Analysis of Gene Expression Data

Jianlin Cheng, PhD

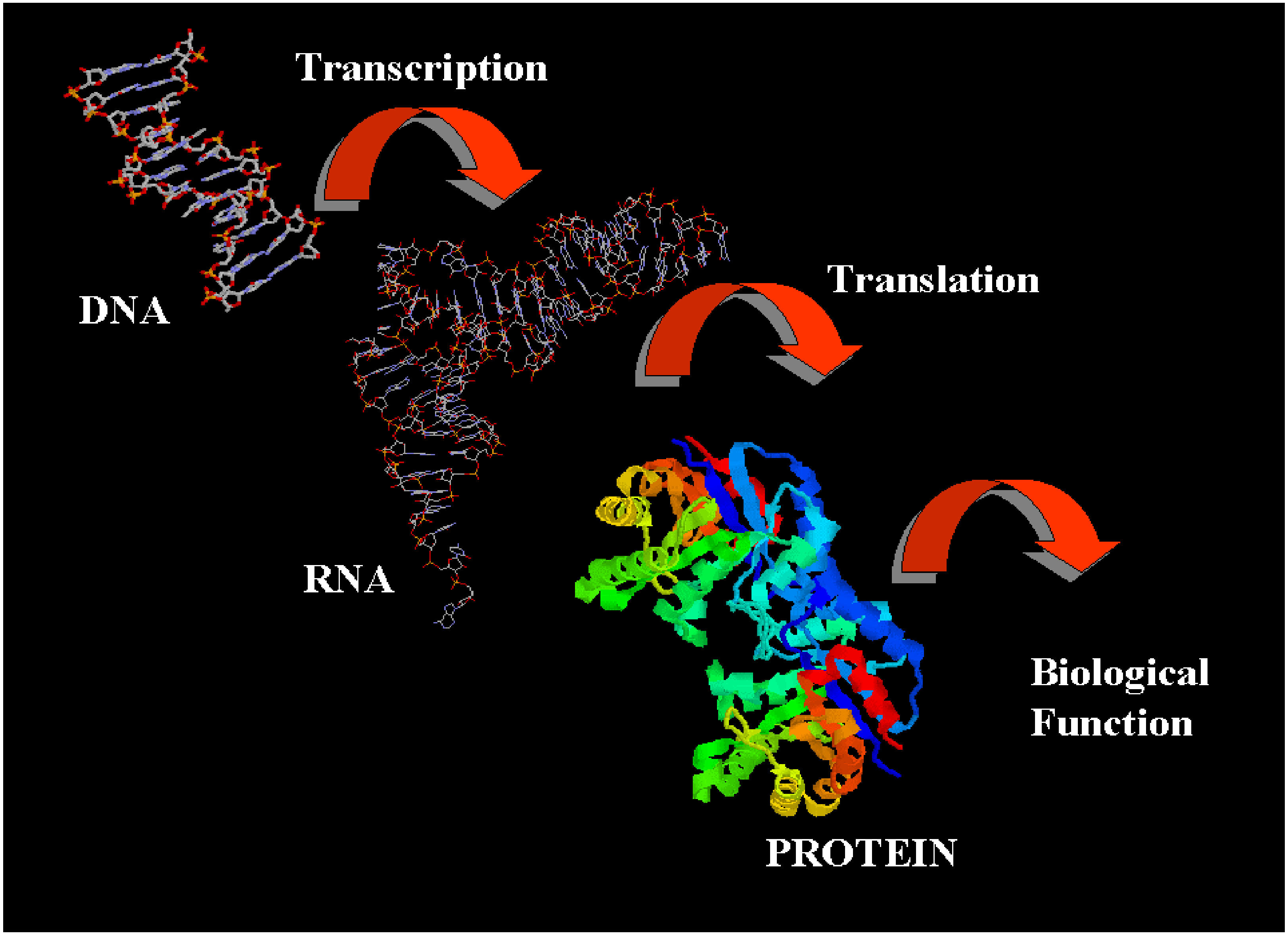
School of Electrical Engineering and Computer Science
University of Central Florida

2006

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering of gene expression data
- Classification of gene expression data
- Inference of gene regulatory networks
- Databases and software

The Central Dogma of Biology



Rainer Breitling, 2005

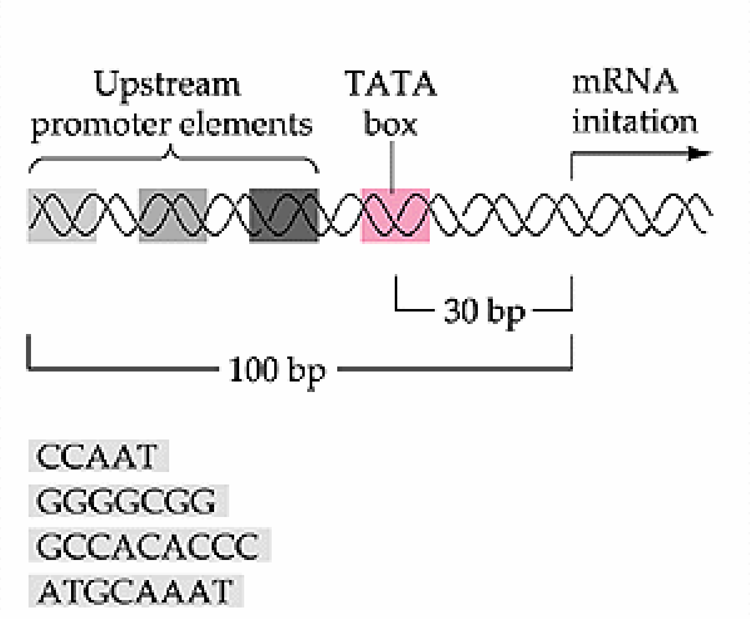
The Dramatic Consequences of Gene Regulation in Biology

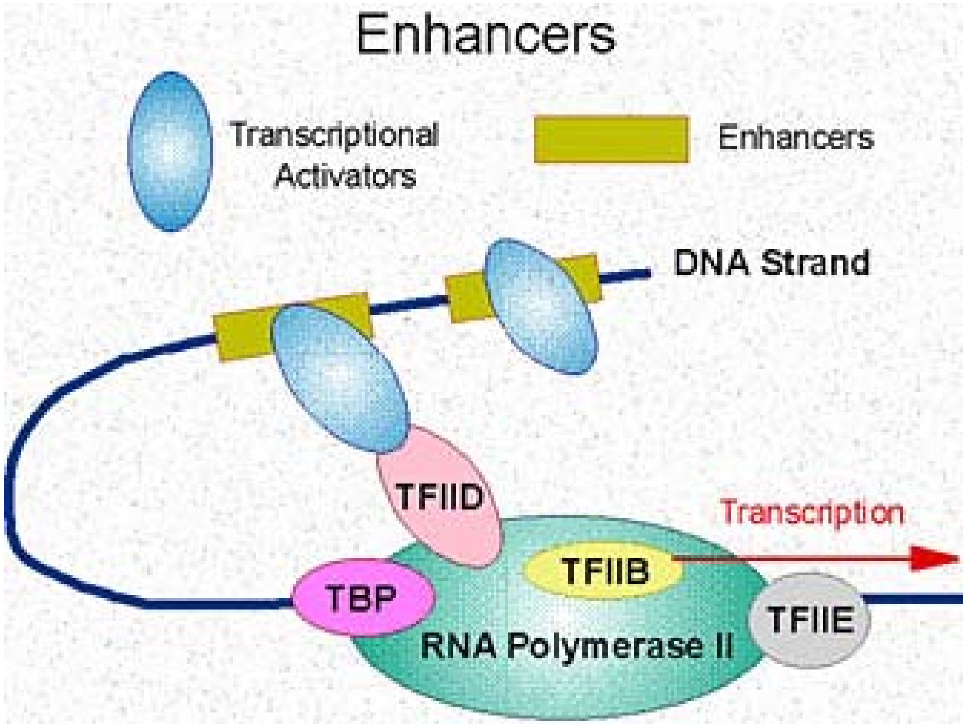
Same genome →
Different tissues
- •Different physiology
- •Different proteome
- •Different expression pattern

Anise swallowtail, *Papilio zelicaon*

Rainer Breitling, 2005

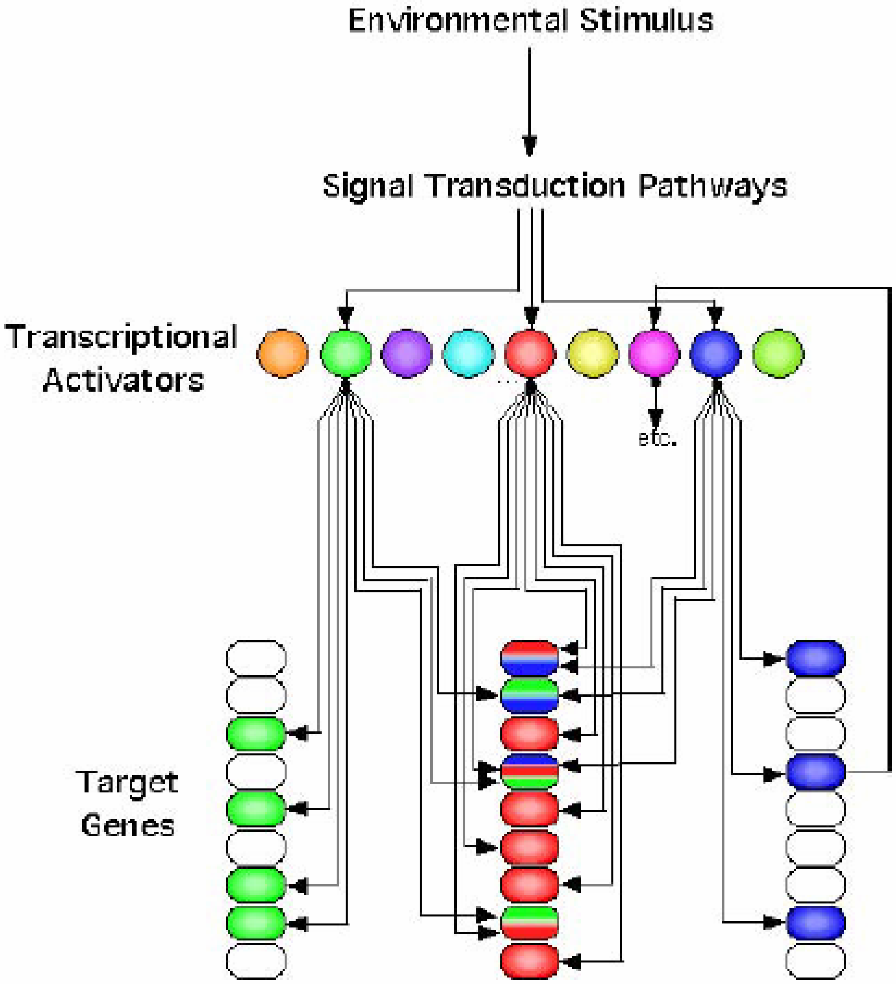
The Complexity of Eukaryotic Gene Expression Regulation





Rainer Breitling, 2005

Transcriptional Regulatory Pathways



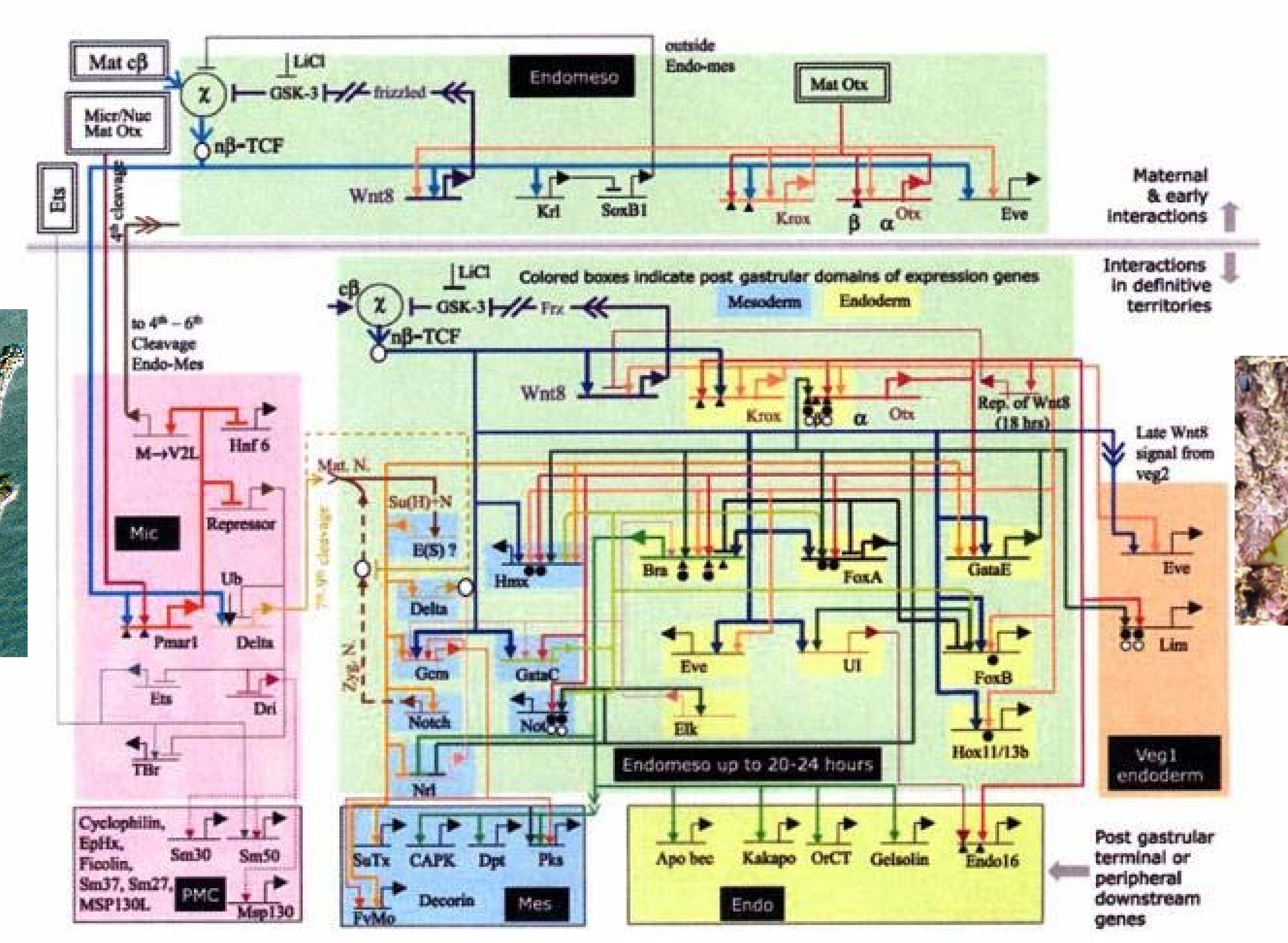
The regulatory pathways that control gene expression programs are uncharted

The mapping of transcriptional regulatory pathways will:

- reveal how cell state, differentiation and response to stimuli are controlled
- suggest new strategies to combat disease

David Gifford, 2005

Regulatory Networks – Integrating It All Together



Genetic regulatory network controlling the development of the body plan of the sea urchin embryo Davidson *et al.*, *Science*, 295(5560):1669-1678.

Rainer Breitling, 2005

Gene Expression Distinguishes...

- Physiological status (nutrition, environment)
- Sex and age
- Various tissues and cell types
- Response to stimuli (drugs, signals, toxins)
- Health and disease
  - underlying pathogenic diversity
  - progression and response to treatment
  - patient classes of varying prospects

Note: about 40% human genes are expressed at a time.

Rainer Breitling, 2005

Gene Expression Measurement

- mRNA expression represents dynamic aspects of cell
- mRNA expression can be measured by DNA Microarrays
- mRNA is isolated and labeled with fluorescent protein
- mRNA is hybridized to the target; level of hybridization corresponds to light emission which is measured with a laser
- DNA Microarray can measure the expression of thousands of genes at the same time (high throughput)

Gene Expression Microarrays

The main types of gene expression microarrays:

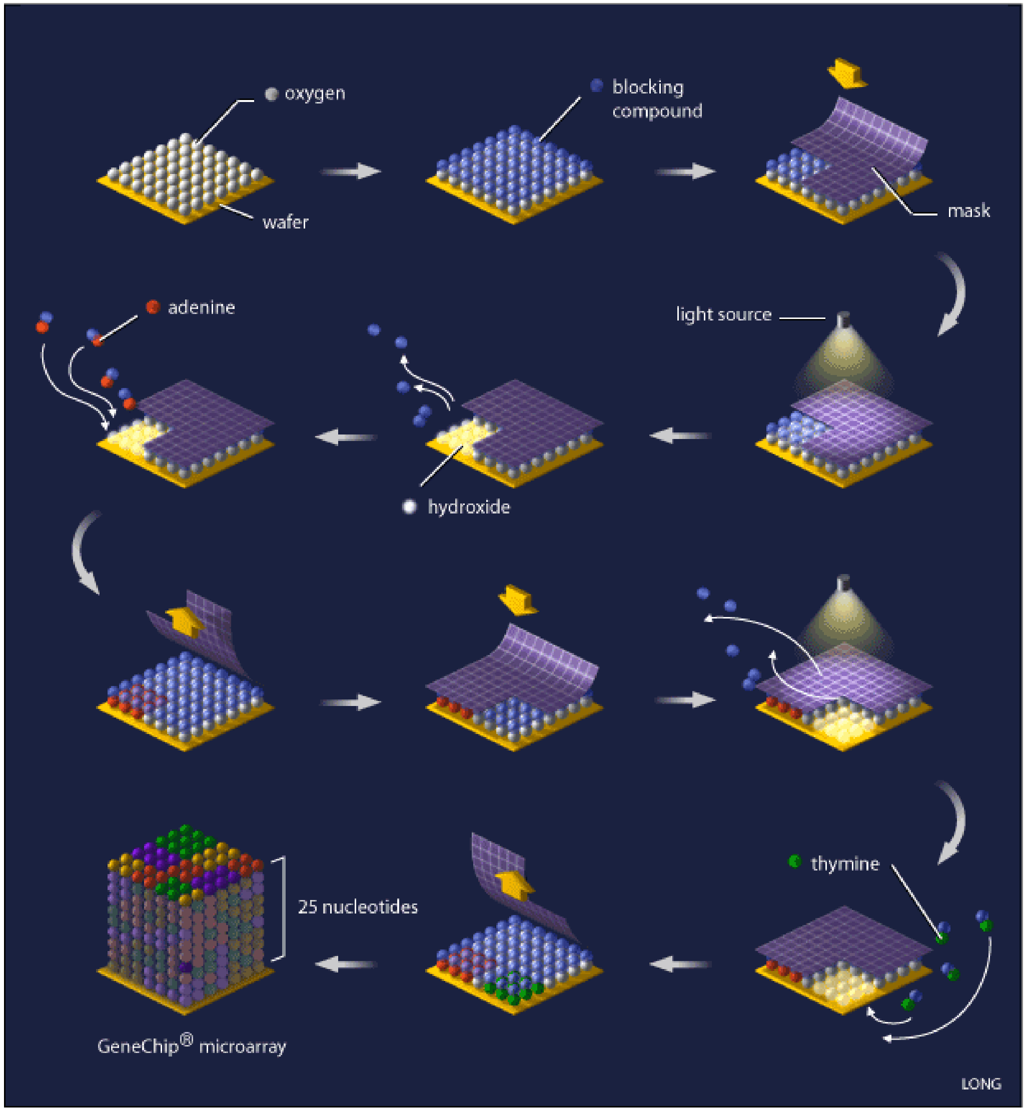
- Short oligonucleotide arrays **(Affymetrix);**
- cDNA or spotted arrays **(Brown/Botstein).**
- Long oligonucleotide arrays (Agilent Inkjet);
- Fiber-optic arrays

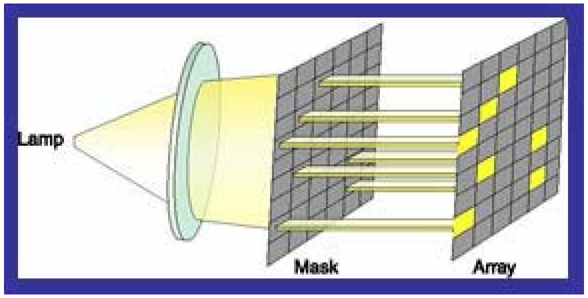
Two-color and one-color Microarrays:

- two color: produce two expression images for experimental and reference environment respectively.
- one color: produce one expression image that reflect the expression levels.

