

Theoretical limitations of massively parallel biology

Genetic network analysis – gene and protein expression measurements

Zoltan Szallasi

Children's Hospital

Informatics Program

Harvard Medical School

www.chip.org

Vipul Periwal

Gene Network Sciences Inc.

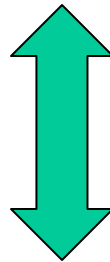
Mattias Wahde

Chalmers University

John Hertz (Nordita)

Greg Klus (USUHS)

How much information is needed to solve a given problem ?



How much information is (or will be) available ?

Conceptual limitations

Practical limitations

- **Finding transcription factor binding sites based on primary sequence information**
- **SNP < > disease association**

What are the problems we want to solve ?

So far the “DNA chip” revolution has been mainly technological:

The principles of measurements (e.g. complementary hybridization) have not changed.

It is not clear yet whether a conceptual revolution is approaching as well ?

potential breakthrough questions:

- can we perform efficient, non-obvious reverse engineering ?

- can we identify non-dominant cooperating factors ?

- can we predict truly new subclasses of tumors based on gene expression patterns ?

- can we perform meaningful (non-obvious & predictive)

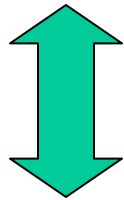
forward modeling

- 1. Reverse engineering time series measurements**
- 2. Identification of novel classes or separators in gene expression matrices in a statistically significant manner**
- 3. Potential use of artificial neural nets (machine learning) in the analysis of gene expression matrices.**

Biological research has been based on the discovery of strong dominant factors.

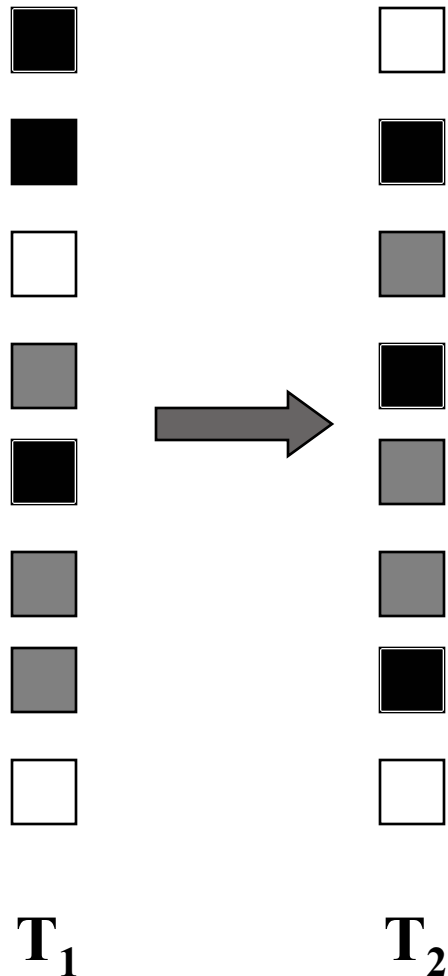
More than methodological issue ?

Robust network based on stochastic processes



Strong dominant factors

The Principle of Reverse Engineering of Genetic Regulatory Networks from time series data:



Determine a set of regulatory rules that can produce the gene expression pattern at T_2 given the gene expression pattern at the previous time point T_1

Continuous modeling:

$$\mathbf{x}_i(t+1) = g(\mathbf{b}_i + \sum_j \mathbf{w}_{ij} \mathbf{x}_j(t))$$

**(Mjolsness et al, 1991 - connectionist model;
Weaver et al., 1999, - weight matrix model;
D'Haeseleer et al., 1999, - linear model;
Wahde & Hertz, 1999 - coarse-grained reverse engineering)**

**at least as many time points as genes: $T-1 > N+2$
(Independently regulated entities)**

For differential equations with r parameters $2r+1$ experiments are enough for identification (E.D.Sontag, 2001)

How much information is needed for reverse engineering?

Boolean fully connected

$$2^N$$

Boolean, connectivity K

$$K 2^K \log(N)$$

Boolean, connectivity K, linearly separable rules

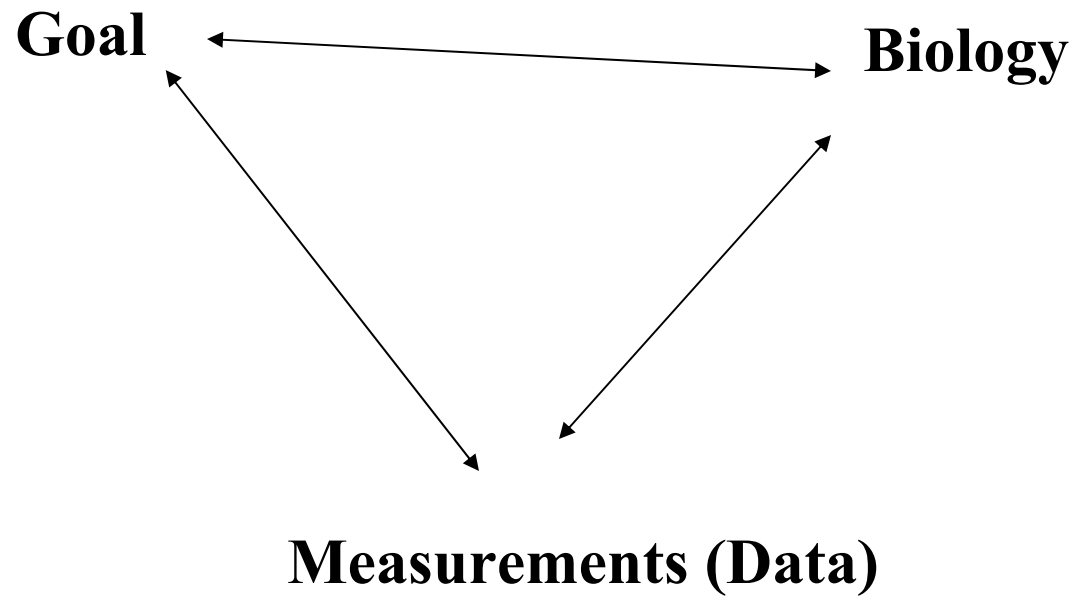
$$K \log(N/K)$$

Pairwise correlation

$$\log(N)$$

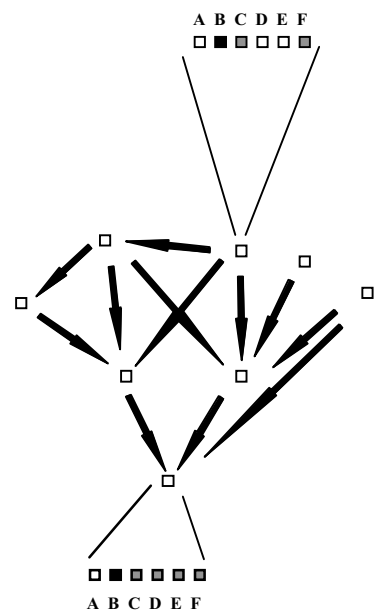
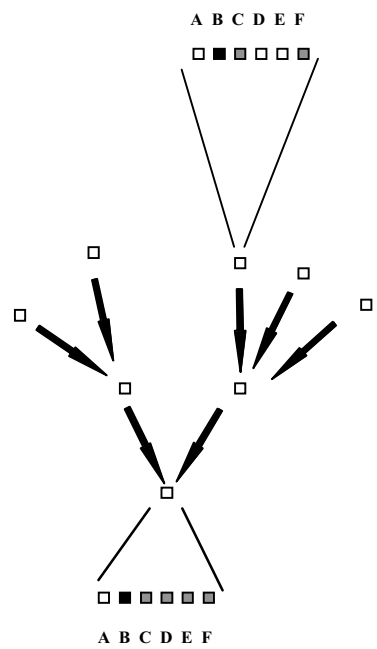
N = number of genes

K = average regulatory input/gene



Biological factors that will influence our ability to perform successful reverse engineering.

