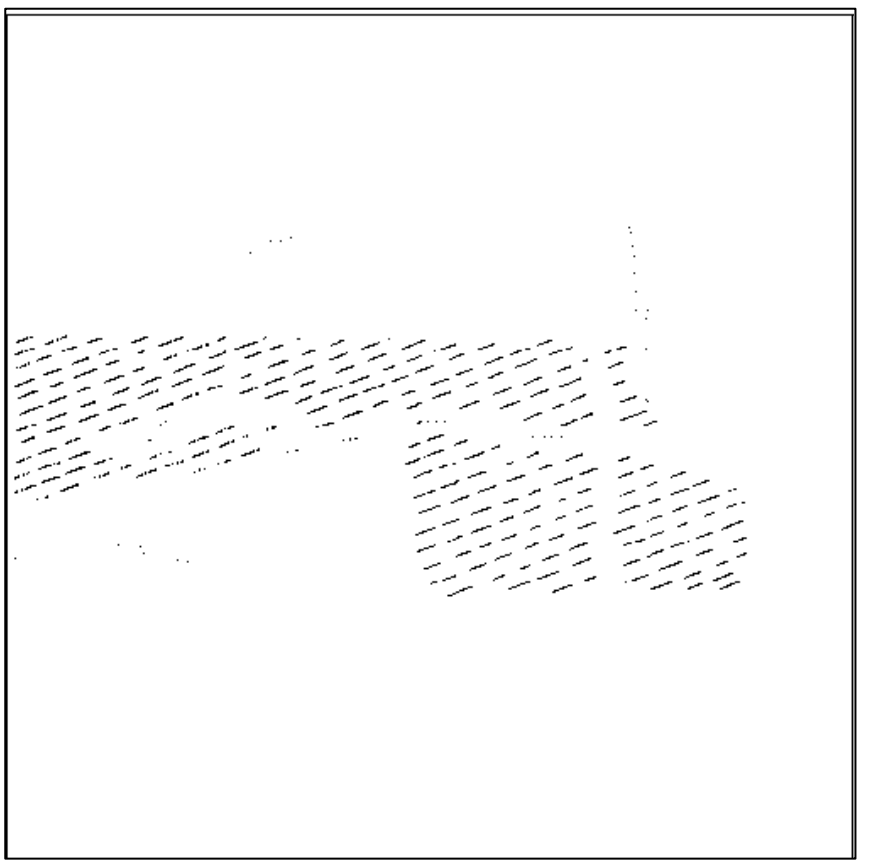
Texture detection





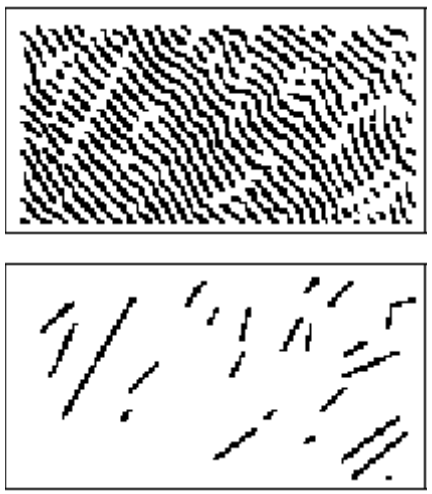
Training images


Guaraçaí
Fazenda Aroeira
Fazenda
Perobal
Fazenda
Fazenda
São Jorge
Fazenda Santana

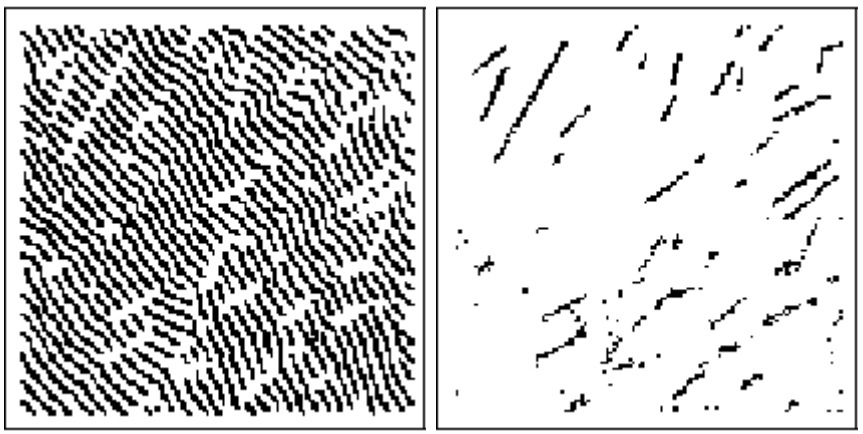


Test images

Fracture detection



Training images



Test images