Y. Guo, V. Curain, H. Morris, 2005

GeneChip® Affymetrix

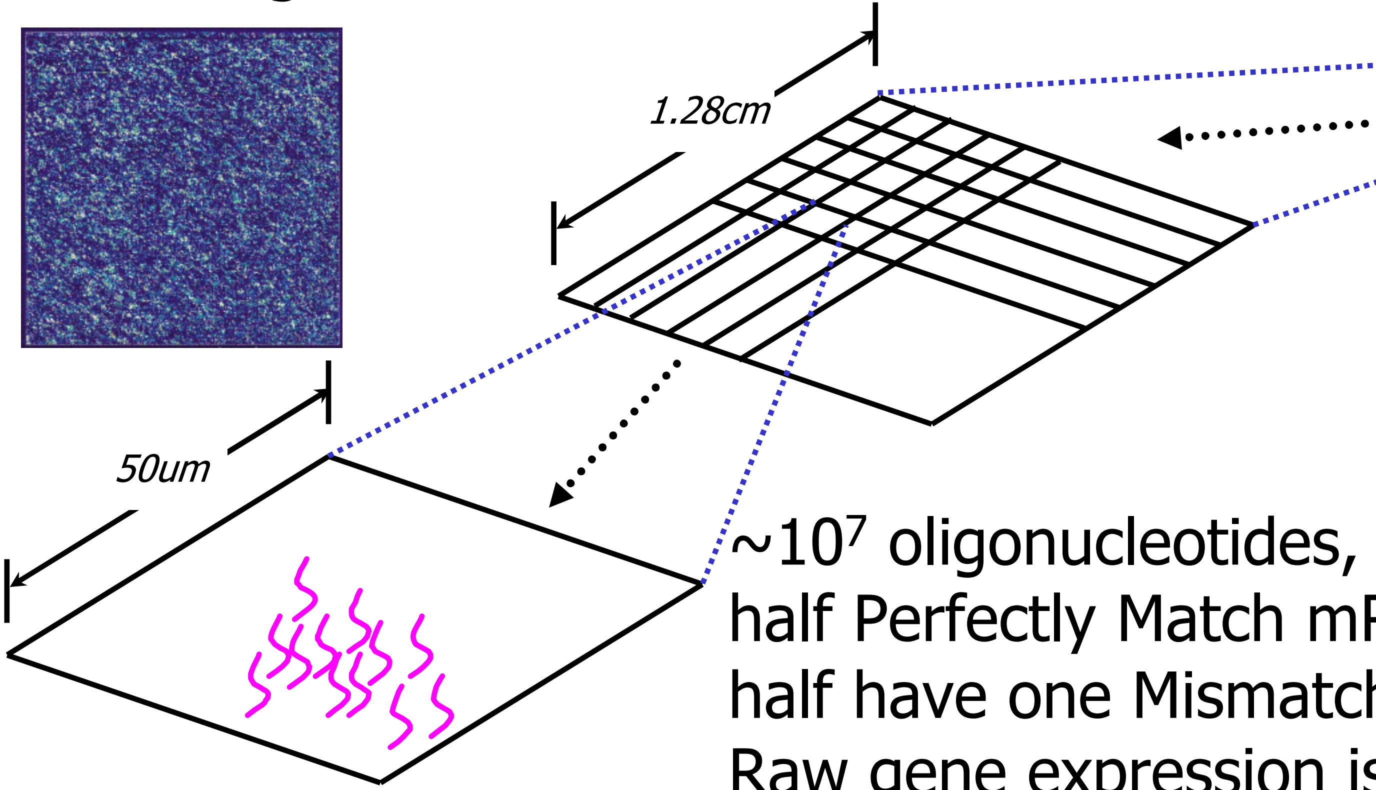


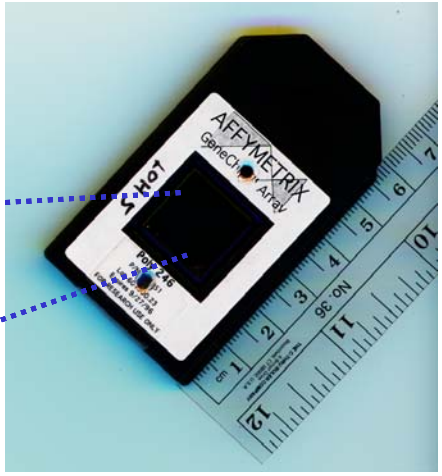


Rainer Breitling, 2005

Affymetrix Microarrays

Raw image





~10^7 oligonucleotides,
half Perfectly Match mRNA (PM),
half have one Mismatch (MM)
Raw gene expression is intensity difference: PM - MM

Rainer Breitling, 2005

GeneChip® Hybridization

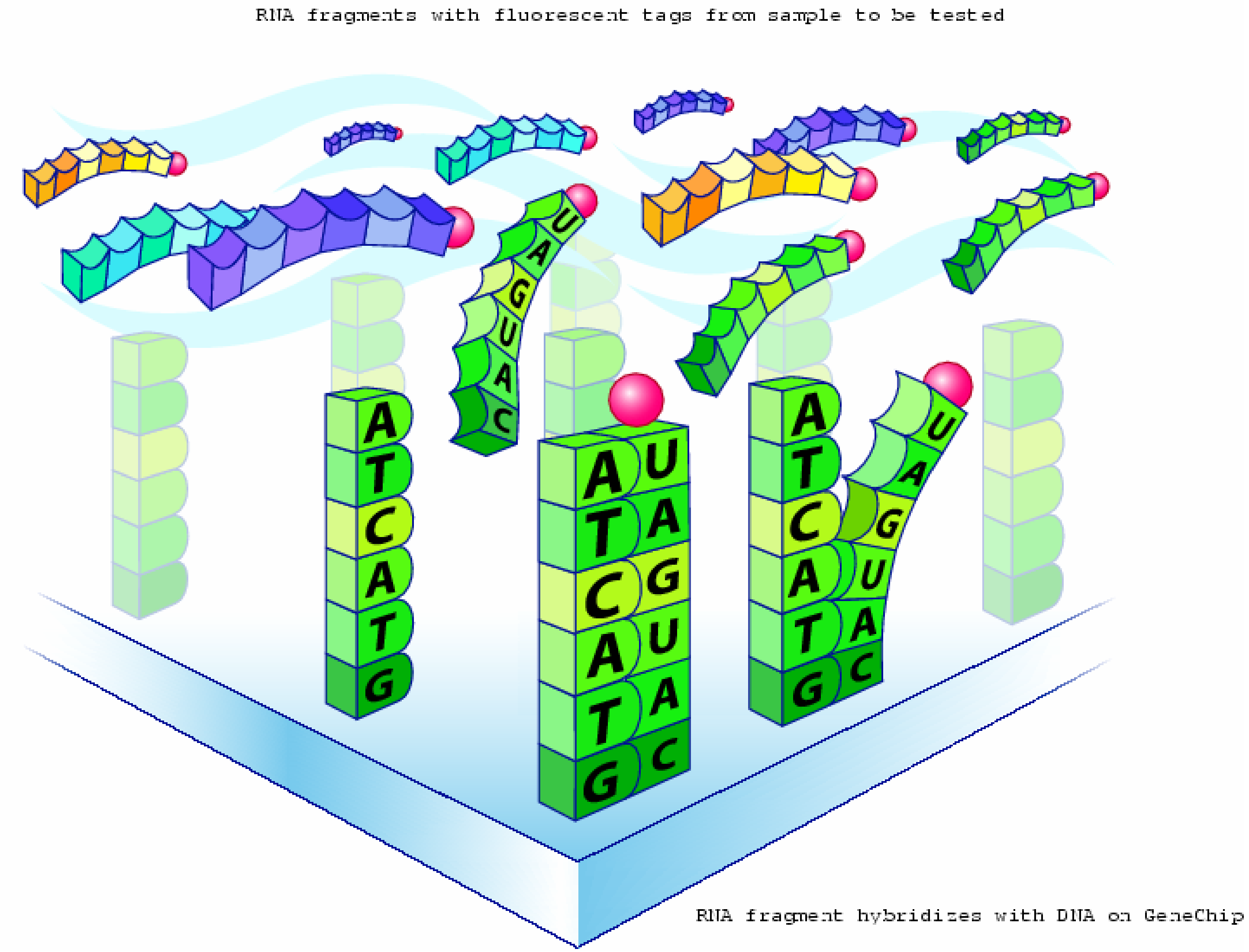
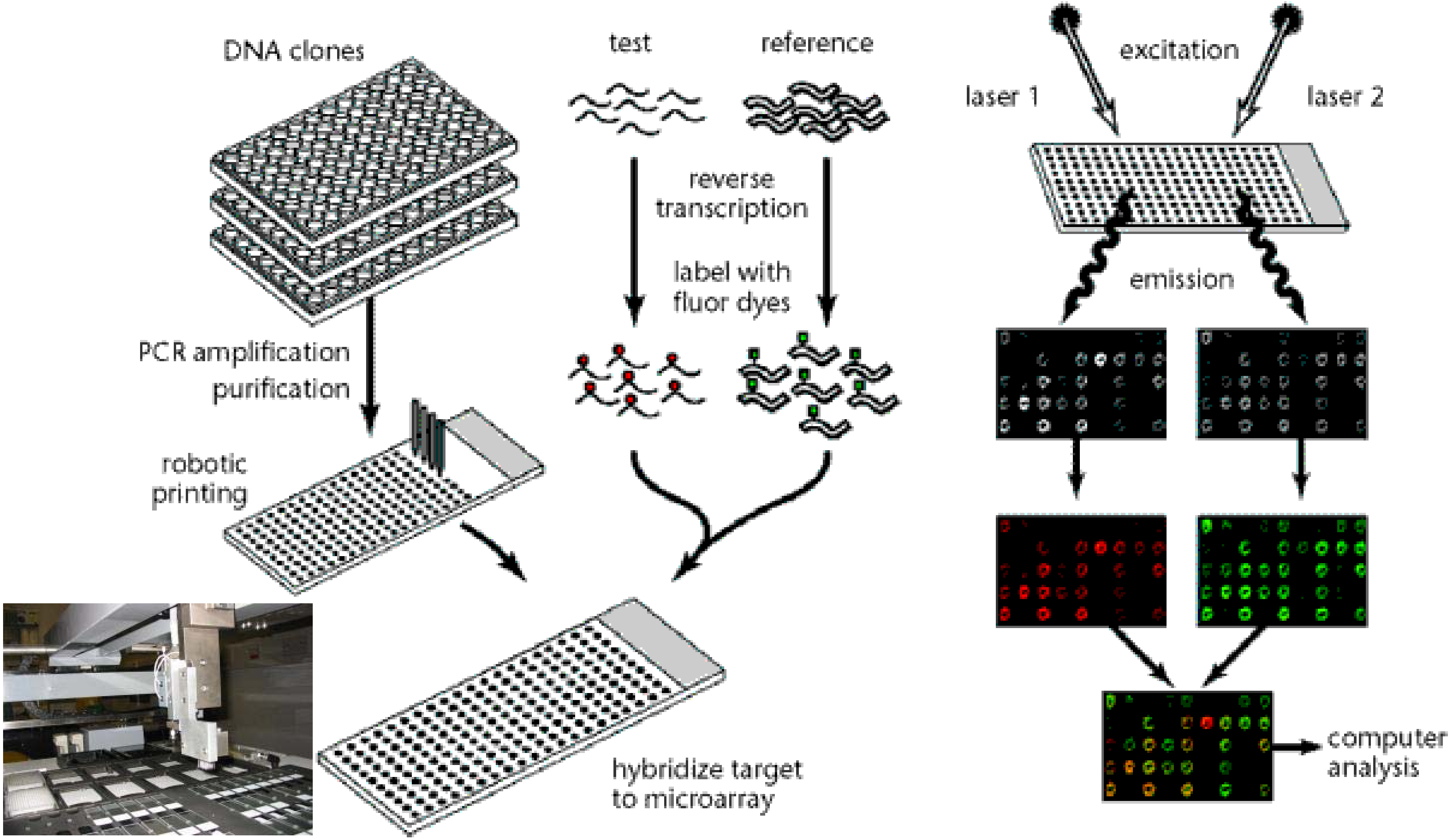

RNA fragments with fluorescent tags from sample to be tested
U A G U A C
A T C A T G
A T C A T G
U A G U A C
A T C A T G
U A G U A C
RNA fragment hybridizes with DNA on GeneChip

Image courtesy of Affymetrix.

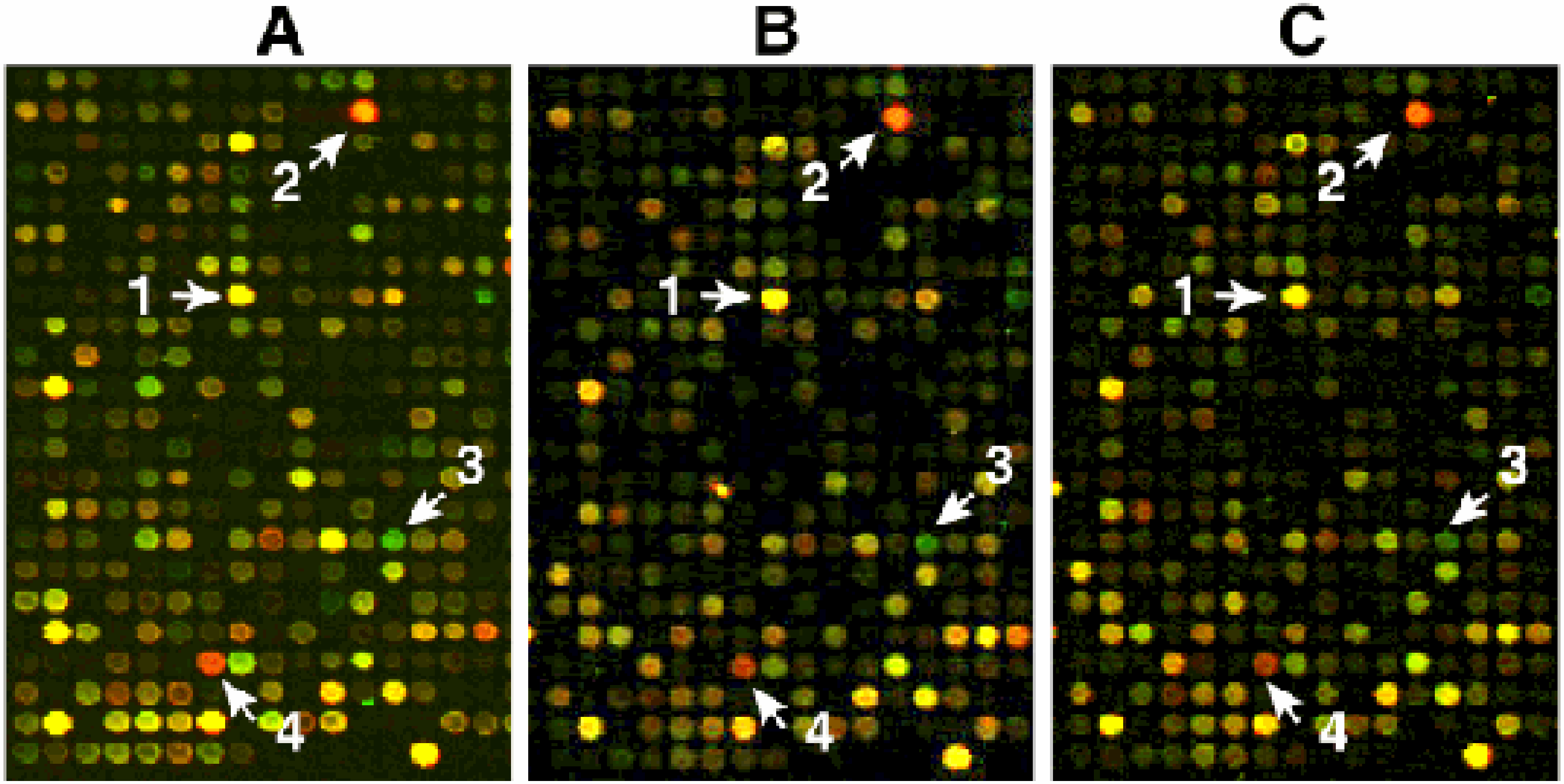
Rainer Breitling, 2005

cDNA Microarray Schema



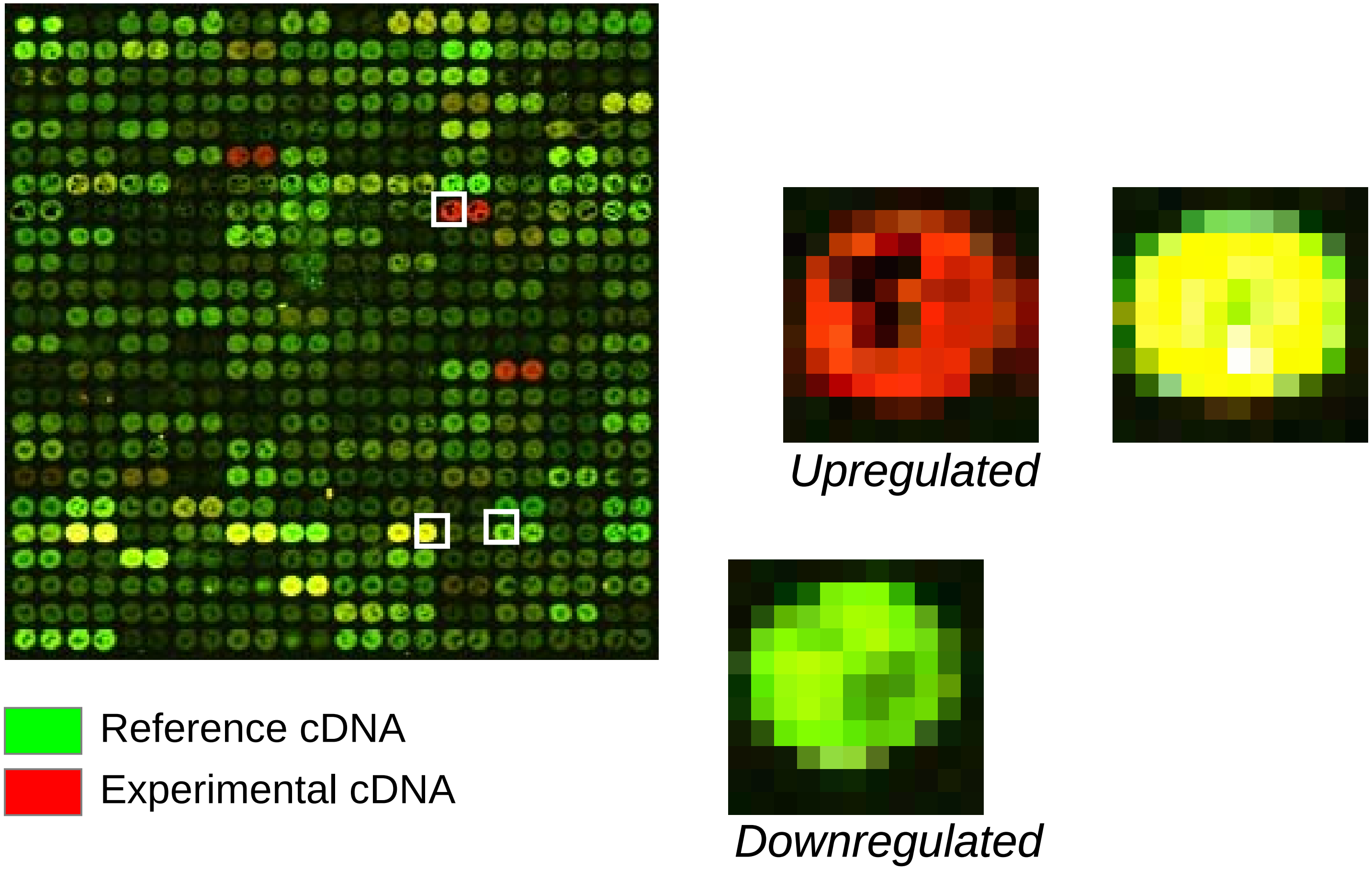
From Duggan *et al. Nature Genetics* **21**, 10 – 14 (1999)

Example of Microarray Image (One Channel / Color)