- (1) the stochastic nature of genetic networks ,**
- (2) the effective size of genetic networks ,**
- (3) the compartmentalization of genetic networks,**



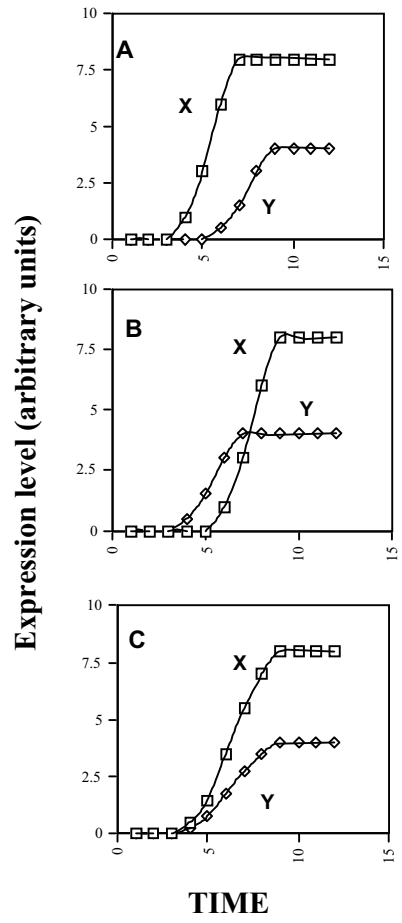
1. The prevailing nature of the genetic network

The effects of stochasticity:

1. It can conceal information (How much ?)

**2. The lack of sharp switch on/off kinetics
can reduce useful information of gene
expression matrices.**

(For practical purposes genetic networks might be considered as deterministic systems ?)



2. The effective size of the genetic network:

How large is our initial directed graph ?

(It is probably not that large.)

We might have a relatively well defined deterministic cellular network with not more than 10 times the number of total genes.

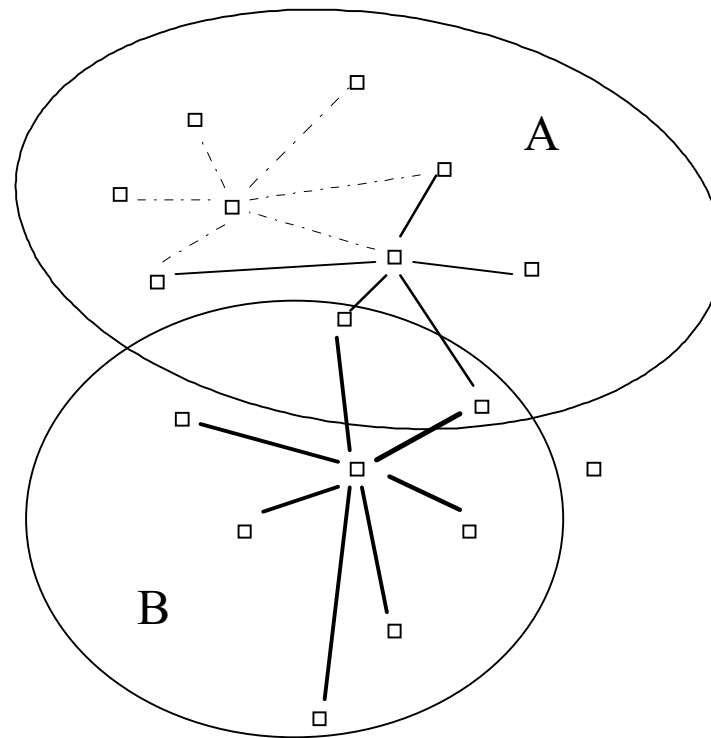
$$N_{\text{bic}} < 10 \times N_{\text{gene}}$$

10,000-20,000 active genes per cell

Splice variants < > modules

3. The compartmentalization (modularity) **of the genetic network:**

The connectivity of the initial directed gene network graph
Low connectivity - better chance for computation.



Genetic networks exhibit:

Scale-free properties (Barabasi et al.)

Modularity

Flatness

**(Useful) Information content of measurements
is influenced by the inherent nature of living
systems**

**We can sample only a subspace of all gene
expression patterns (gene expression space),
because:**

**1. the system has to survive
(83% of the genes can be knocked out in *S. cerevisiae*)**

**2. Gene-expression matrices (i.e. experiments)
are coupled**

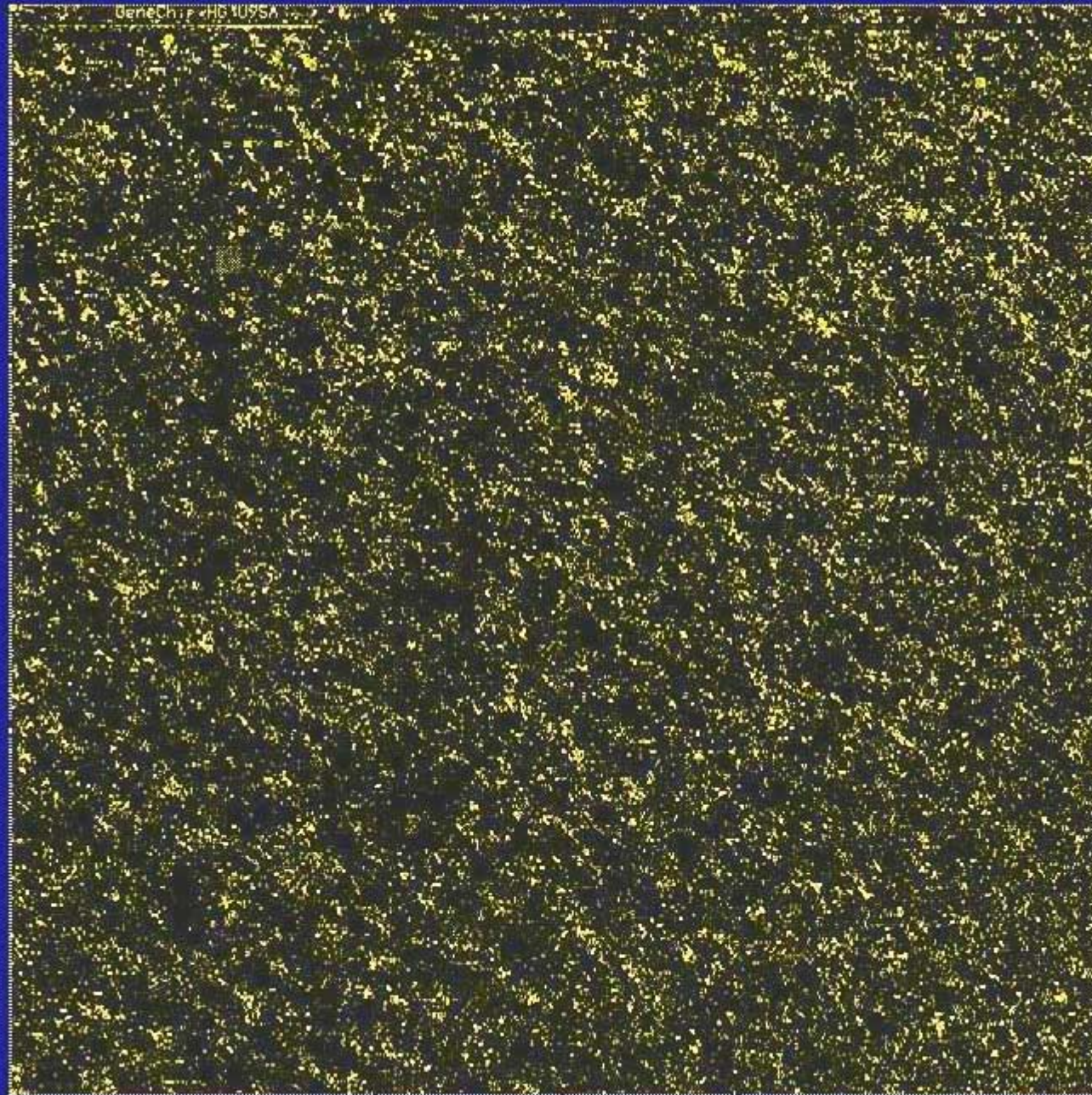
Cell cycle of yeast under different conditions