R. Murphy, 2005

Microarray Images -> Differential Expression



Upregulated

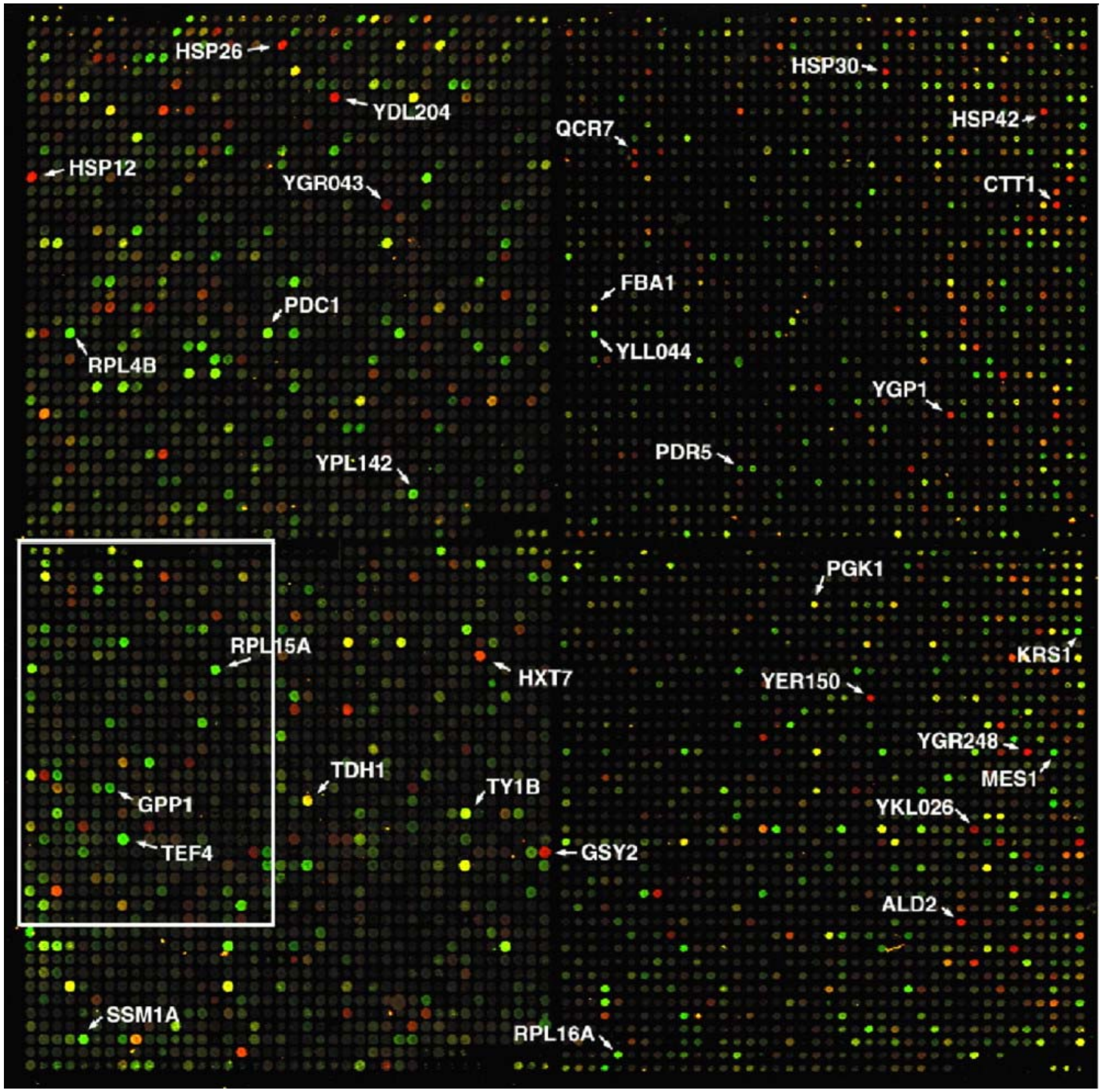
Downregulated

🟩 Reference cDNA

🟥 Experimental cDNA

A. Singh, 2005

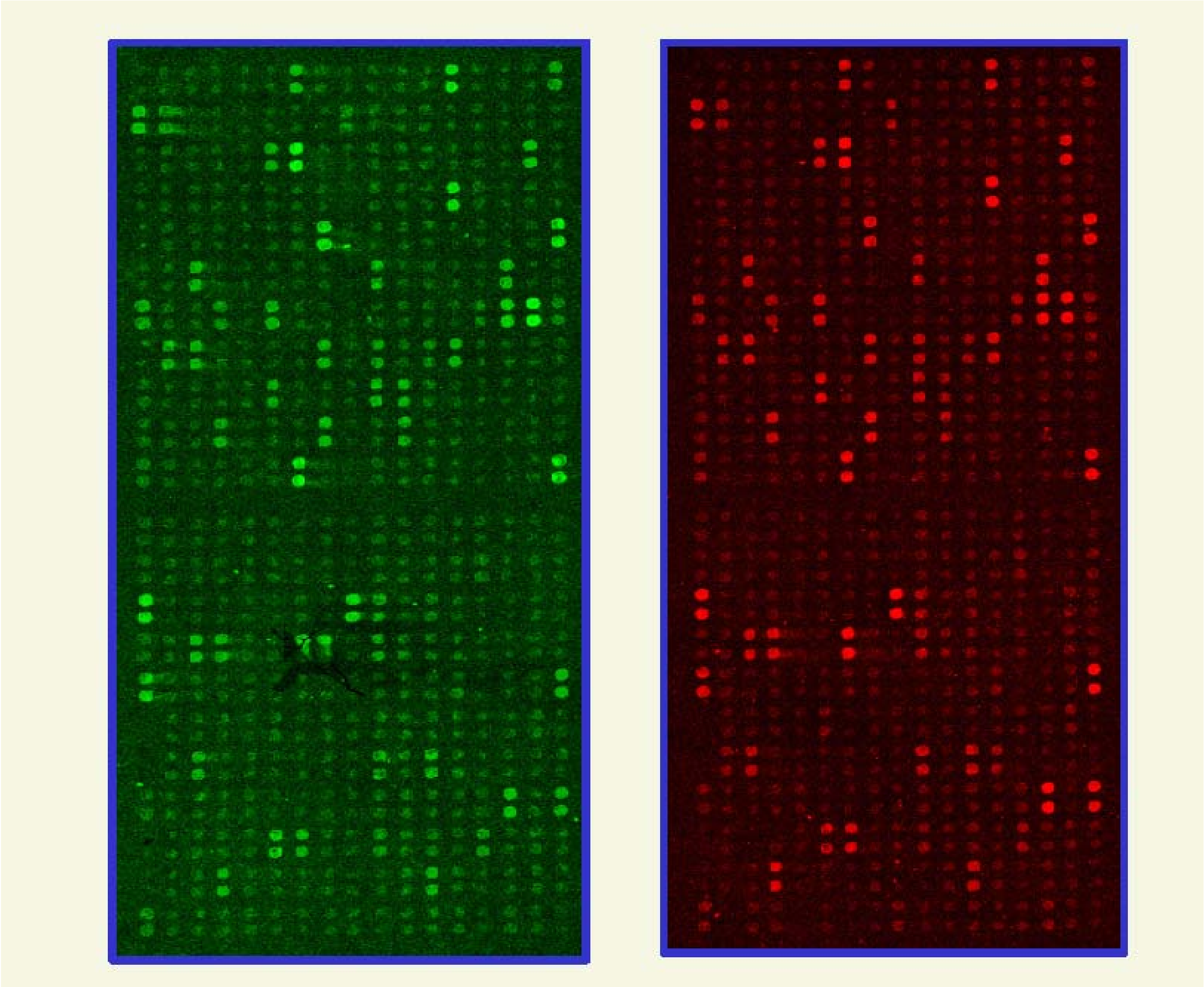
cDNA Microarray raw data



- can be custom-made in the laboratory
- always compares two samples
- relatively cheap
- up to about 20,000 mRNAs measured per array

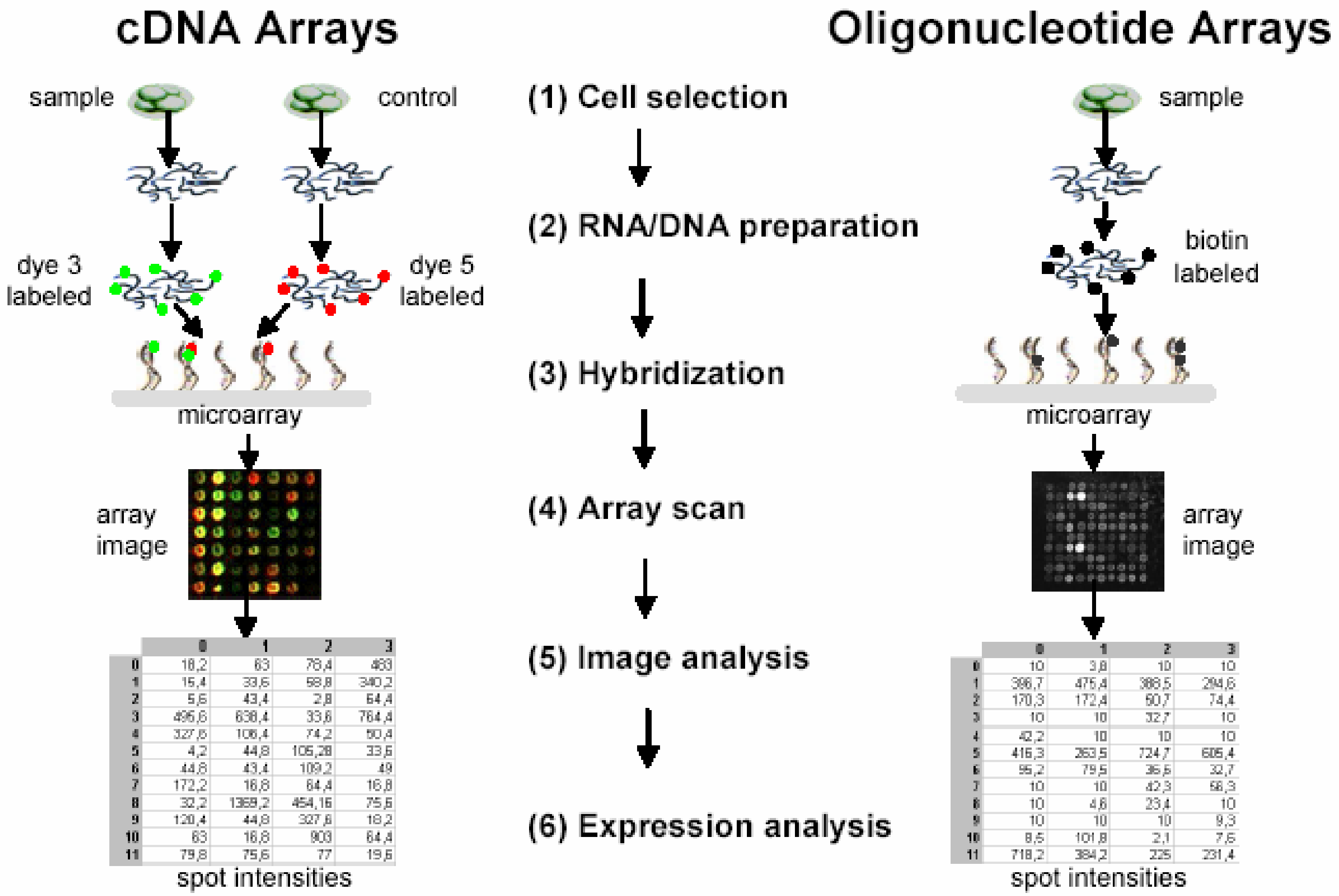
Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. (DeRisi, Iyer & Brown, Science, 268: 680-687, 1997)

Raw Image of Two Channels / Colors



F. Hong, 2005

Microarray Experiment



cDNA Arrays

	0	1	2	3
0	18,2	63	78,4	483
1	15,4	33,6	58,8	340,2
2	5,6	43,4	2,8	64,4
3	495,6	638,4	33,6	764,4
4	327,6	106,4	74,2	50,4
5	4,2	44,8	105,28	33,6
6	44,8	43,4	109,2	49
7	172,2	16,8	64,4	16,8
8	32,2	1369,2	454,16	75,6
9	120,4	44,8	327,6	18,2
10	63	16,8	903	64,4
11	79,8	75,6	77	19,6

spot intensities

(1) Cell selection

(2) RNA/DNA preparation

(3) Hybridization

(4) Array scan

(5) Image analysis

(6) Expression analysis

Oligonucleotide Arrays

	0	1	2	3
0	10	3,8	10	10
1	396,7	475,4	388,5	294,6
2	170,3	172,4	50,7	74,4
3	10	10	32,7	10
4	42,2	10	10	10
5	416,3	263,5	724,7	605,4
6	95,2	79,5	36,6	32,7
7	10	10	42,3	56,3
8	10	4,6	23,4	10
9	10	10	10	9,3
10	8,5	101,8	2,1	7,6
11	718,2	384,2	225	231,4

spot intensities

H. Do, T. Kirsten, E. Rahm, 2003

Image Processing

- Gridding
  - – Identifying spot locations
- Segmentation
  - – Identifying foreground and background
- Removal of outliers
- Absolute measurements
  - – cDNA microarray
    - Intensity level of red and green channels
  - – Affymetrix chips
    - Average difference of PM and MM spots

Data Extraction

One Color

- Calculate ratio of red to green fluorescence
- Convert to $\log_2$ and round to integer

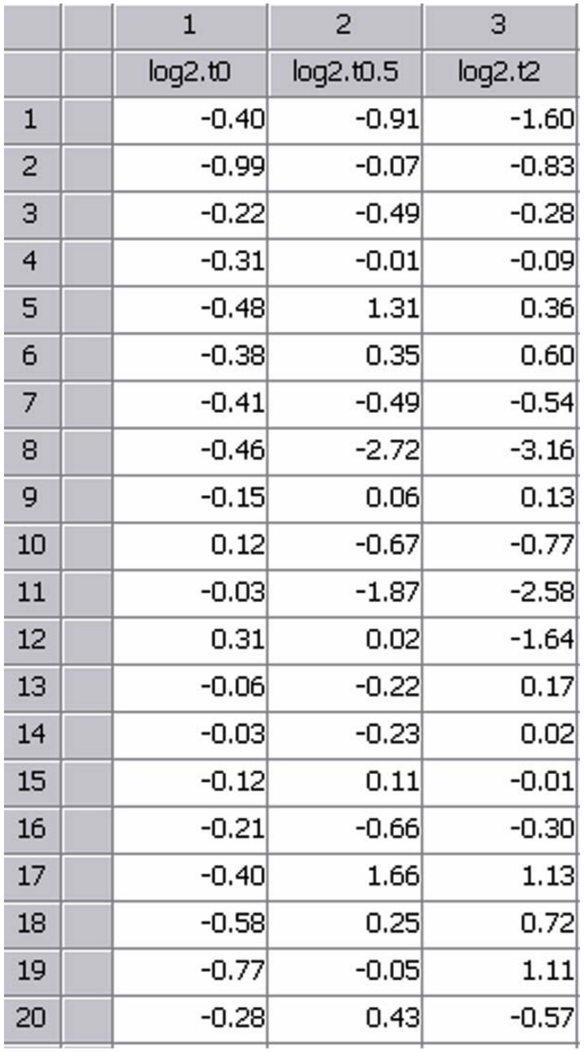
Two-Color

- Calculate log R and log G.

Microarray Data Example

Time Points

Genes



		1	2	3
		log2.t0	log2.t0.5	log2.t2
1		-0.40	-0.91	-1.60
2		-0.99	-0.07	-0.83
3		-0.22	-0.49	-0.28
4		-0.31	-0.01	-0.09
5		-0.48	1.31	0.36
6		-0.38	0.35	0.60
7		-0.41	-0.49	-0.54
8		-0.46	-2.72	-3.16
9		-0.15	0.06	0.13
10		0.12	-0.67	-0.77
11		-0.03	-1.87	-2.58
12		0.31	0.02	-1.64
13		-0.06	-0.22	0.17
14		-0.03	-0.23	0.02
15		-0.12	0.11	-0.01
16		-0.21	-0.66	-0.30
17		-0.40	1.66	1.13
18		-0.58	0.25	0.72
19		-0.77	-0.05	1.11
20		-0.28	0.43	-0.57

Typically, there are many genes (>> 10,000) and few samples (~ 10)

J. Pevsner, 2005

Characteristics of Microarray Data

- Extremely high dimensionality
  - Experiment = (gene$_1$, gene$_2$, …, gene$_N$)
  - Gene = (experiment$_1$, experiment$_2$, …, experiment$_M$)
  - *N* is often on the order of 10^4
  - *M* is often on the order of 10^1
- Noisy data
  - Normalization and thresholding are important
- Missing data
  - For some experiments a given gene may have failed to hybridize

A. Singh, 2005

Data Mining Challenges

- Too few experiments (samples), usually < 100
- Too many rows (genes), usually > 1,000
- Model needs to be explainable to biologists

Five Main Problems

1. Data pre-processing (normalization)
2. Identify differentially expressed genes in normal and non-normal situations.
3. Clustering genes according to expression data
4. Use gene expression data to classify samples (e.g., diagnosis of cancer)
5. Infer biological networks

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

Microarray Data Analysis: Preprocessing

Observed differences in gene expression could be due to transcriptional changes, or they could be caused by artifacts such as:

- different labeling efficiencies of Cy3, Cy5
- uneven spotting of DNA onto an array surface
- variations in RNA purity or quantity
- variations in washing efficiency
- variations in scanning efficiency

J. Pevnser, 2005

Microarray data analysis: preprocessing

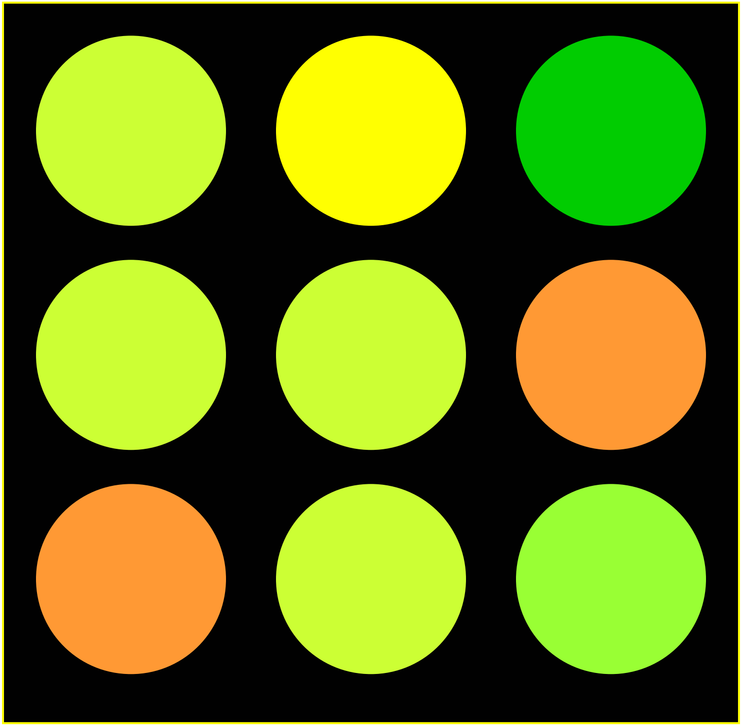
The main goal of data preprocessing is to remove the systematic bias in the data as completely as possible, while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription.

A basic assumption of most normalization procedures is that the average gene expression level does not change in an experiment.

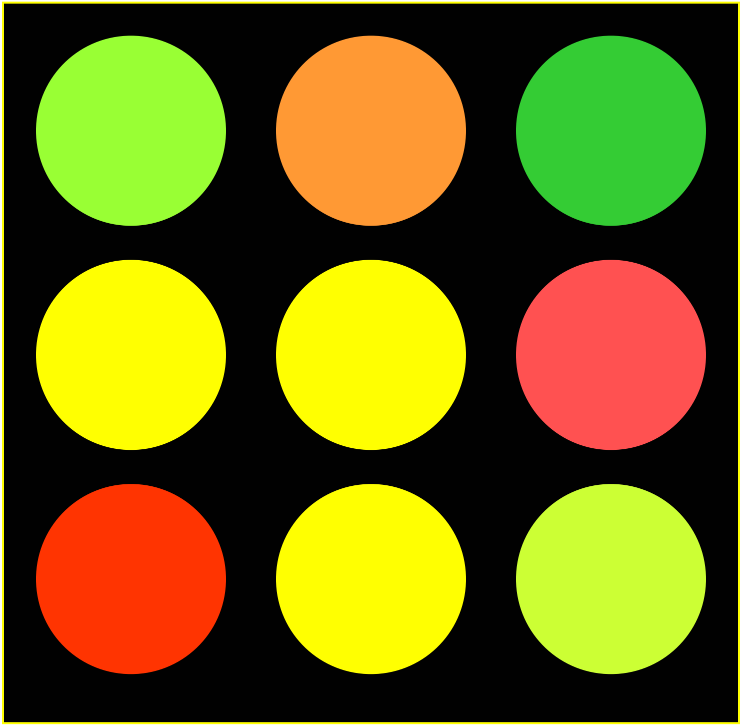
J. Pevnser, 2005

Data normalization

Uncalibrated, red light under detected



Calibrated, red and green equally detected



A. Singh, 2005

Data analysis: global normalization

Global normalization procedure

Step 1: subtract background intensity values
(use a blank region of the array)

Step 2: globally normalize so that the average ratio = 1

Some researchers use housekeeping genes for
global normalization

J. Pevnser, 2005

Normalization: global

- Normalization based on a *global adjustment*

$$\log_2 R/G \;\rightarrow\; \log_2 R/G - c = \log_2 R/(kG)$$

- Common choices for k or c = $\log_2 k$ are c = *median* or *mean* of log ratios for a particular gene set (e.g. all genes, or control or housekeeping genes)

http://ludwig-sun2.unil.ch/~darlene/

Gene expression data example

Data on *m* genes for *n* samples

mRNA samples

		sample1	sample2	sample3	sample4	sample5	…
Genes	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene *i* in mRNA sample *j*

= (normalized) Log(Red intensity / Green intensity)

http://ludwig-sun2.unil.ch/~darlene/

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

Scatter plots

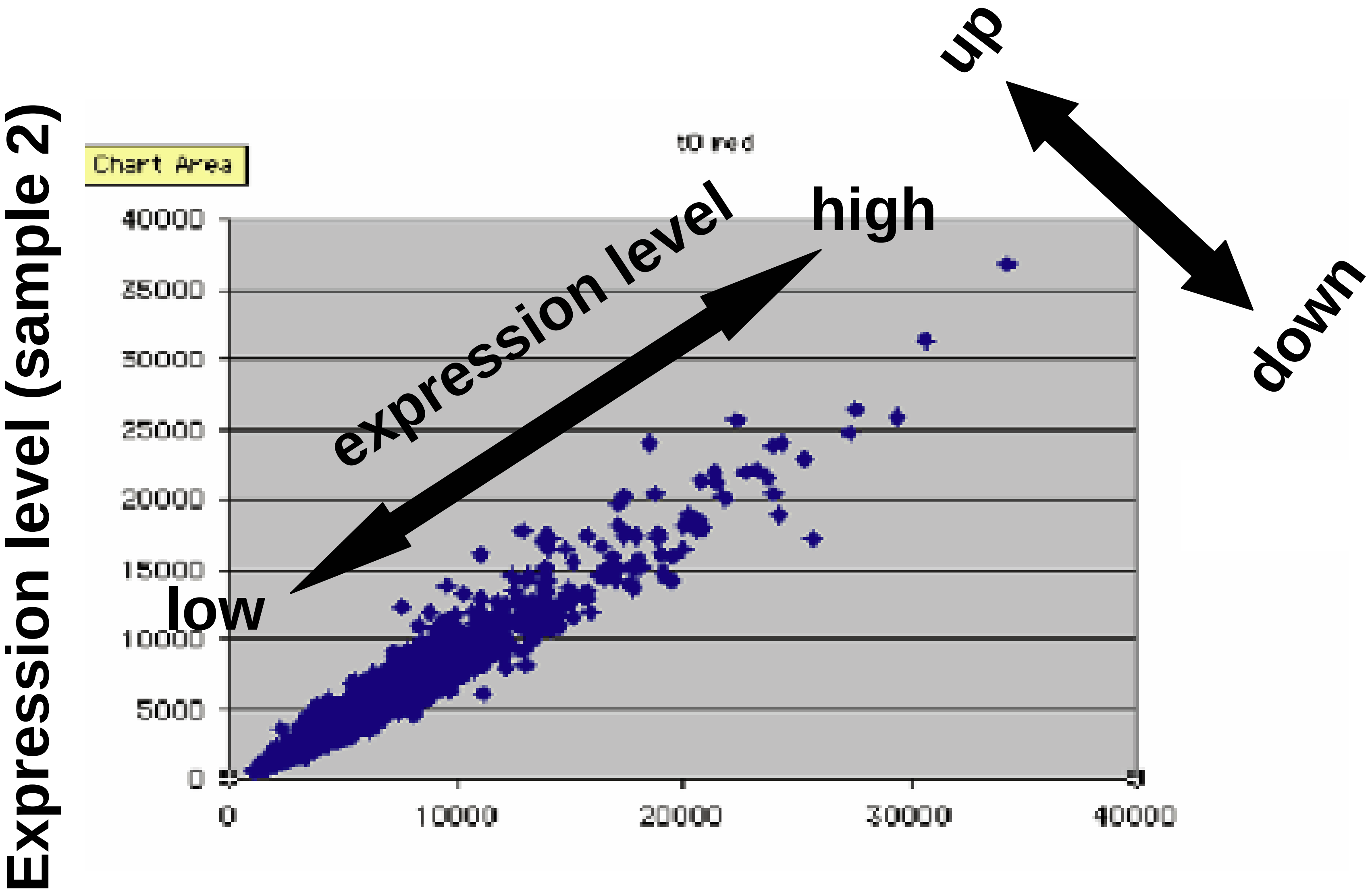
Useful to represent gene expression values (logarithm) from two microarray experiments (e.g. control, experimental)

Each dot corresponds to a gene expression value (logarithm)

Most dots fall along a line

Outliers represent up-regulated or down-regulated genes

J. Pevsner, 2005



Expression level (sample 2)

expression level

high

low

up

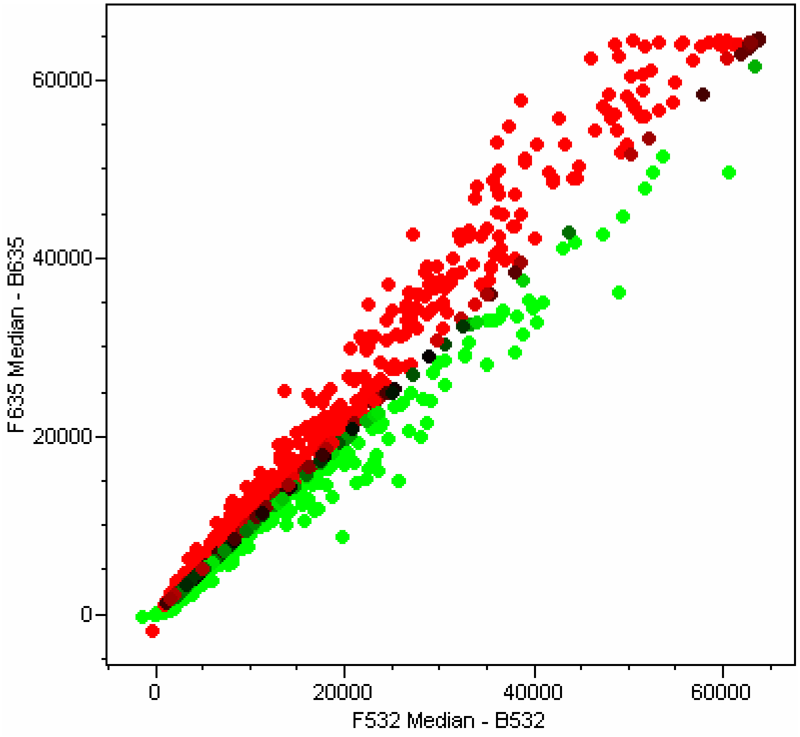
down

Expression level (sample 1)

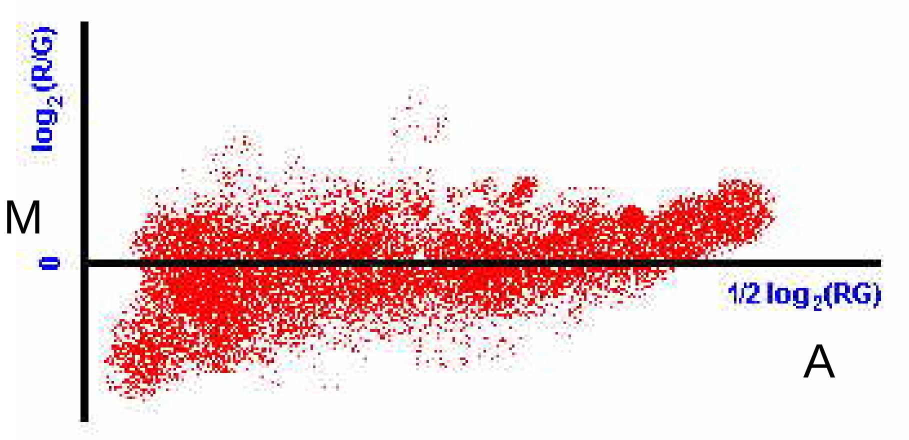
J. Pevsner, 2005

Scatter plots

classical scatter plot



M-A plot for microarray analysis



$$M = \log_2 \left({}^{R}\!/_{G} \right)$$

$$A = \log_2 \sqrt{RG} \quad \text{OR} \quad {}^{1}\!/_{2} \log_2 RG$$

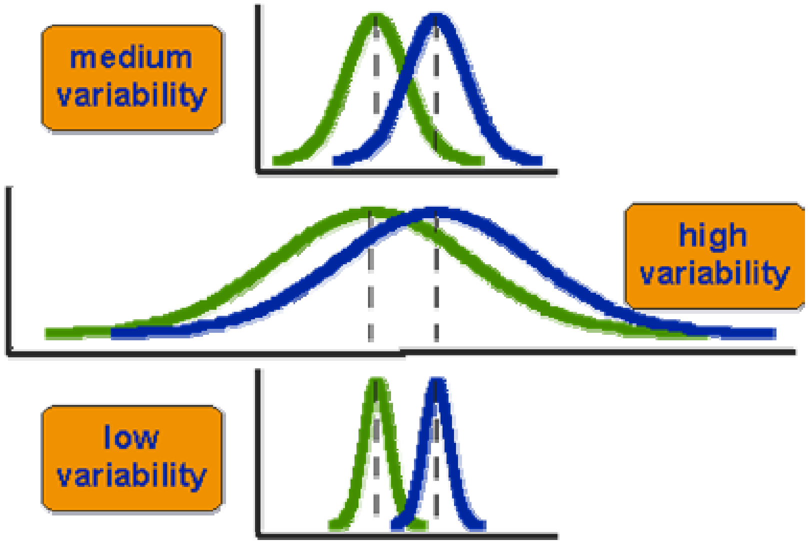
Differentially expressed genes are higher (or lower) in one of the samples

Use an appropriate cut-off (‘distance’ from diagonal) to select relevant genes → **highly arbitrary!**

Rainer Breitling, 2005

t-test = statistical significance of observed difference

- requires independent experimental replication
- assumes the data are identically normally distributed



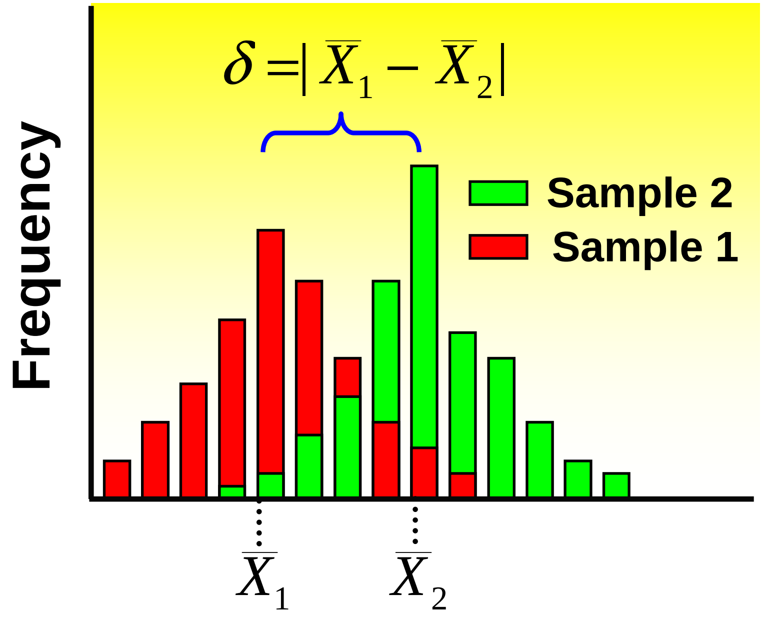
$$t = \frac{\text{difference of means}}{\text{variabilit y}}$$

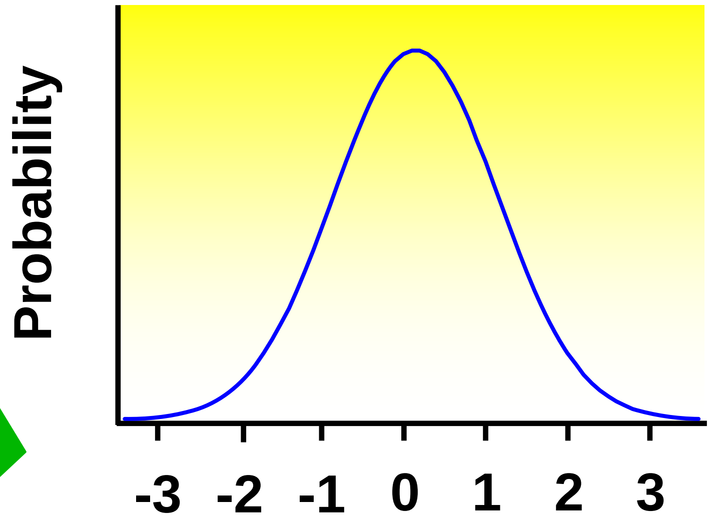
$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

Rainer Breitling, 2005

Testing an intrinsic hypothesis

- Two samples (1, 2) with mean expression that differ by some amount δ.
- If $H_0 : \delta = 0$ is true, then the expected distribution of the test statistic t is

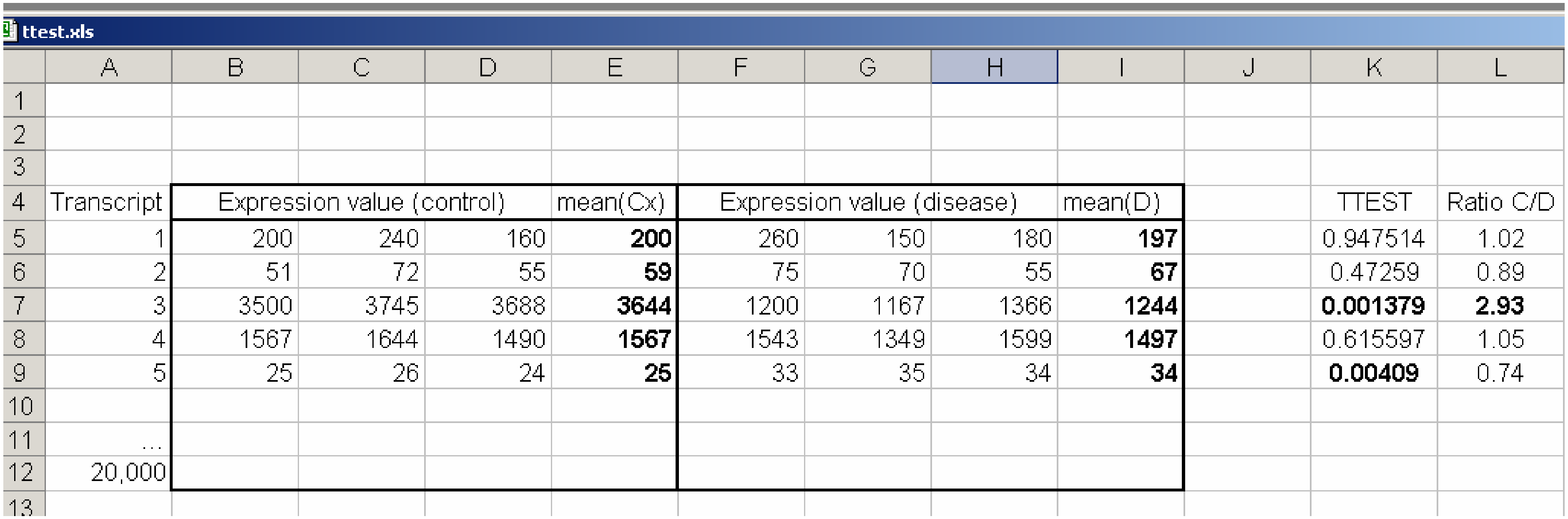




$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

Rainer Breitling, 2005

T-test Example



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3												
4	Transcript	Expression value (control)			mean(Cx)	Expression value (disease)			mean(D)		TTEST	Ratio C/D
5	1	200	240	160	**200**	260	150	180	**197**		0.947514	1.02
6	2	51	72	55	**59**	75	70	55	**67**		0.47259	0.89
7	3	3500	3745	3688	**3644**	1200	1167	1366	**1244**		**0.001379**	**2.93**
8	4	1567	1644	1490	**1567**	1543	1349	1599	**1497**		0.615597	1.05
9	5	25	26	24	**25**	33	35	34	**34**		**0.00409**	0.74
10												
11	...											
12	20,000											
13												

J. Pevsner, 2005

The result of “differential expression” statistical analysis → a long list of genes!

	Fold-Change	Gene Symbol	Gene Title
1	26.45	TNFAIP6	tumor necrosis factor, alpha-induced protein 6
2	25.79	THBS1	thrombospondin 1
3	23.08	SERPINE2	serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2
4	21.5	PTX3	pentaxin-related gene, rapidly induced by IL-1 beta
5	18.82	THBS1	thrombospondin 1
6	16.68	CXCL10	chemokine (C-X-C motif) ligand 10
7	18.23	CCL4	chemokine (C-C motif) ligand 4
8	14.85	SOD2	superoxide dismutase 2, mitochondrial
9	13.62	IL1B	interleukin 1, beta
10	11.53	CCL20	chemokine (C-C motif) ligand 20
11	11.82	CCL3	chemokine (C-C motif) ligand 3
12	11.27	SOD2	superoxide dismutase 2, mitochondrial
13	10.89	GCH1	GTP cyclohydrolase 1 (dopa-responsive dystonia)
14	10.73	IL8	interleukin 8
15	9.98	ICAM1	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
16	9.97	SLC2A6	solute carrier family 2 (facilitated glucose transporter), member 6
17	8.36	BCL2A1	BCL2-related protein A1
18	7.33	TNFAIP2	tumor necrosis factor, alpha-induced protein 2
19	6.97	SERPINB2	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 2
20	6.69	MAFB	v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian)

Rainer Breitling, 2005

Biological Interpretation Strategy

- Are certain types of genes more common at the top of the list and is that significant?
- Challenges:
  - Some types of genes are more common in the genome/on the array
  - The list of genes usually stops at an arbitrary cut-off (“significantly changed genes”)
  - Classifying genes according to “gene type” is a tedious task
  - Expectations and focused expertise might bias the interpretation

- Solution: Automated procedure using available annotations

Rainer Breitling, 2005

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

Clustering goals

- Find natural classes in the data, un-supervised learning
- Identify gene classes / gene correlations / gene functions
- Support biological analysis / discovery (regulatory sites)
- Different Methods
  - Hierarchical clustering, SOM, k-means, PCA

Two Components of Clustering Algorithms

- Similarity / Distance Measures

- Clustering Methods

Similarity / Distance Measures

Pearson correlation (looks for similarity in shape of the response profile, not the absolute values)

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma_x} \right) \left(\frac{y_i - \overline{y}}{\sigma_y} \right)$$

Euclidean distance

takes absolute expression level into account

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan (or city-block) **distance**

$$d = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Rainer Breitling, 2005

Euclidean distance:

The distance between two vectors is the square root of the sum of the squared differences over all coordinates.

$$d_E(x_1, x_2) = \sqrt{(2\text{-}2/4)^2 + (4\text{-}4/4)^2 + (5\text{-}5/4)^2 + (6\text{-}6/4)^2} = 3\sqrt{3/4} \approx 2.598$$

$x_1 = (2, 4, 5, 6)$

$x_2 = (2/4, 4/4, 5/4, 6/4)$

$x_3 = (6/4, 4/4, 3/4, 2/4)$

$x_4 = (2.5, 3.5, 4.5, 1)$

0	**2.60**	**2.75**	**2.25**
2.60	**0**	**1.23**	**2.14**
2.75	**1.23**	**0**	**2.15**
2.25	**2.14**	**2.15**	**0**

Matrix of pairwise distances

Manhattan distance:

The distance between two vectors is the sum of the absolute (unsquared) differences over all coordinates.

$$d_M(x_1, x_2) = |2-2/4| + |4-4/4| + |5-5/4| + |6-6/4| = 51/4 = 12.75$$

$x_1 = (2, 4, 5, 6)$

$x_2 = (2/4, 4/4, 5/4, 6/4)$

$x_3 = (6/4, 4/4, 3/4, 2/4)$

$x_4 = (2.5, 3.5, 4.5, 1)$

0	**12.75**	**13.25**	**6.50**
12.75	**0**	**2.50**	**8.25**
13.25	**2.50**	**0**	**7.75**
6.50	**8.25**	**7.75**	**0**

Matrix of pairwise distances

Correlation distance:

Distance between two vectors is 1-ρ, where ρ is the Pearson correlation of the two vectors.

$$d_C(x_1, x_2) = 1 - \frac{(2-\frac{17}{4})(\frac{2}{4}-\frac{17}{16}) + (4-\frac{17}{4})(\frac{4}{4}-\frac{17}{16}) + (5-\frac{17}{4})(\frac{5}{4}-\frac{17}{16}) + (6-\frac{17}{4})(\frac{6}{4}-\frac{17}{16})}{\sqrt{(2-\frac{17}{4})^2 + (4-\frac{17}{4})^2 + (5-\frac{17}{4})^2 + (6-\frac{17}{4})^2}\sqrt{(\frac{2}{4}-\frac{17}{16})^2 + (\frac{4}{4}-\frac{17}{16})^2 + (\frac{5}{4}-\frac{17}{16})^2 + (\frac{6}{4}-\frac{17}{16})^2}}$$

x_1 = (2, 4, 5, 6)

x_2 = (2/4, 4/4, 5/4, 6/4)

x_3 = (6/4, 4/4, 3/4, 2/4)

x_4 = (2.5, 3.5, 4.5, 1)

0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

Matrix of pairwise distances

Clustering Methods

- Hierarchical
  - Single, Complete and Average Linkage

- Divisive
  - K-means
  - Self Organizing Maps (SOM)

- Dimension Reduction
  - Principal Component Analysis (PCA / SVD)