Data:

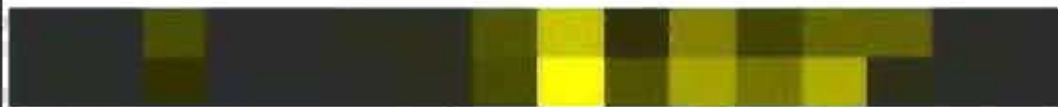
A reliable detection of 2-fold differences seems to be the practical limit of massively parallel quantitation.

(estimate: optimistic and not cross-platform)

Population averaged measurements

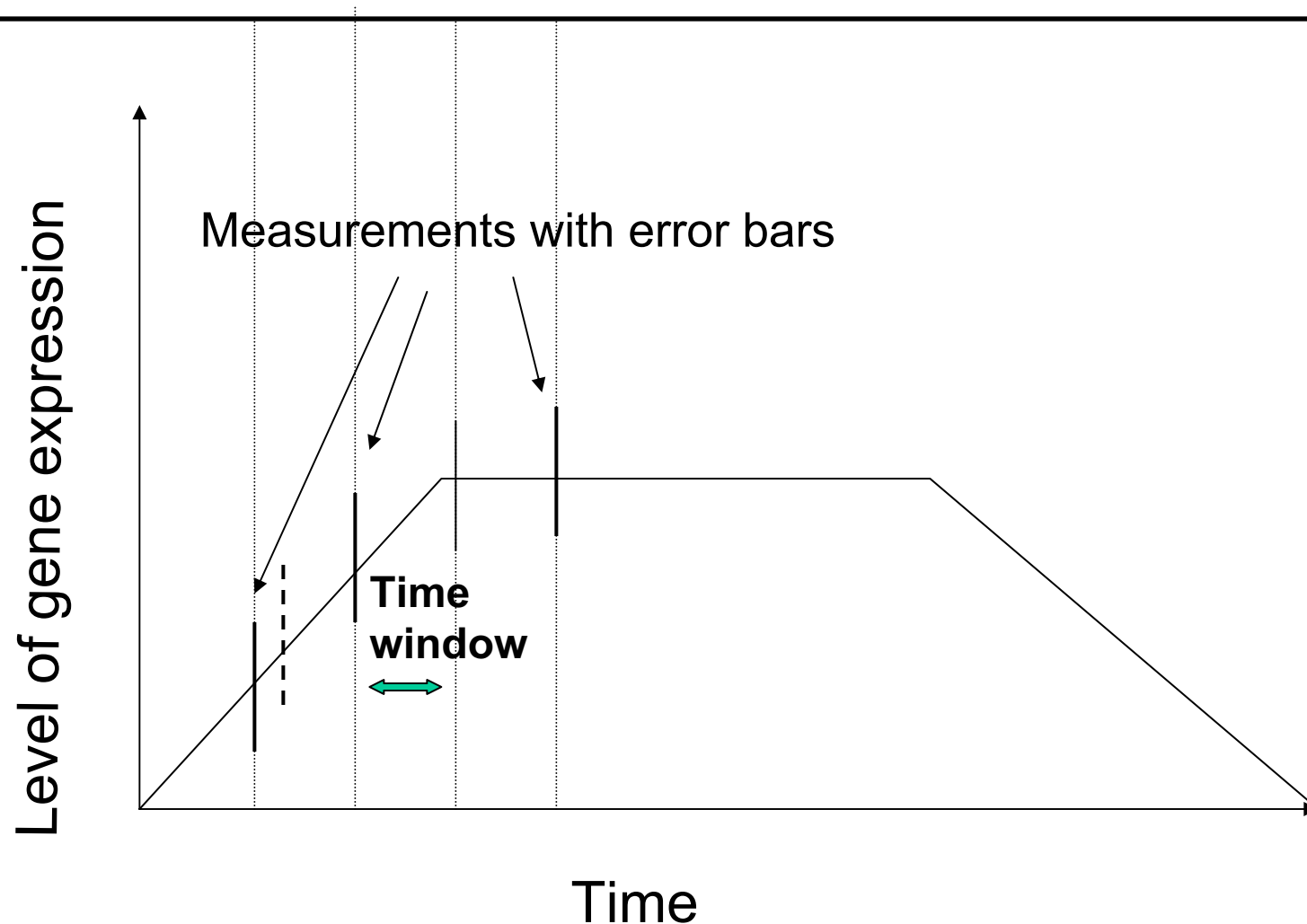


[Gene 5000 33700_all] (#196 C)



The useful information content of time series measurements depend on:

- 1. Measurement error (conceptual and technical limitations, such as normalization)**
- 2. Kinetics of gene expression level changes (lack of sharp switch on/off kinetics - stochasticity ?)**
- 3. Number of genes changing their expression level.**
- 4. The time frame of the experiment.**



A rational experiment will sample gene-expression according to a time-series in which each consecutive time point is expected to produce at least as large expression level difference as the error of measurement: approximately 5 min intervals in yeast, 15-30 min intervals in mammalian cells.

$$P = K \log(N/K) \text{ (John Hertz, Nordita)}$$

P : gene expression states

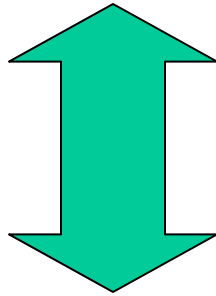
N: size of network

K: average number of regulatory interactions

Applying all this to cell cycle dependent gene expression measurements by cDNA microarray one can obtain 1-2 orders of magnitude less information than expected in an ideal situation. (Szallasi, 1998)

Can we identify non-dominant cooperating factors ?

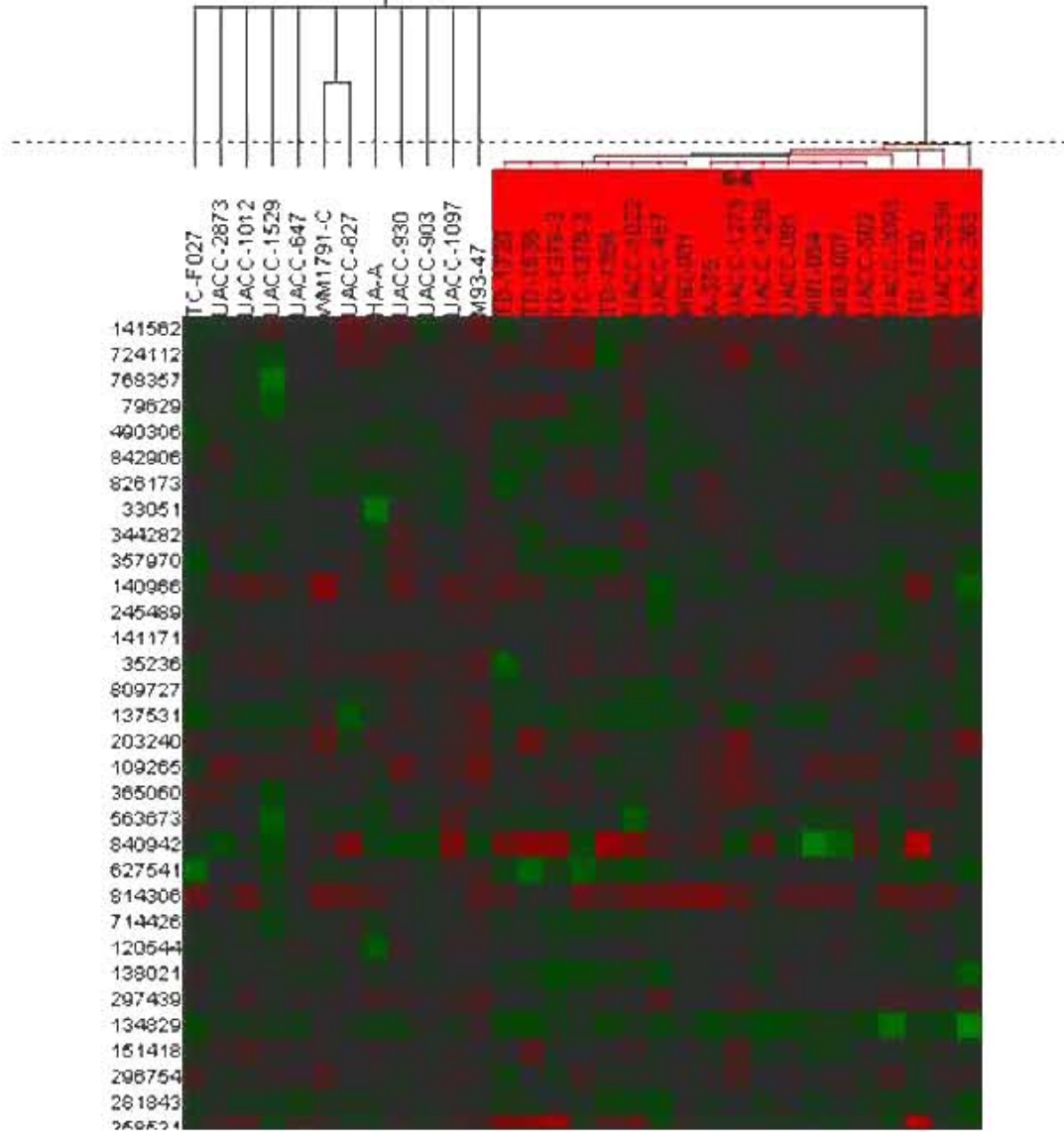
Can we predict truly new subclasses of tumors based on gene expression patterns ?



How much data is needed ?

How much data will be available ?

Samples



Analysis of massively parallel data sets

Unsupervised - avoiding artifacts in random data sets

**avoiding artifacts in data sets retaining
the internal data structure**

Supervised

INFORMATION REQUIREMENT

Consistently mis-regulated genes in random matrices

“E” different samples

“N”-gene microarray

M_i genes mis-regulated in the “i”-th sample,

K consistently mis-regulated across all E samples.

What is the probability that (at least) K genes were mis-regulated by chance ?

$$P(E, k \geq K) = 1 - \sum_{i=0}^{K-1} P(E, i)$$

Where $P(E, k)$ is the probability that exactly k genes are consistently mis-regulated

$$P(E, k) = \sum_{j=k}^M \frac{\binom{j}{k} * \binom{N-j}{M-k} * P(E-1, j)}{\binom{N}{M}}$$

$$P(2, k) = \frac{\binom{M}{k} * \binom{N-M}{M-k}}{\binom{N}{M}}$$

If $N \gg M$, then

$$P(E, k \geq K) \approx \frac{\binom{N}{K} * \binom{N-K}{M-K}^E}{\binom{N}{M}^E}$$

or

$$P(E, k \geq K) \approx \binom{N}{k} * (q^E)^k * (1 - q^E)^{N-k}$$

For a K gene separator:

$$n_K \binom{N}{K} * (1 - q^K)^E$$

N	M	E	K	n_K simulated	n_K calculated
500	100	4	3	1172455±123637	1174430
500	100	8	3	69630 ± 17487	66605
300	50	15	3	760 ± 579	785
200	40	20	4	2032 ± 1639	1713

how many cell lines do we need in order to avoid accidental separators ?