Hierarchical Clustering

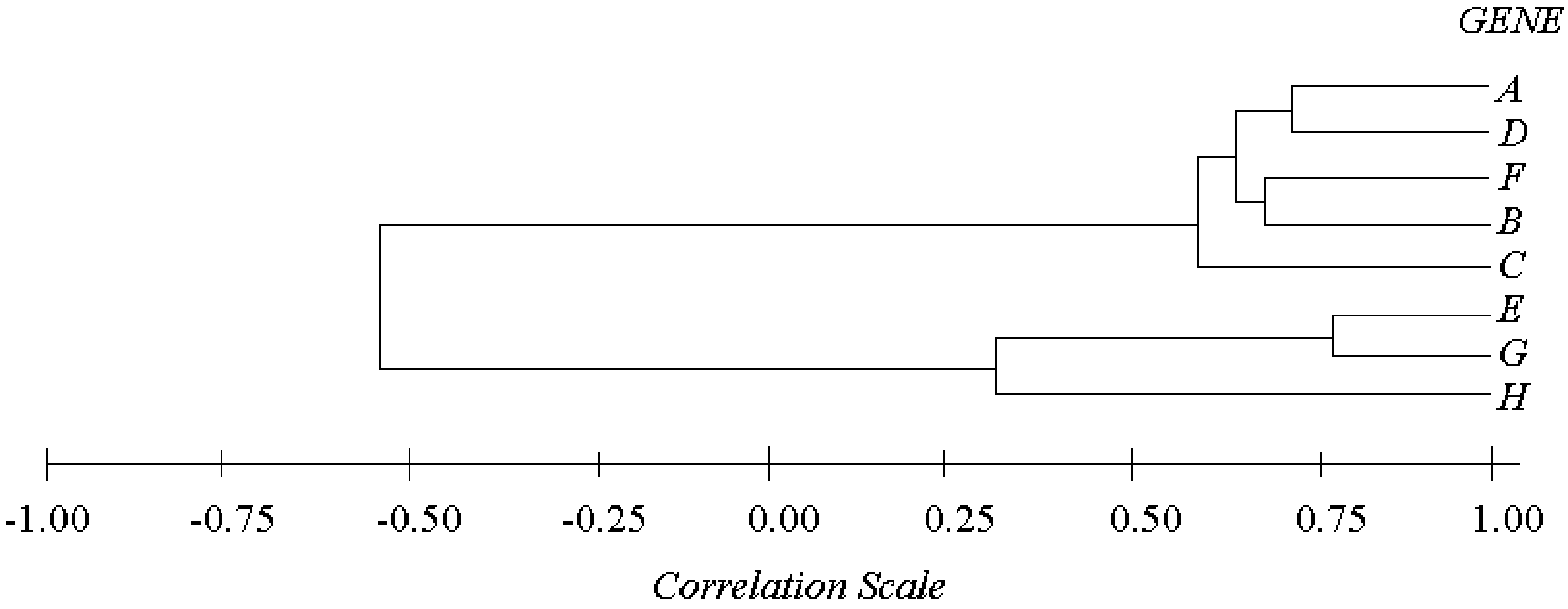
- The first algorithm used in gene expression data clustering (Eisen et al., 1998)
- Algorithm
  - Assign each data point into its own cluster (node)
  - Repeat
    - Select two closest clusters are joined. Replace them with a new parent node in the clustering tree.
    - Update the distance matrix by computing the distances between the new node with other nodes.
  - Until there is only one node (root) left.

Three Ways to Compute Distance Between Groups / Clusters

- Average Linkage: average distance
- Single Linkage: smallest distance
- Complete Linkage: largest distance

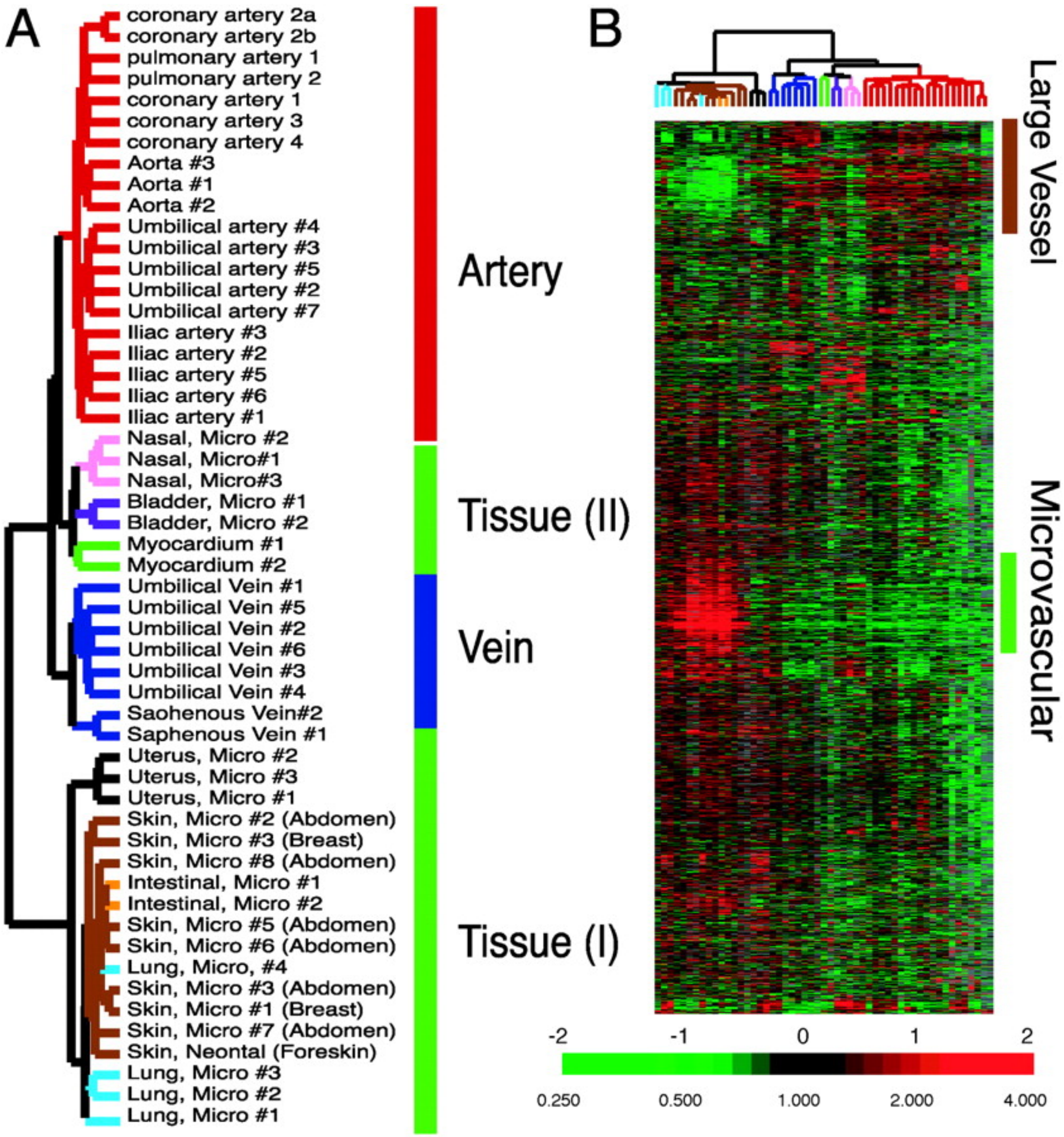
Hierarchical Clustering

Combine most similar genes into agglomerative clusters, build tree of genes



Rainer Breitling, 2005

Hierarchical clustering results



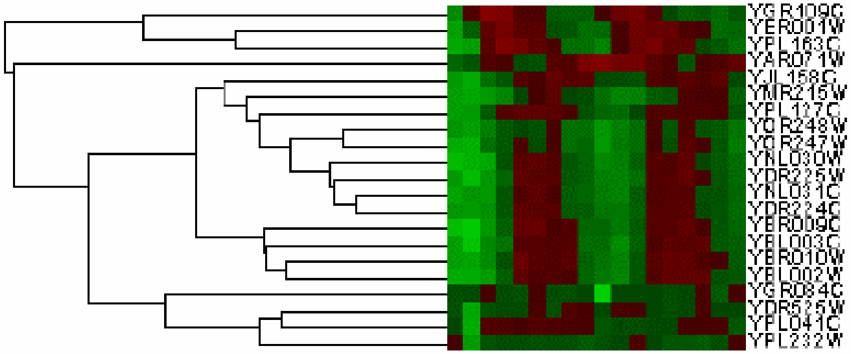
Chi et al., PNAS | **September 16, 2003** | vol. 100 | no. 19 | **10623-10628**

“Endothelial cell diversity revealed by global expression profiling”

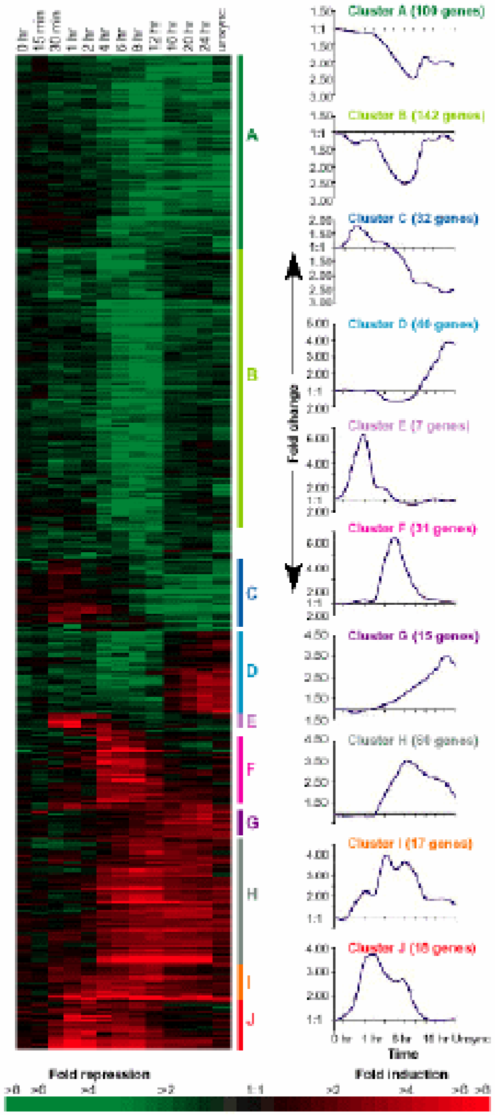
Rainer Breitling, 2005

Iyer et al., Science, Jan 1999:

Genes from functional classes are clustered together.



Cluster dendrogram

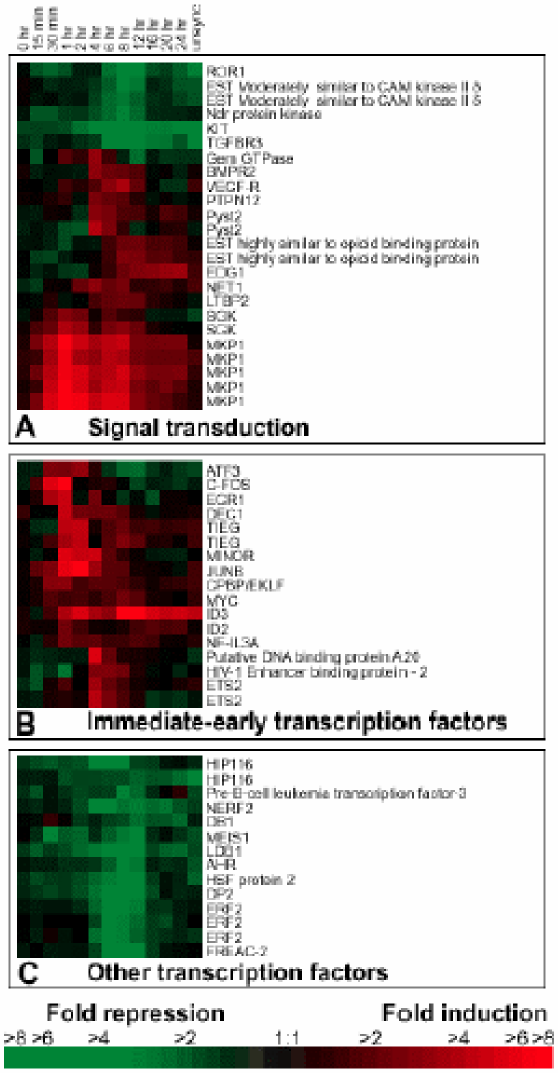


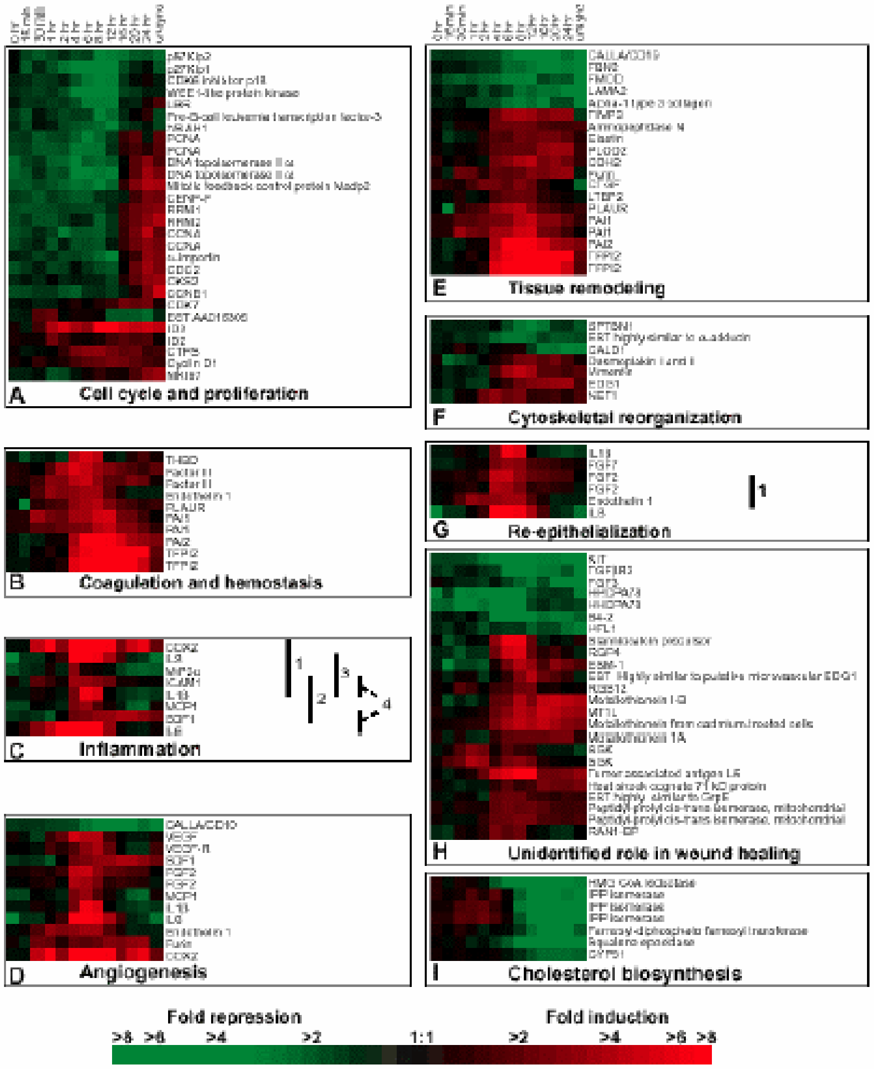
Cluster average patterns

Iyer et al., Science, Jan 1999:

Genes from functional classes are clustered together (sometimes!).

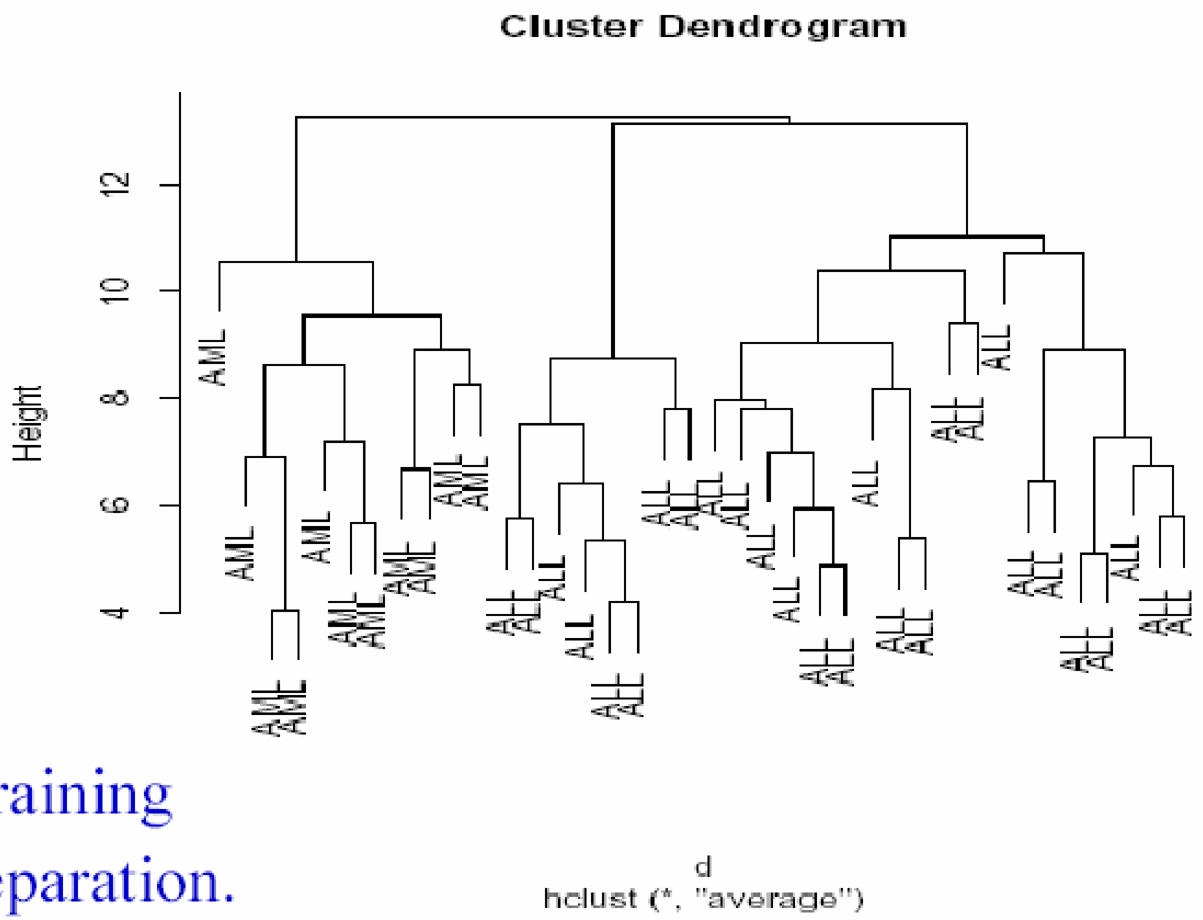
Careful interpretation neccessary!



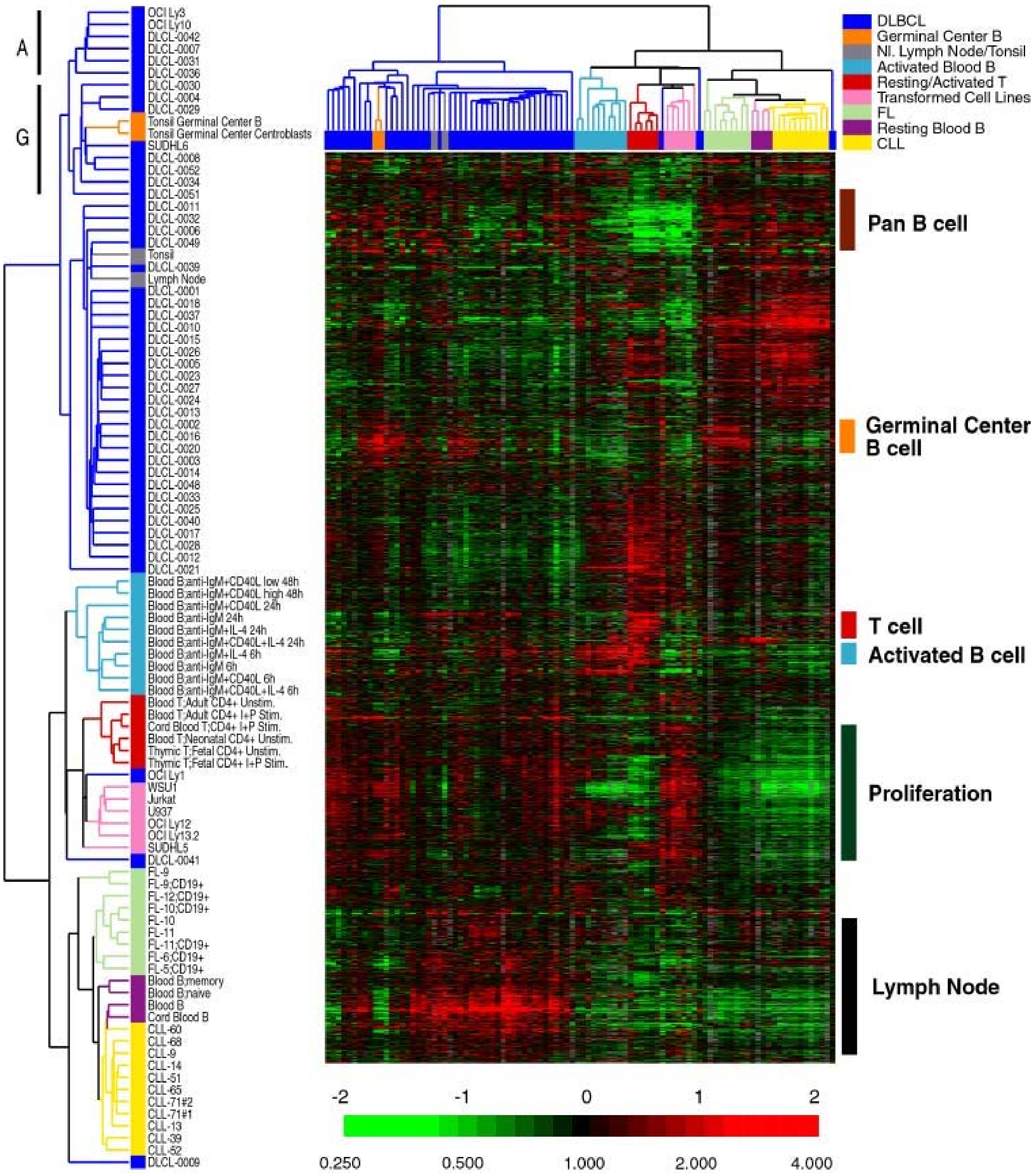


Golub et al.: Leukemia dataset, *http://www.genome.wi.mit.edu/MPR*

3 cancer classes:
25 acute myeloid leukemia (AML),
47 acute lympho-blastic leukemia (ALL), the latter
9 T-cell and 38 B-cell.