for N=10000 M=1000 for $p < 0.001$

K=1 E=7

Higher order separator

K=2 E=15

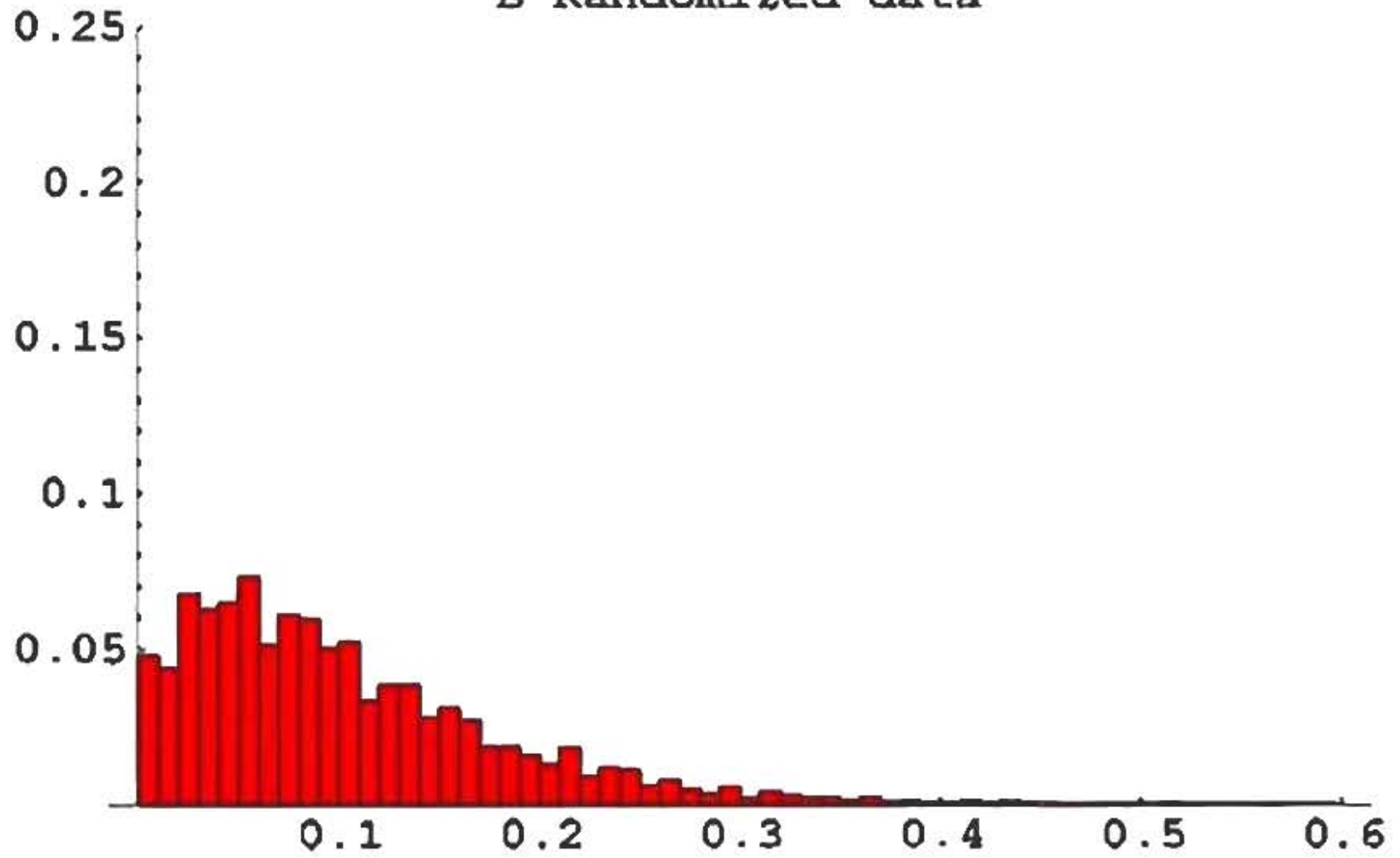
K=3 E=25

K=4 E=38

K=5 E=54

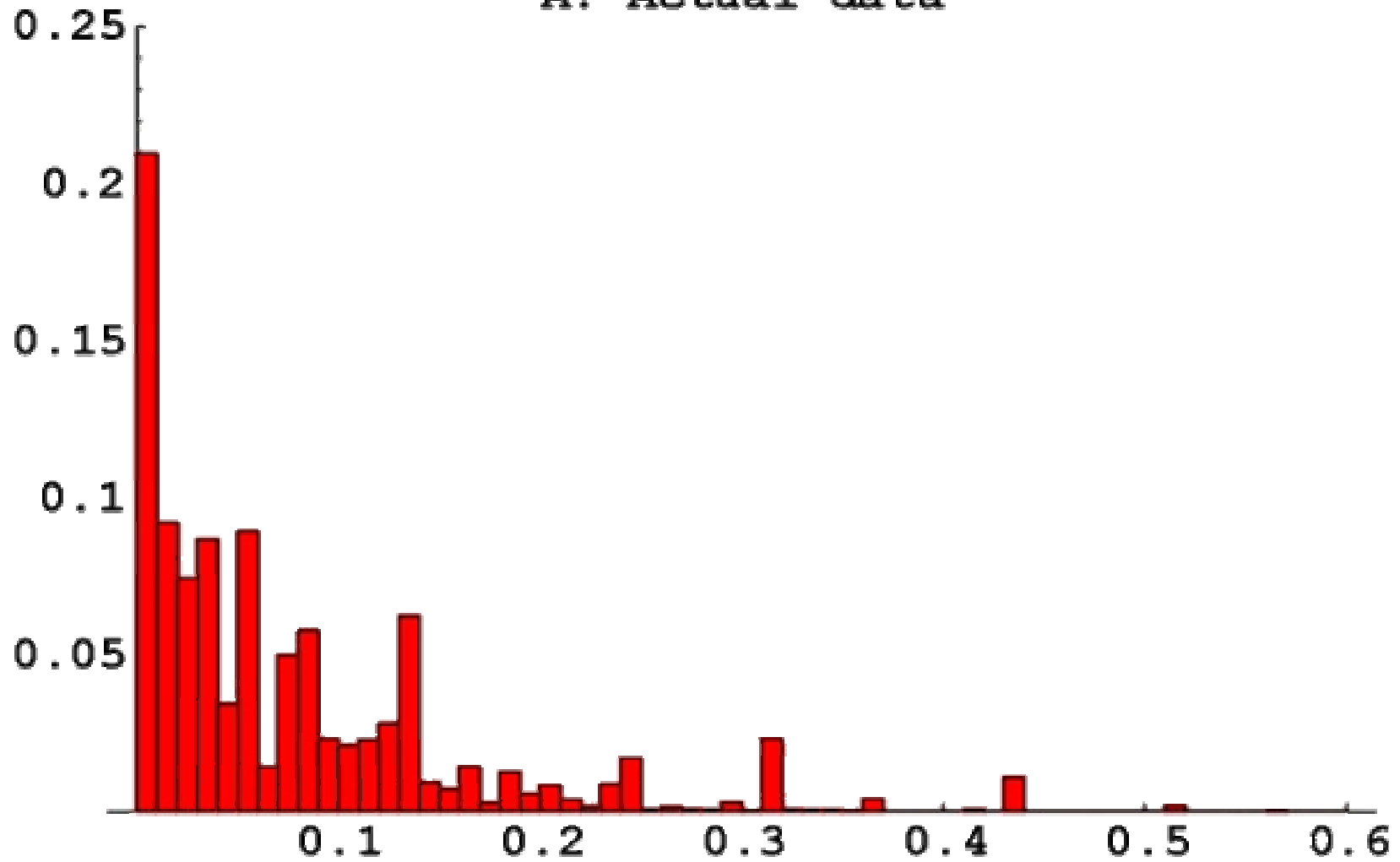
K=6 E=73

B Randomized data



Genes are not independently regulated

A. Actual data



Generative models (gene expression operator) will simulate realistic looking gene expression matrices ?

- the number of genes that can be mis-regulated
- the independence of gene mis-regulation.

	N_1	N_2	N_3	N_i	T_1	T_2	T_3	T_i
gene ₁	0	0	1		0	1	1	1	0
gene ₂	0	0	0		0	0	1	0	1
gene ₃	0	0	0		0	1	0	1	0
gene ₄	0	1	1		0	0	0	1	1

Algorithm to extract Boolean separators from a gene expression matrix.

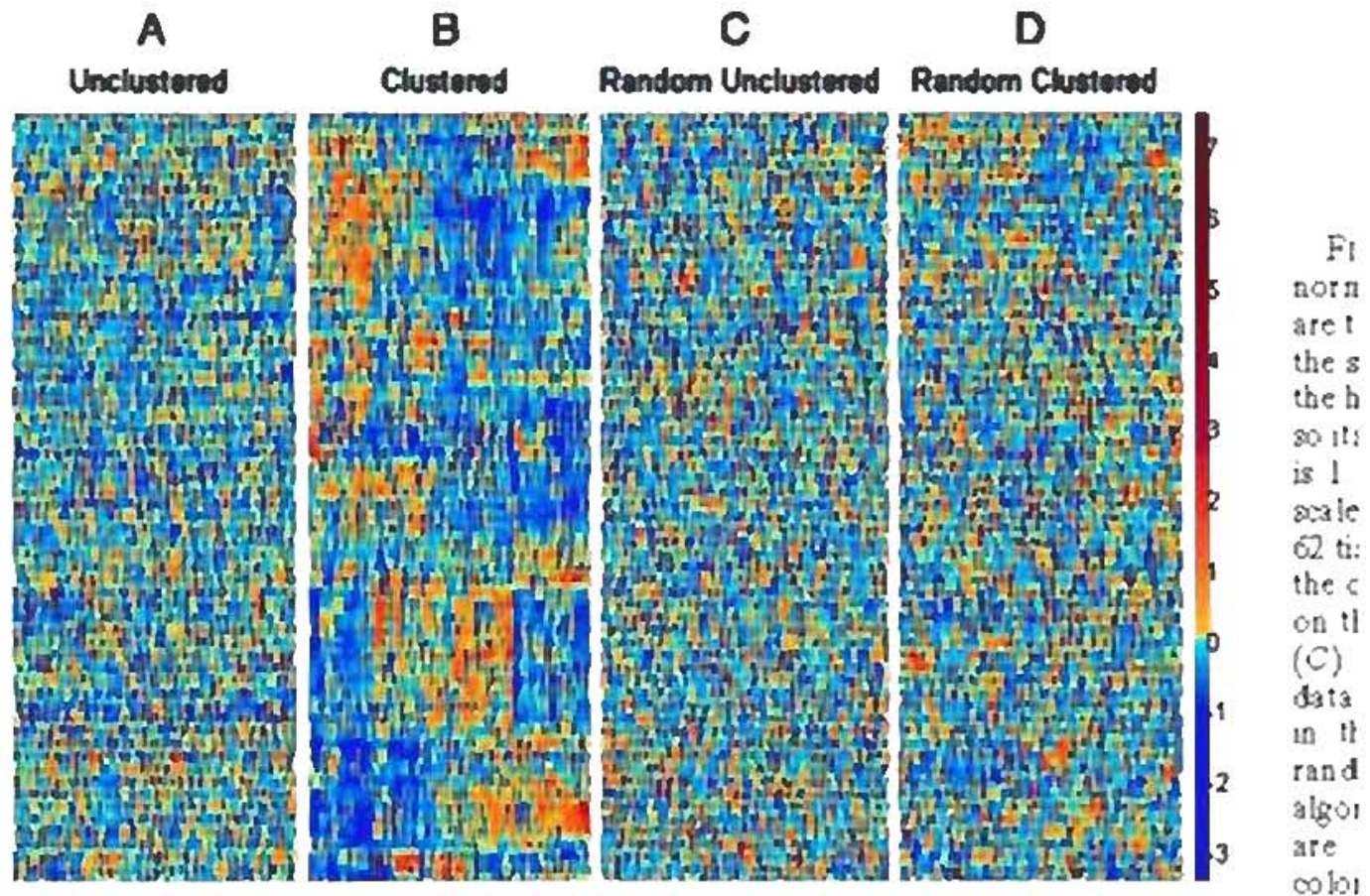
U. Alon data set (colon tumors) : $N=2000$, $M_{\text{average}}=180$ $K=2$