Dendrogram for 38 training data shows perfect separation.



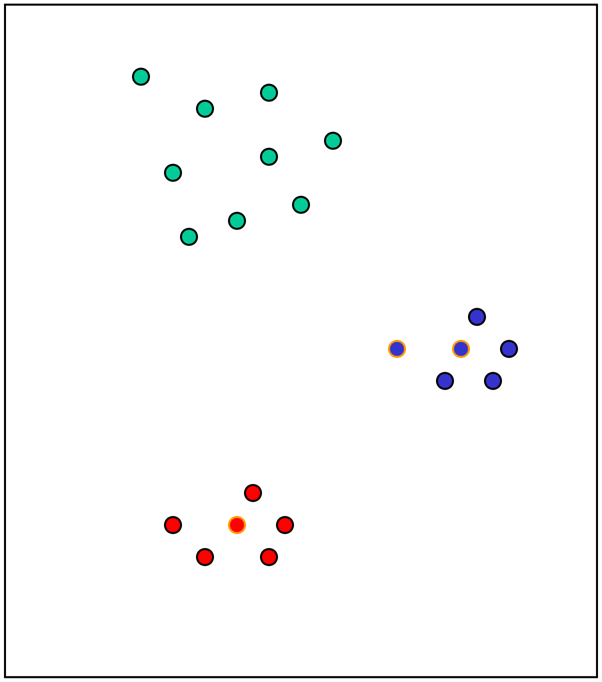
Two-way clustering of genes (y-axis) and cell lines (x-axis) (Alizadeh et al., 2000)

J. Pevsner, 2005

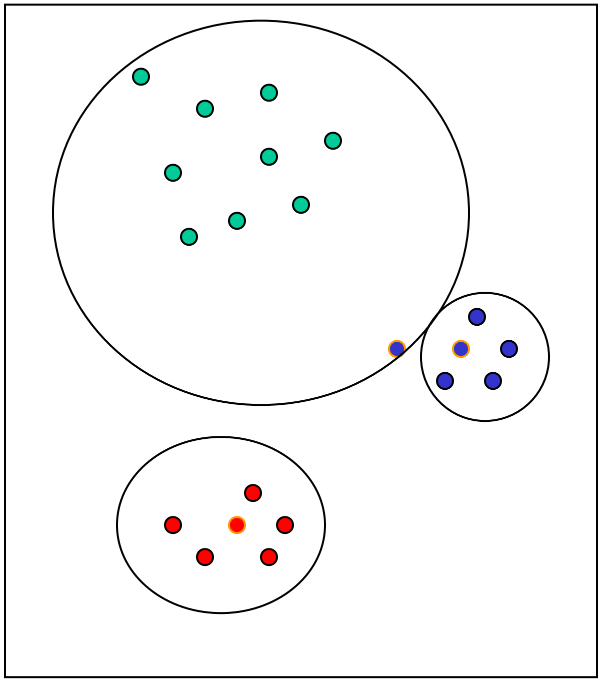
K-Means Clustering

- Randomly select k data points as the centrods of k clusters. Assign points to k clusters with the closest centroids.
- Repeat
  - Compute centroid (mean) of each cluster
  - Assign each point to its nearest cluster (use centroid of clusters to compute distance / similarity)
- Until assignment of data points is not changed

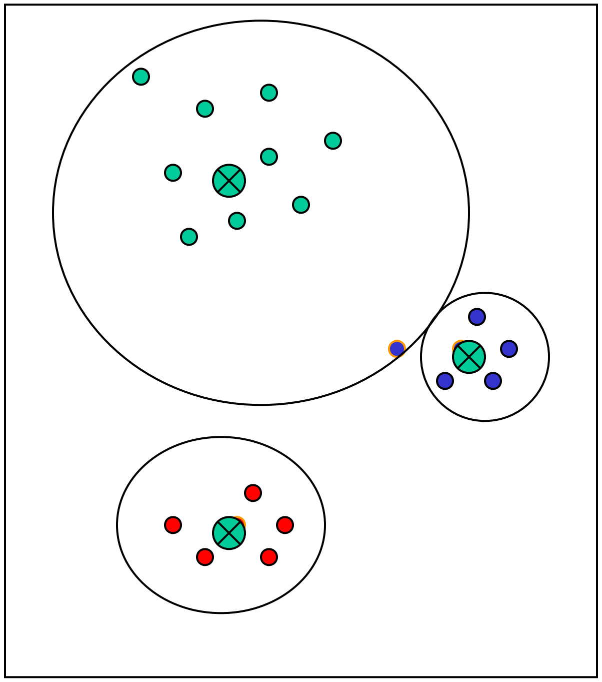
K-Means Clustering Example



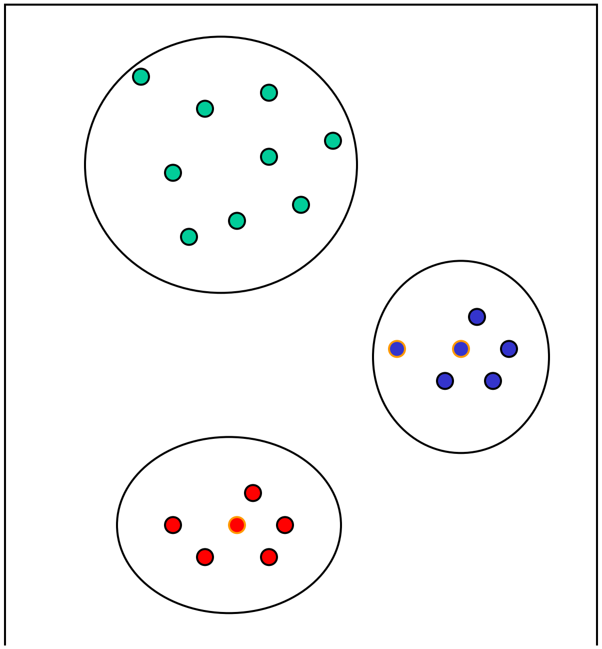
Initialization



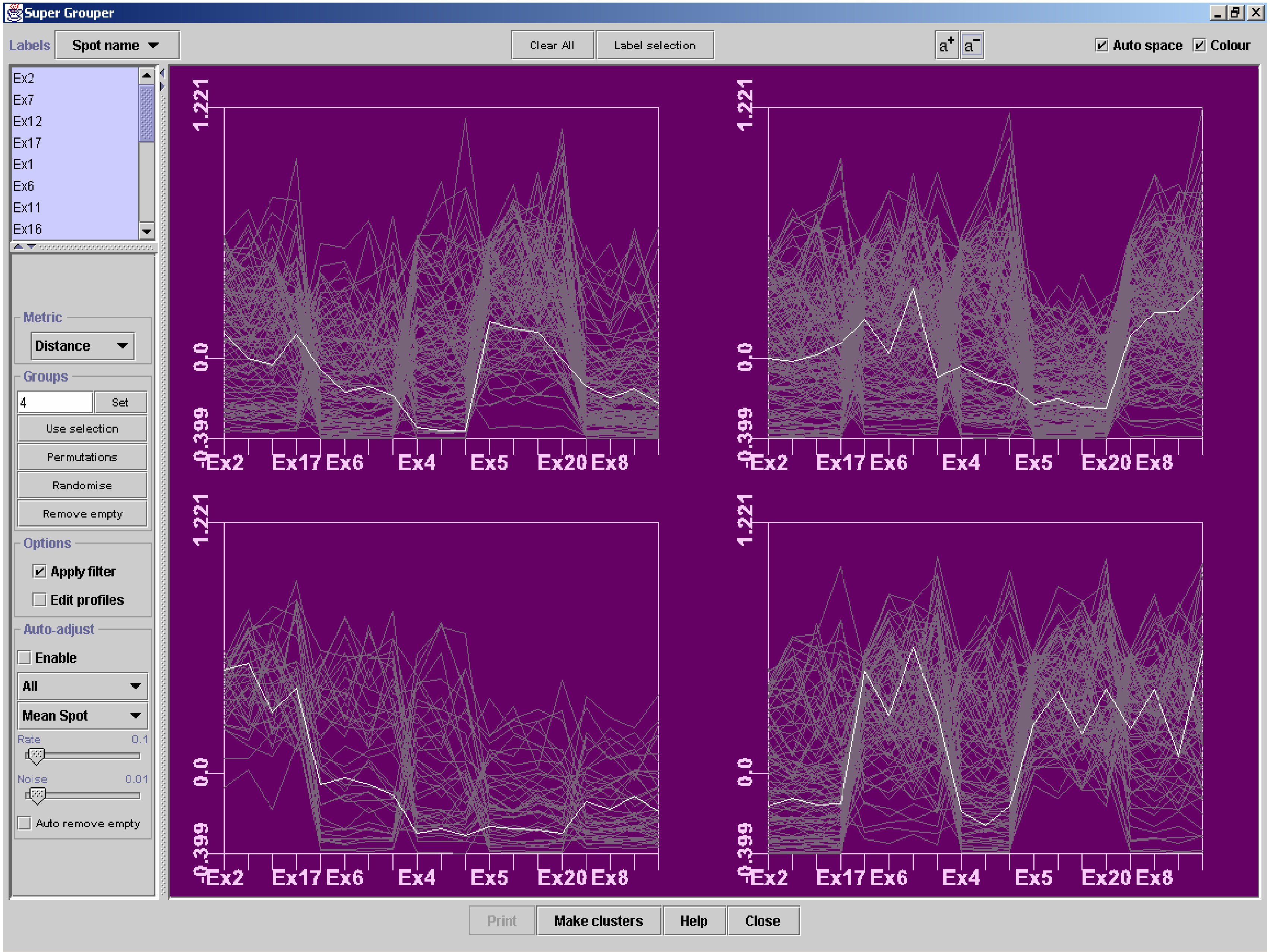
Round 1: Assign data



Compute centroids



K Means Example



Labels: Spot name

☑ Auto space ☑ Colour

Metric: Distance

Groups: 4

Options: ☑ Apply filter ☐ Edit profiles

Auto-adjust: ☐ Enable

All

Mean Spot

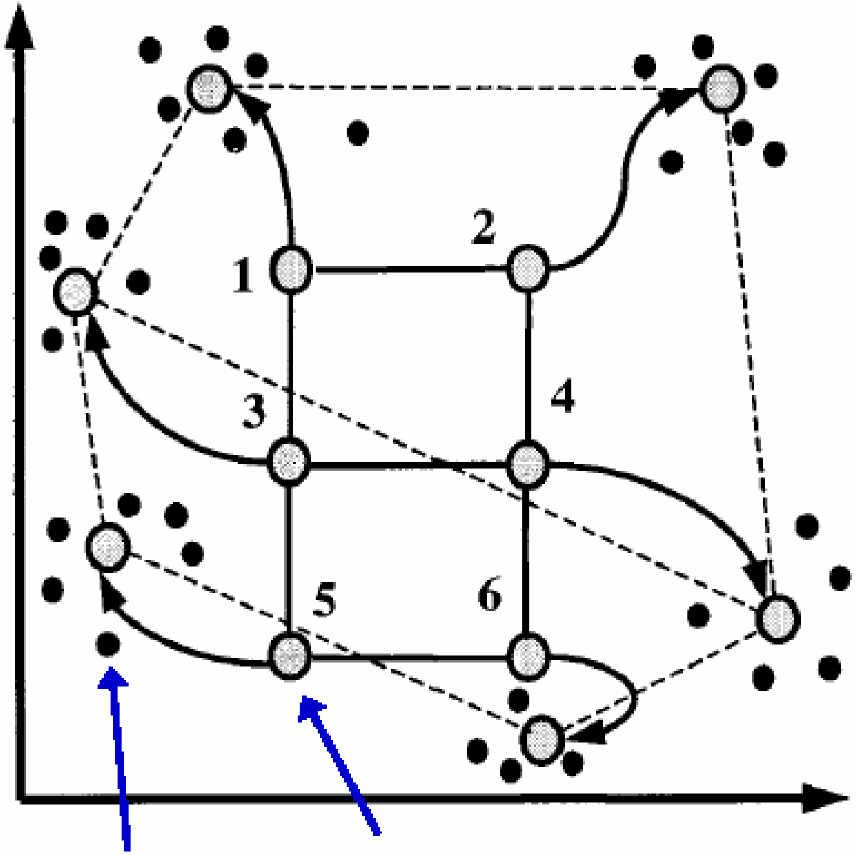
Rate 0.1

Noise 0.01

☐ Auto remove empty

M. Ahmed, 2004

- **SOM's** are similar to k-means, but with additional **constraints**.
- Mapping from input space onto one or two-dimensional array of **k** total nodes.
- Iteration steps (20000-50000):
  - Pick data point P at random
  - Move all nodes in direction of P, the closest node most, the further a node is in network topology, the less
  - Decrease amount of movement with iteration steps



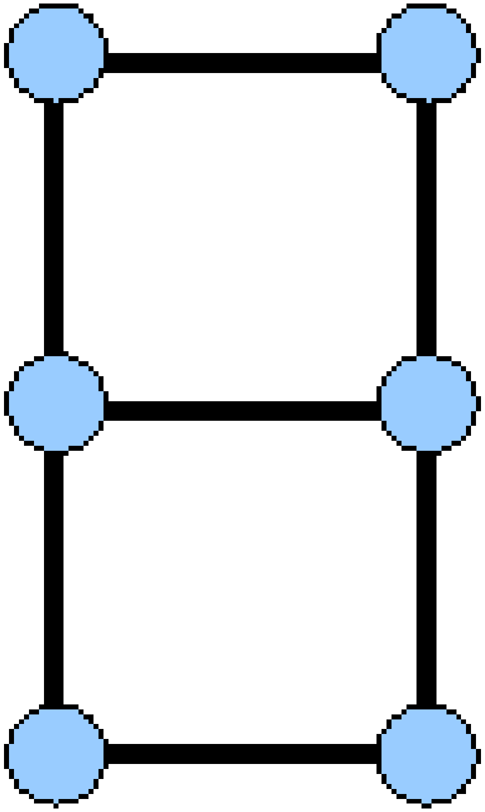
Data point

Node (cluster prototypes)

Tamayo et al. (1999): First use of SOM's for gene clustering from microarrays

Self-organizing maps (SOM)

One chooses a geometry of 'nodes'-for example, a 3x2 grid

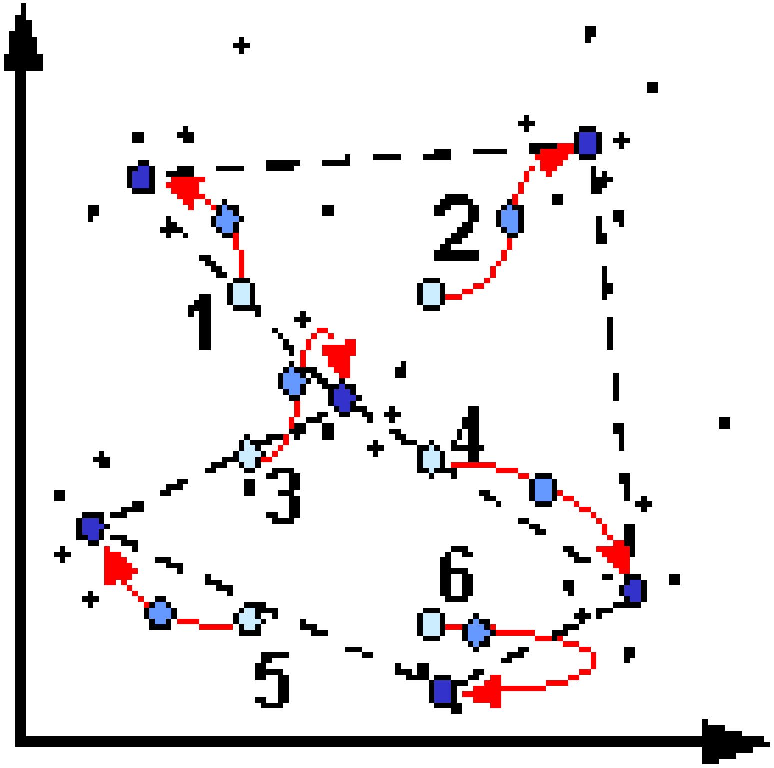


J. Pevsner, 2005

Formerly http://www.genome.wi.mit.edu/MPR/SOM.html

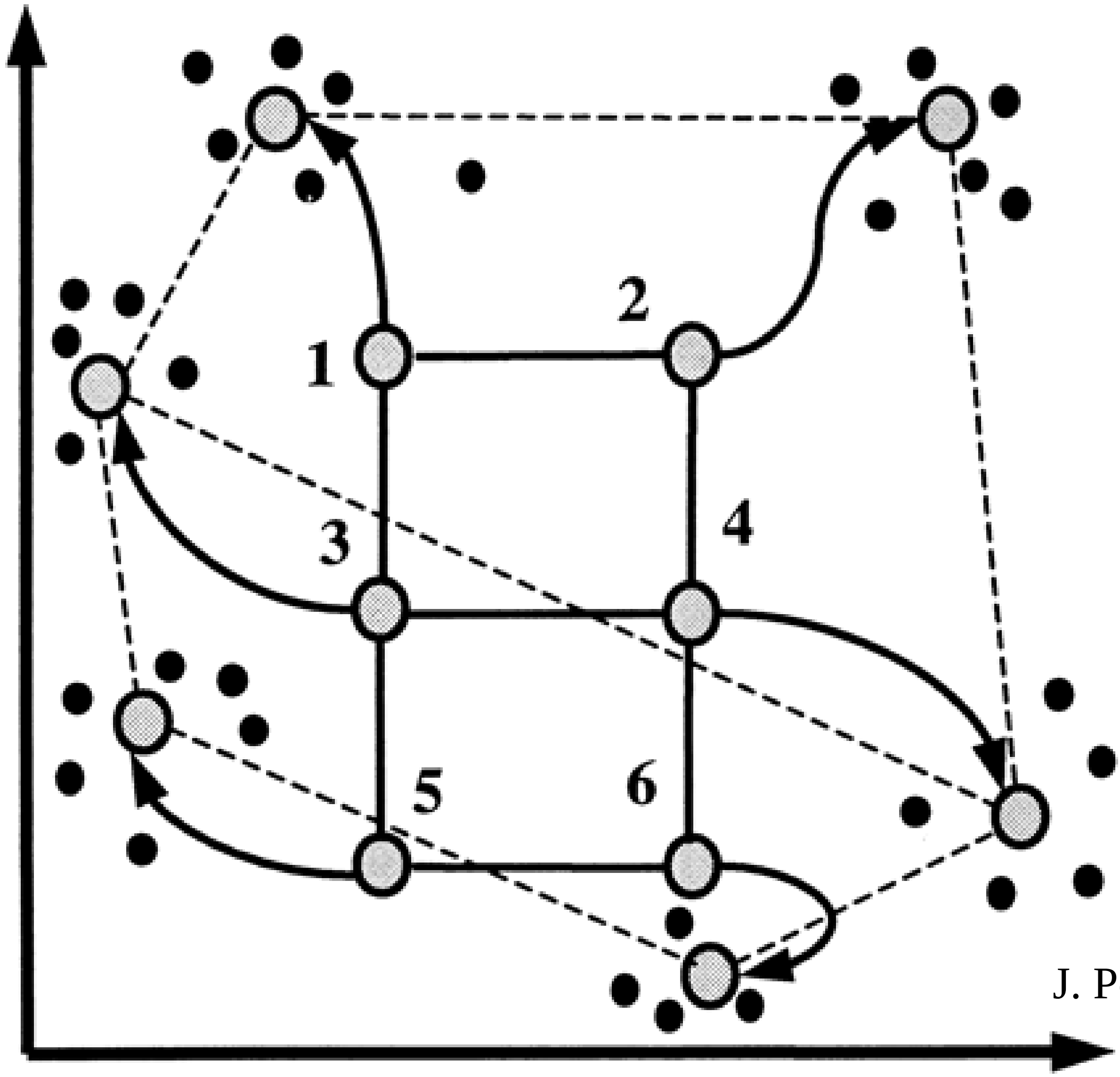
Self-organizing maps (SOM)

The nodes are mapped into k-dimensional space, initially at random and then successively adjusted.