E	Alon data	calc.	Num. sim.
10	708	131	130
11	120	~ 1	1
12	45	8.6×10^{-3}	8.6×10^{-3}
13	3	7.0×10^{-5}	-
14	3	5.6×10^{-7}	-
15	1	4.6×10^{-9}	-
16	1	3.7×10^{-11}	-

Generative model: 4 ± 2 separators

features, while preserving step-like changes in intensity. The features were arranged in the order they appear in the EST sequence, the PM-MM intensities in a moving window of five features were sorted, and the filtered intensity was given by the mean of the middle three sorted intensities. The total intensity

coefficient of variation. The binary tree to reorganize the data at this end, we include a deterministic



Pearson-disproportion of an array:

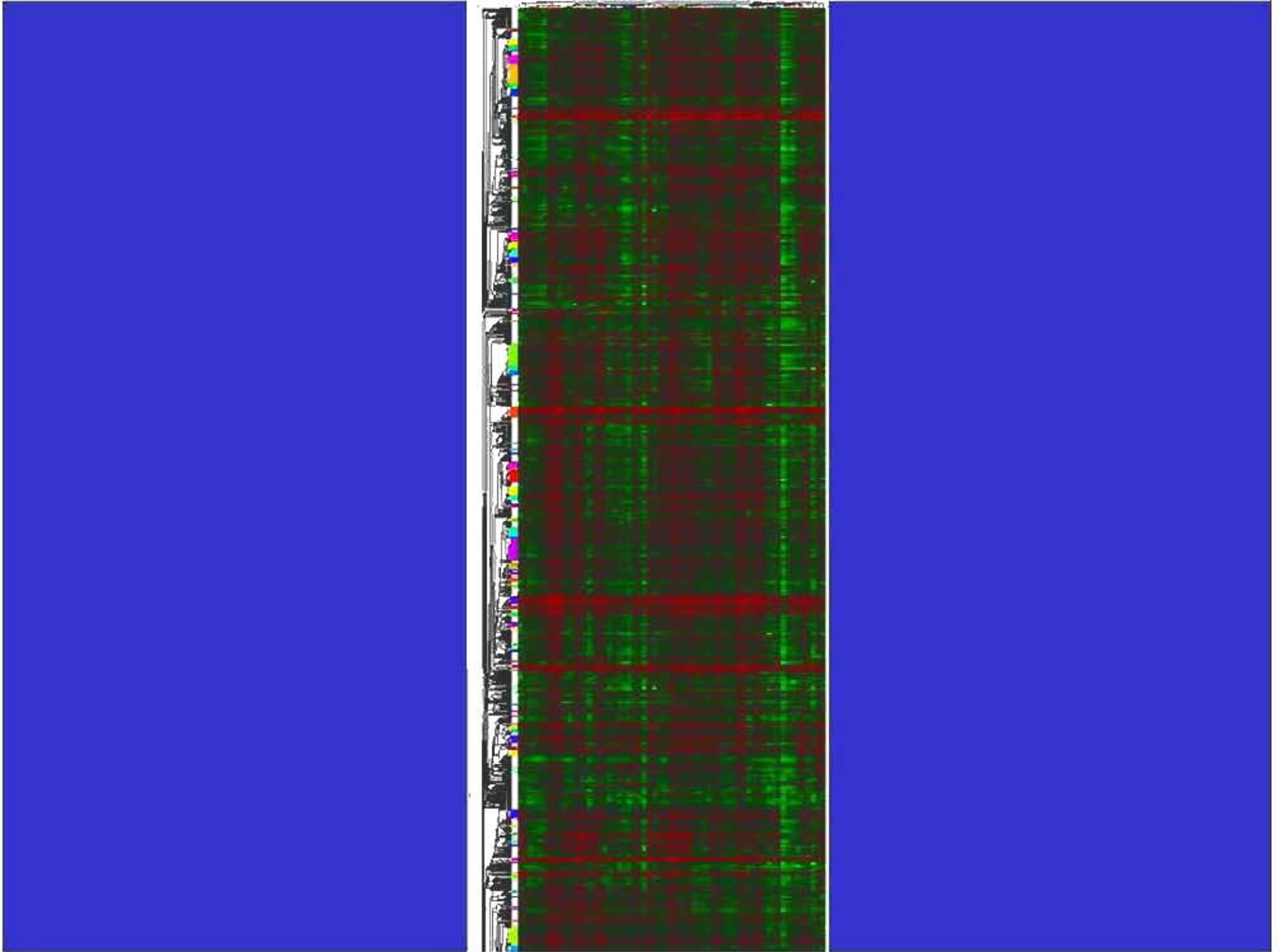
$$PE(y) = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(y_{ij} - \frac{m_i n_j}{N}\right)^2}{\frac{m_i n_j}{N}}$$