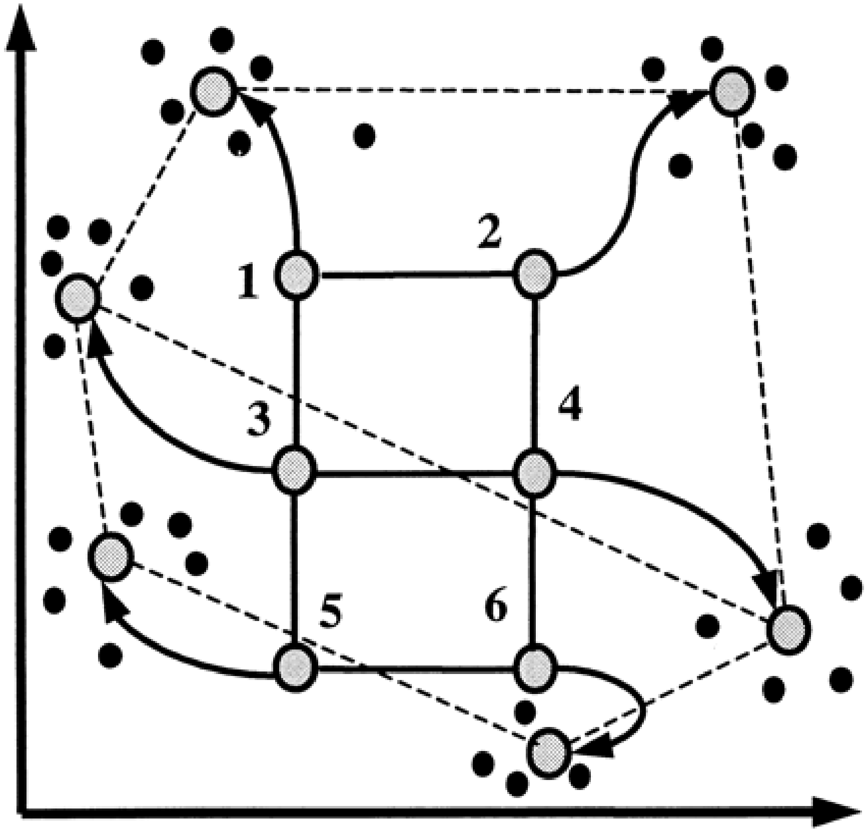
J. Pevsner, 2005

Self-organizing maps (SOM)



J. Pevsner, 2005

Unlike k-means clustering, which is unstructured, SOMs allow one to impose partial structure on the clusters. The principle of SOMs is as follows.
One chooses an initial geometry of "nodes" such as a 3 x 2 rectangular grid (indicated by solid lines in the figure connecting the nodes). Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.



J. Pevsner, 2005

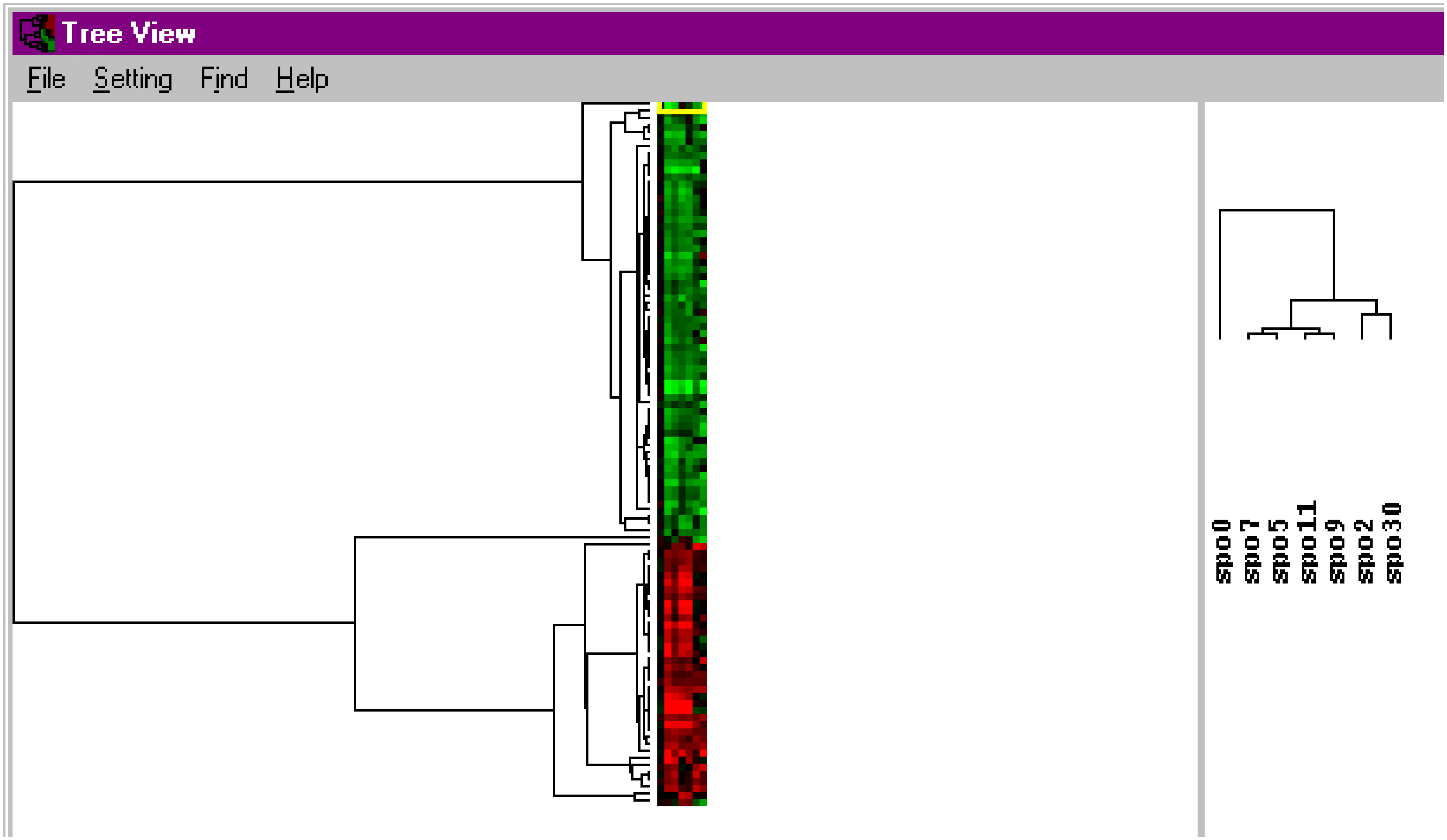
Self-organizing maps (SOM)

To download GeneCluster:

http://www.genome.wi.mit.edu/MPR/software.html

J. Pevsner, 2005

Cluster and TreeView (Visualization)



TreeView is associated with GeneCluster software.

J. Pevsner, 2005

One Key Issue of Clustering

- How many clusters are there?

  Unfortunately, there is no general rule. Usually one tries different number of clusters. Use each number (K) to cluster data many times. If the clustering results are rather consistent, K may be a good choice.

Principal components analysis (PCA)

An exploratory technique used to reduce the dimensionality of the data set to 2D or 3D

For a matrix of *m* genes × *d* samples, create a new covariance matrix of size *d* x *d*

Thus transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).

Also called SVD (Singular Value Decomposition)

J. Pevsner, 2005

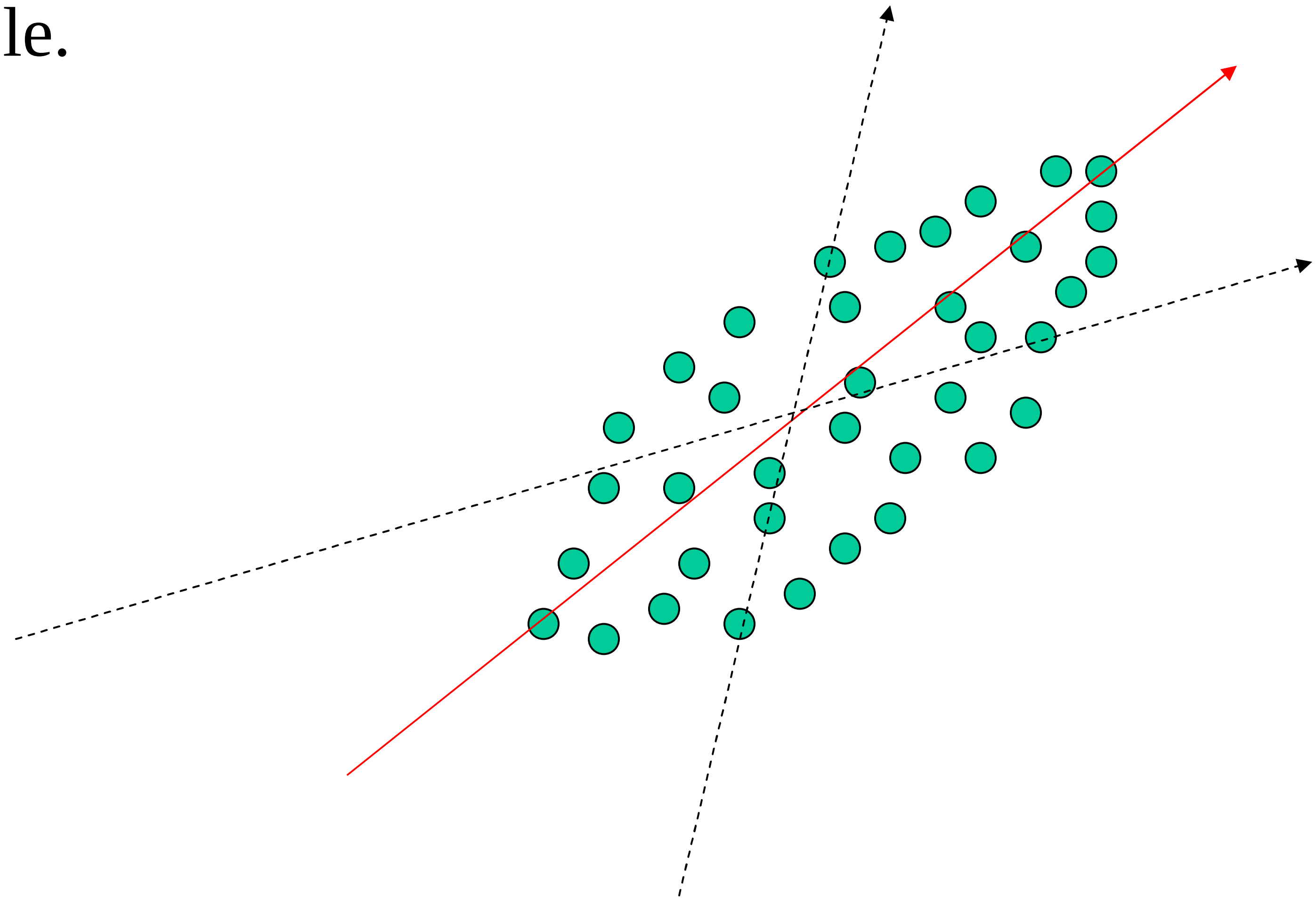
Objectives of PCA

- Reduce dimensionality
- Determine the linear combination of variables
- Choose the most useful variables (features)
- Visualize multidimensional data
- Identify groups of objects (e.g. genes/samples)
- Identify outliers

J. Pevsner, 2005

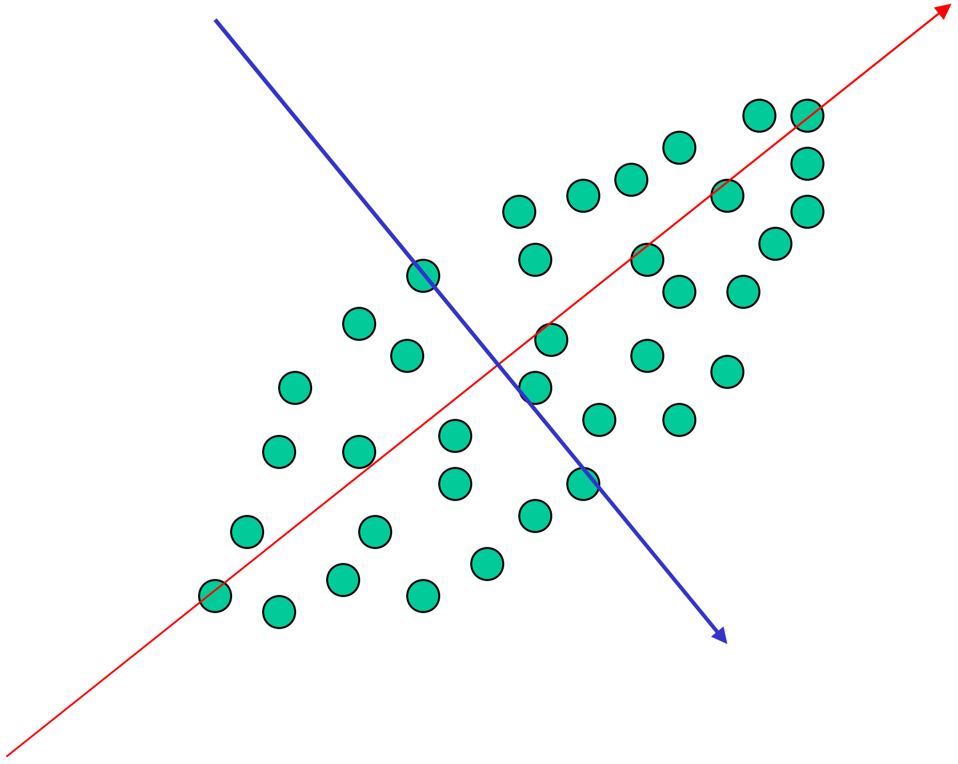
Basic Idea of PCA

Goal: Map data points into a few dimension while trying to preserve the variance of data as much as possible.



Basic Idea of PCA

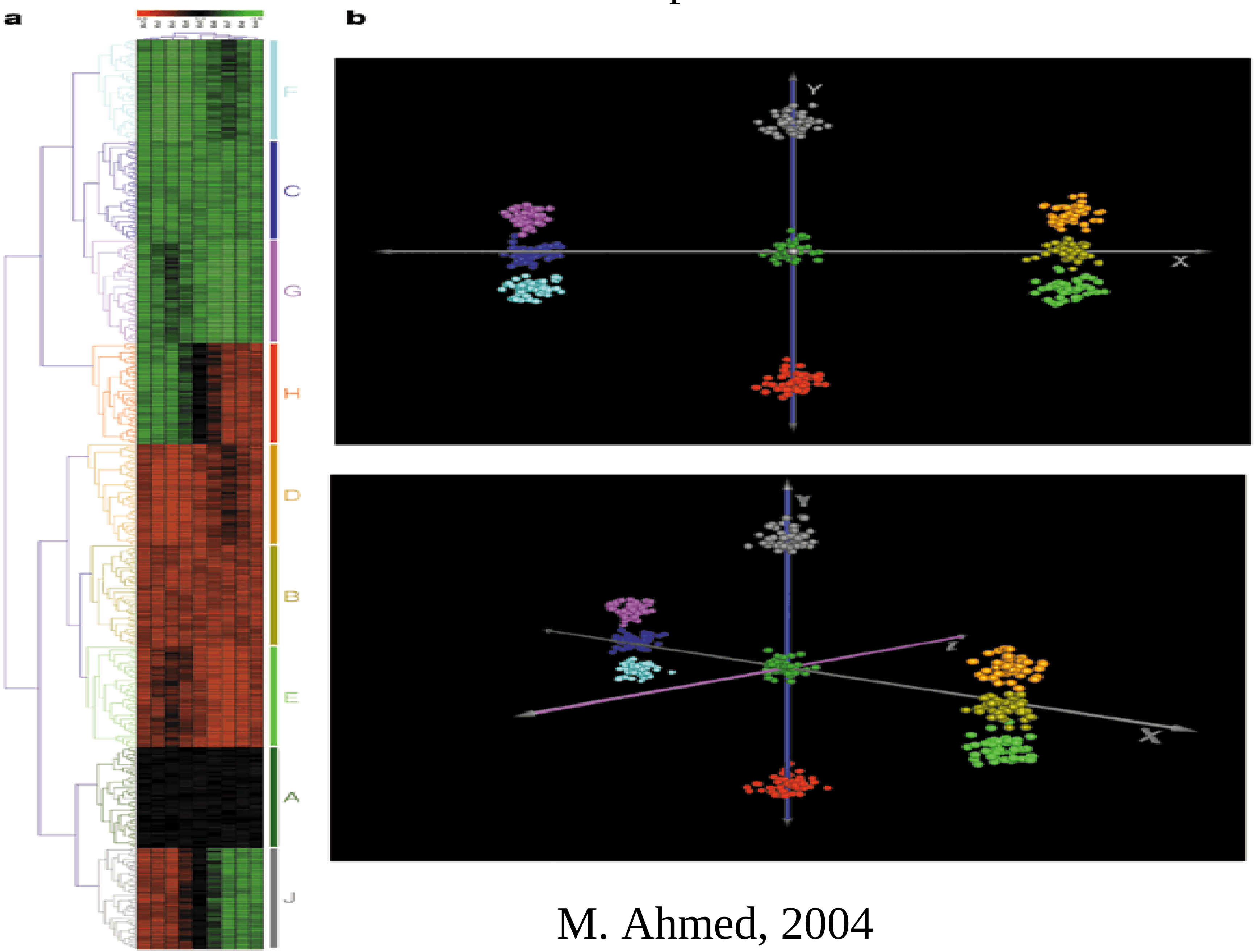
Goal: Map data points into a few dimension while trying to preserve the variance of data as much as possible.



PCA Method

- Given a data matrix X (n × d, n data points, d dimension).
- Normalize X by subtracting mean from each data point
- Construct a covariance matrix $C=X^TX / n$. (d × d)
- Calculate the eigenvectors and eigenvalues of the covariance matrix C. ($C v = v \lambda$).
- Sort eigenvectors by eigenvalues in decreasing order
- Map data point *x* to the direction *v* by computing the dot product.
- A well studied problem. Implementation in many software such as MatLab.

PCA Example



M. Ahmed, 2004

Nature Reviews | Genetics

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

Classification Methods

- Decision Tree
- K-nearest neighbor
- Neural Nets
- Support Vector Machines (SVM)

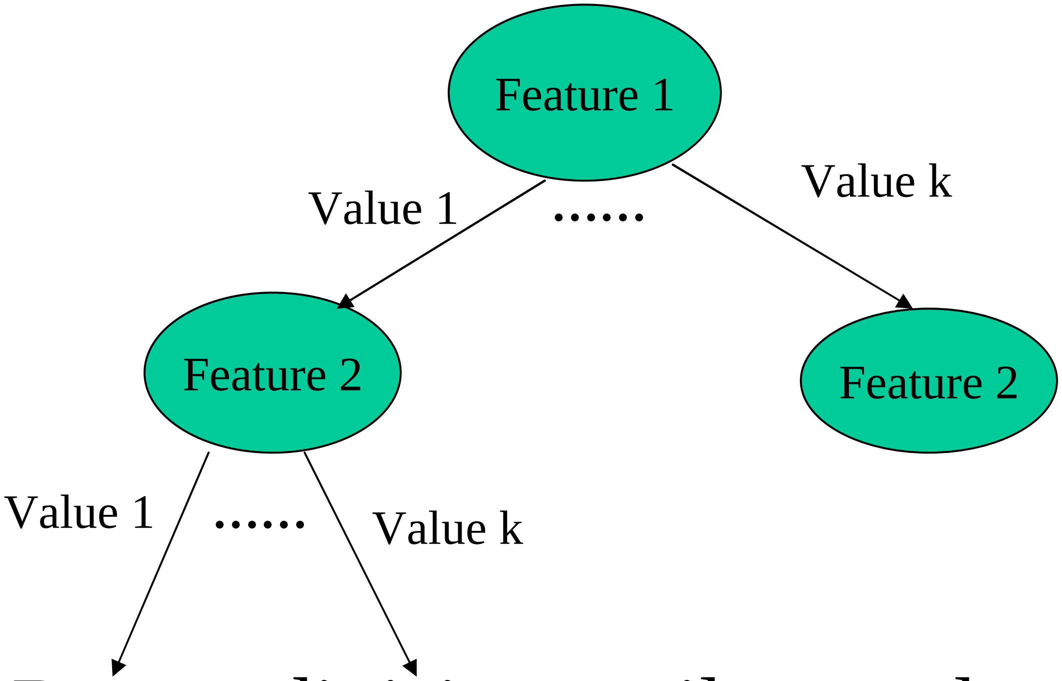
Tradeoff:

Decision tree is easy to understand, but usually less accurate

Neural Nets and SVM have higher accuracy, but hard to understand the model (black box).