y_{ij} = gene expression level in the i th row and j th column

$$m_i = \sum_{j=1}^c y_{ij} \quad n_j = \sum_{i=1}^r y_{ij}$$

$$N = \sum_i m_i = \sum_j n_j$$

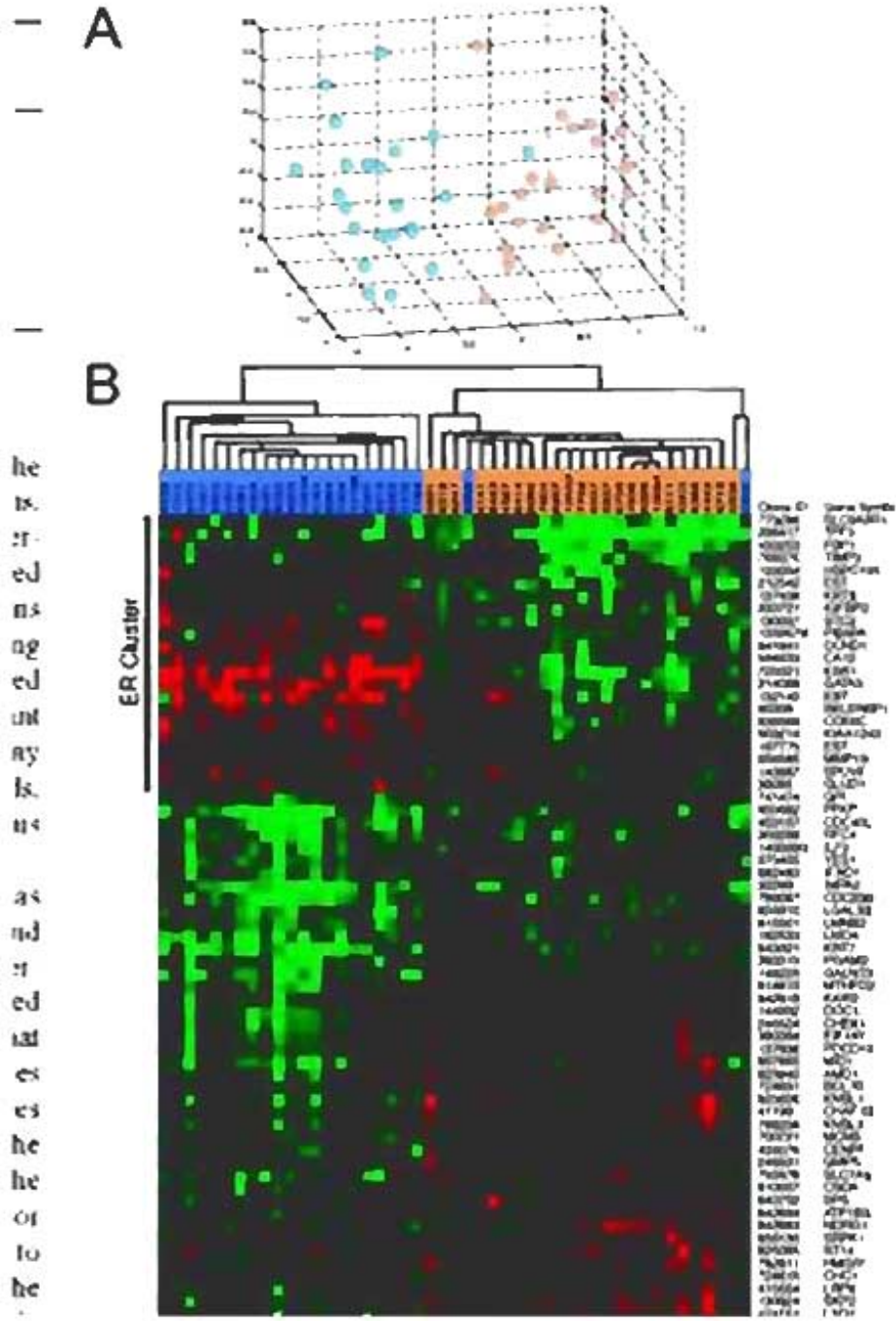
Random matrices with the same intensity distribution and same (or larger) disproportion measure as the original matrix (Monte Carlo simulations)



Generative models (random matrices retaining internal data structure) will help to determine the required sample number for statistically meaningful identification of classes and separators.

Machine learning – Artificial Neural Nets in the analysis Cancer associated gene expression matrices

DIFFERENTIAL GENE EXPRESSION PATTERNS



shuffling was redone 200 times, and for each shuffling we analyzed three ANN res

P. Meltzer,
J. Trent
M. Bittner

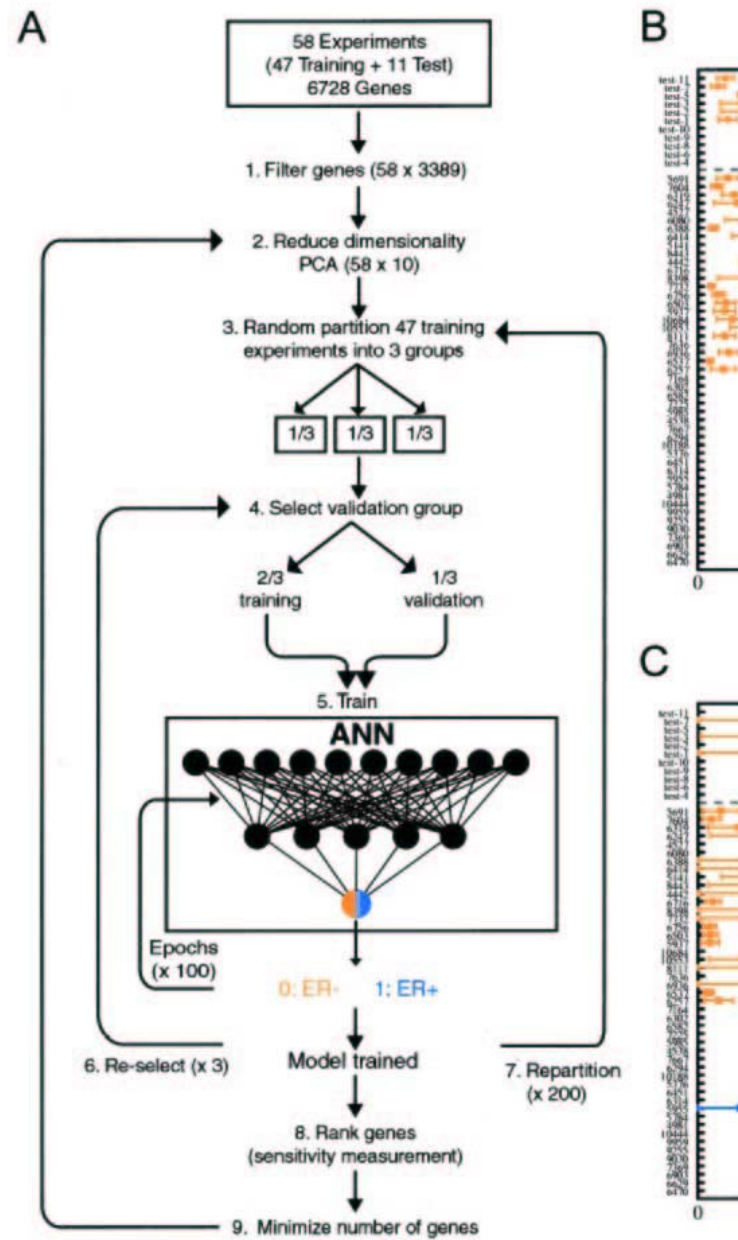


Fig. 1. Classification of ER+ and ER- tumors using ANNs and gene expression patterns. *A* expression and spot area reduced the number of genes to 3389 (1). PCA further reduced the dime into three groups (3). Two of these groups were used for training and one for validation (4) using process was repeated so that all three groups were used for validation (6). The random partition

ANN (artificial neural nets) work well when a large number of samples is available relative to the number of variables

(e.g. for the pattern recognition of hand written digits one can create a huge number of sufficiently different samples).

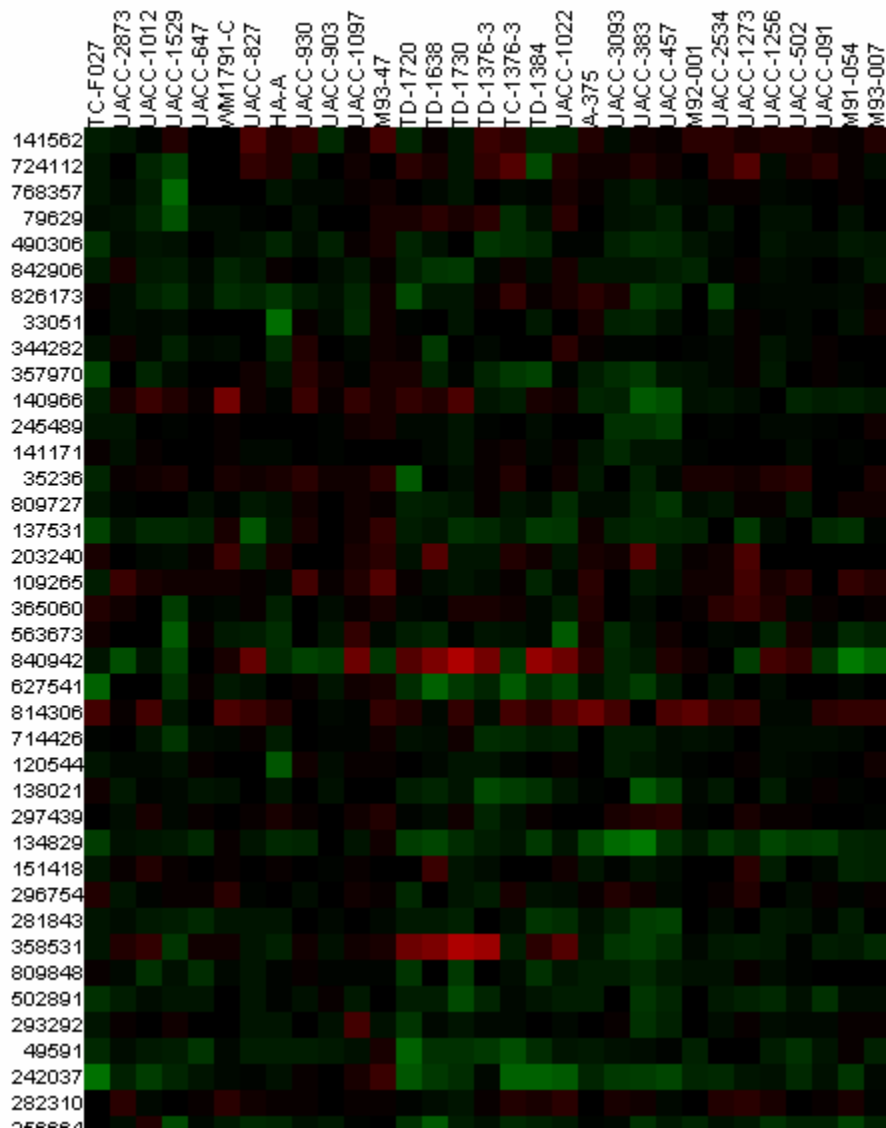
In biology there might be two limitations:

- 1. the number of samples might be quite limited, at least relative to the complexity of the problems (The cell has to survive)**
- 2. There might be a practical limit to collecting certain types of samples**

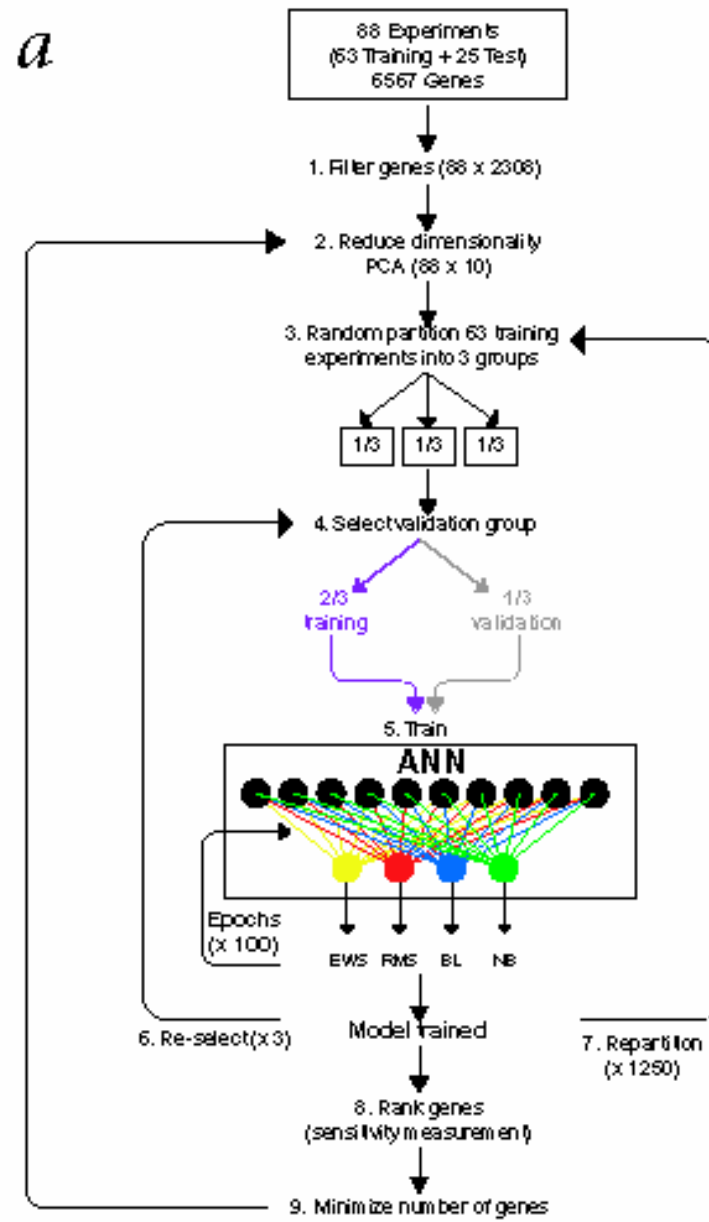
< 100 samples

Samples

> 1000



tic
ca-
ed
re-
to
ix-
its
its
re
nd
IN
n-
at-
re
es
3
in
ig
ili-
ig
d
re
si-
ed
ix-
all
a-
or



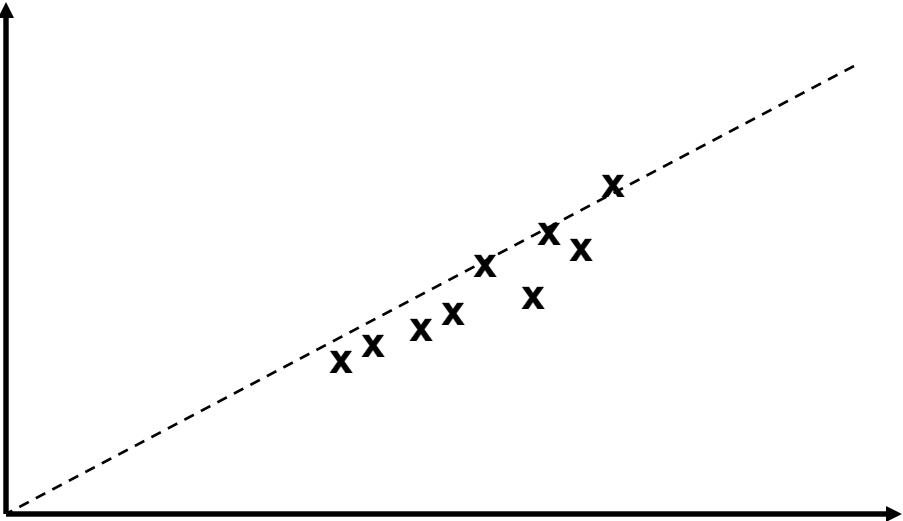
b

Summed square error

c

Number of misclassified samples

Reducing dimensionality
Principal component analysis
retain variance



The risk of reducing dimensionality by PCA