Decision Tree Classification

- Divide and Conquer Technique



- Repeat division until most data points in the in the nodes are in the same class
- What is the key issue here?

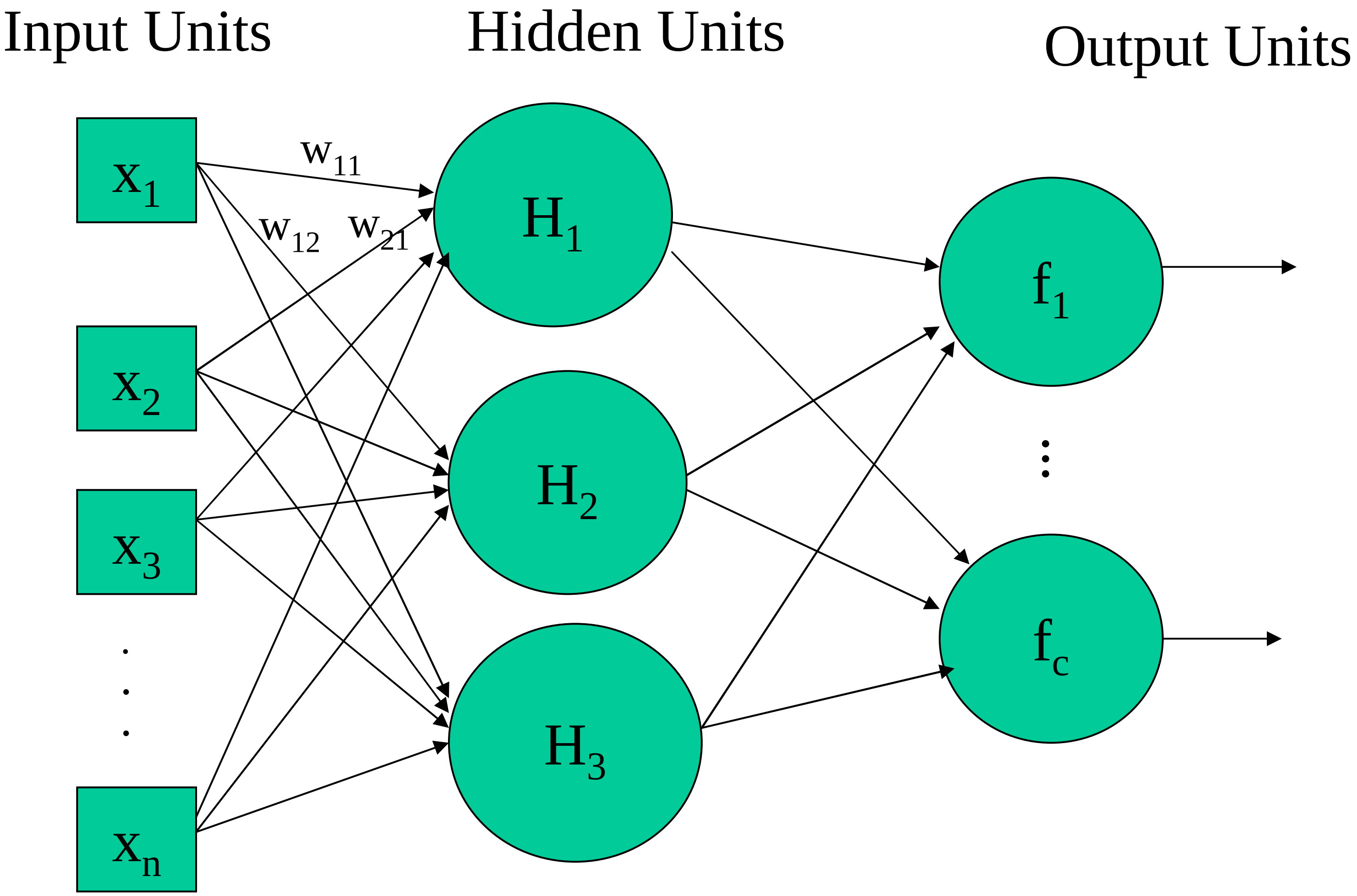
Key Issue of Decision Tree

- Which feature is selected at each step?
- We want to select most informative feature at each step
- Use Information Gain Measure
- Use a feature to divide data and check how entropy changes. Select the feature reducing entropy most.

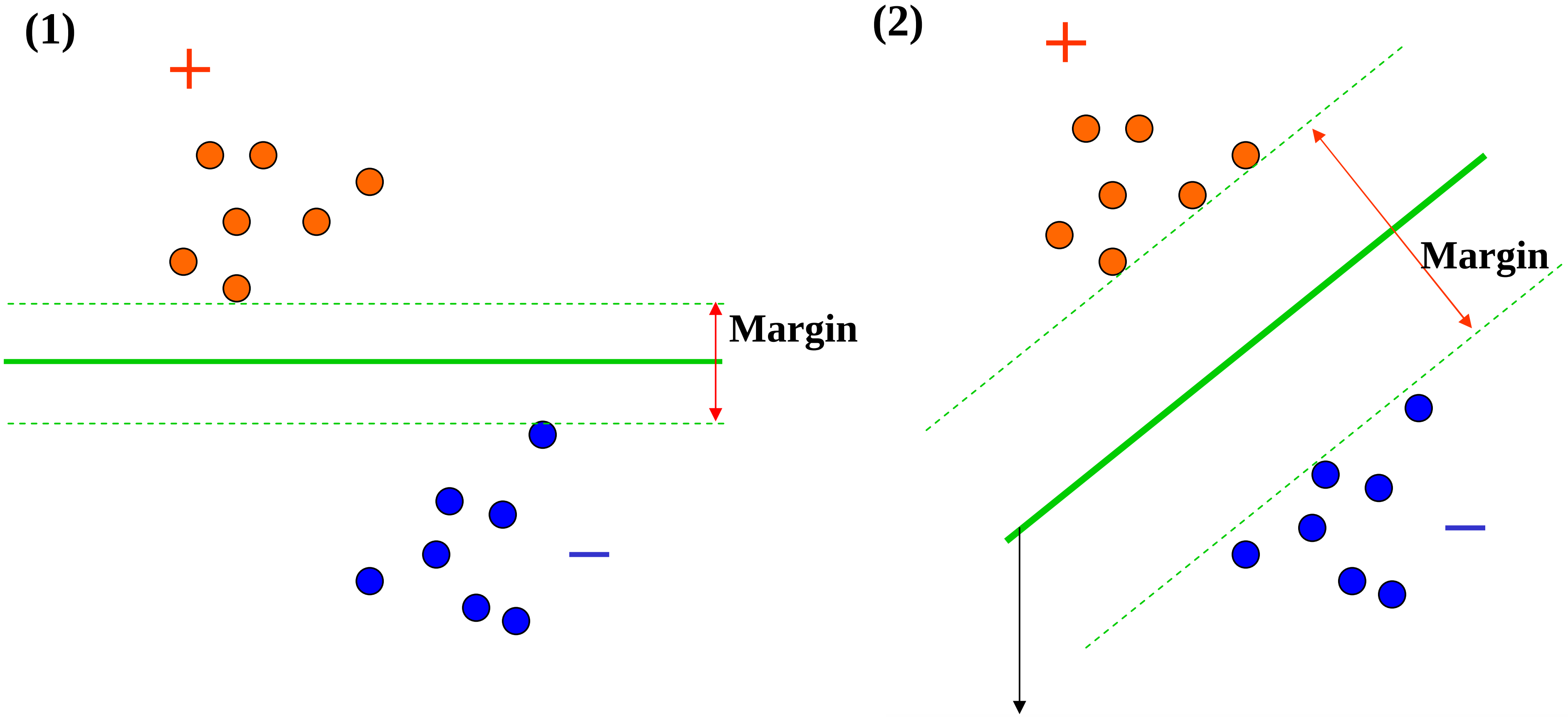
K Nearest Neighbor (KNN)

- Given a data *x*, compute its distance (or similarity) to all data points with known classes.
- Select k closest neighbors
- Use majority classes of the k neighbors to predict the label of *x*.

Neural Network



Support Vector Machine Learning



$$f(x) = \sum_{x_i \in S} \alpha_i y_i K(x, x_i) + b$$

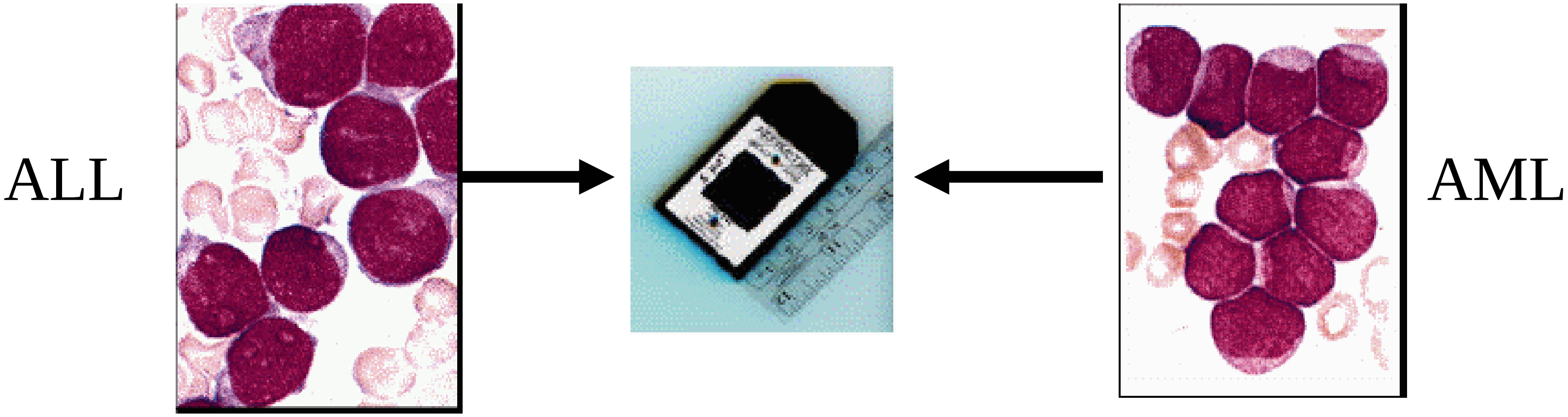
K is Gaussian Kernel: $e^{-\gamma ||x-y||^2}$.

Two Classification Problems

- Classify samples using expression levels of a set of genes as features. (discriminate different known cell types. e.g. tumor cell vs normal cell).
- Classify genes using expression levels of genes across multiple samples or experiments. A gene class may correspond to a functional category or biological process.

A Sample Classification Example

- Leukemia: Acute Lymphoblastic (ALL) vs Acute Myeloid (AML), Golub et al, Science, v.286, 1999
  - 72 examples (38 train, 34 test), about 7,000 genes
  - Gene expression values are features



Visually similar, but genetically very different

Y. Guo, V. Curan, H. Morris, 2005

Results on the Test Data

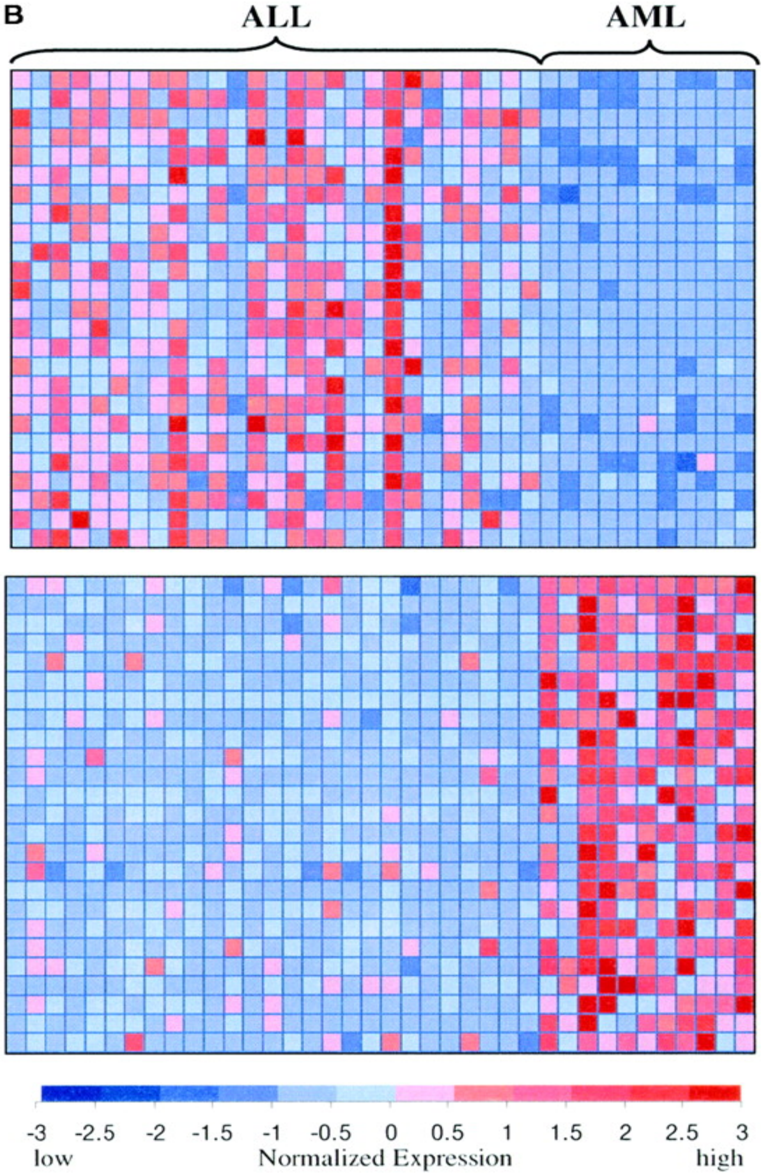
- Select genes (Feature selection)
- Best neural net model used 10 genes per class
- Evaluation on test data (34 samples) gives 1 or 2 errors (94-97% accuracy) using most classification methods

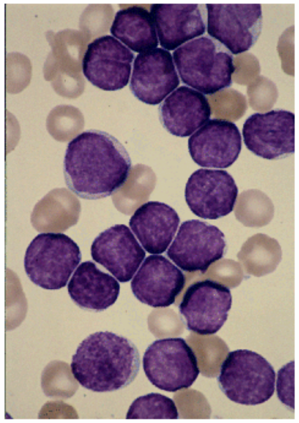
Y. Guo, V. Curan, H. Morris, 2005

Classical study of cancer subtypes

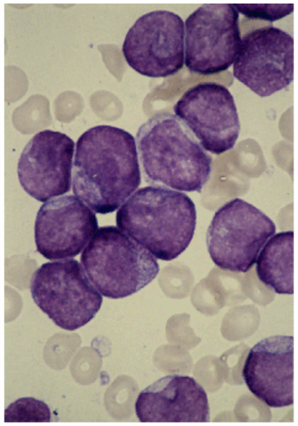
Golub et al. (1999)

identification of diagnostic genes





ALL
acute lymphoblastic leukemia
(lymphoid precursors)



AML
acute myeloid leukemia
(myeloid precursor)

Rainer Breitling, 2005

Some Common Feature Selection Methods

- Information Gain
- Forward Selection
- Backward Selection

Outline

- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

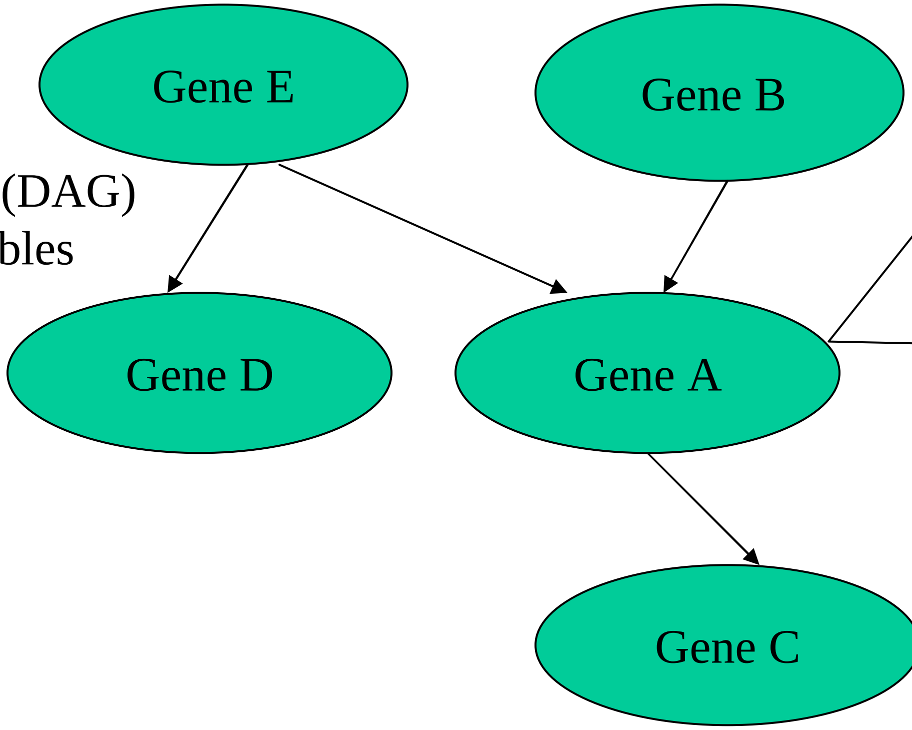
Discovery of Regulatory Mechanism of Gene Expression

- A long term goal of Systems Biology is to discover the causal processes among genes, proteins, and other molecules in cells
- Can this be done (in part) by using data from high throughput experiments, such as microarrays?
- Clustering can group genes with similar expression patterns, but does not reveal structural relations between genes
- Bayesian Network (BN) is a probabilistic framework capable of learning complex relations between genes

Bayesian Networks

- A Bayesian Network (BN) is a graphical representation of a probability distribution

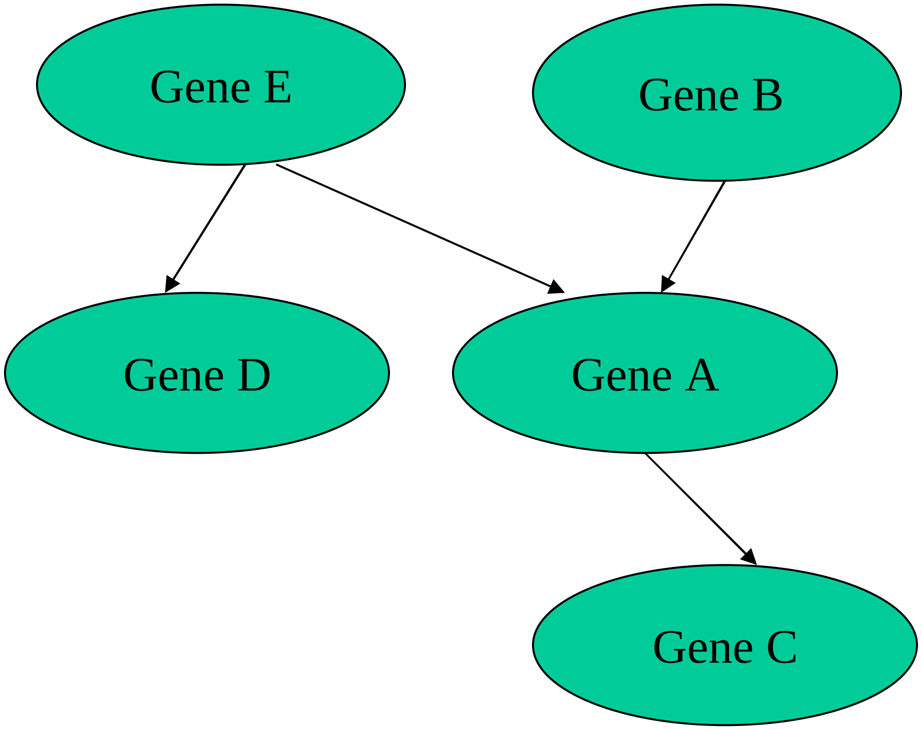
Qualitative Part:
Directed acyclic Graph (DAG)
• Nodes – random variables
•Edges – direct (causal) influence



E	B	P(A|E,B) 1	0
0	1	0.9	0.1
1	0	0.2	0.8
1	1	0.9	0.1
0	0	0.01	0.99

Quantitative part
•Local conditional probability

Key Features of BN



- Conditional Independence (decomposition, simplification)

P(A, B, C, D, E) = P(E) * P(B) * P(D|E) * P(A|E, B) * P(C|A)

If each variable can have two different values, how many parameters are required represent P(A, B, C, D, E)?

How many parameters are needed using Bayesian network at the left?

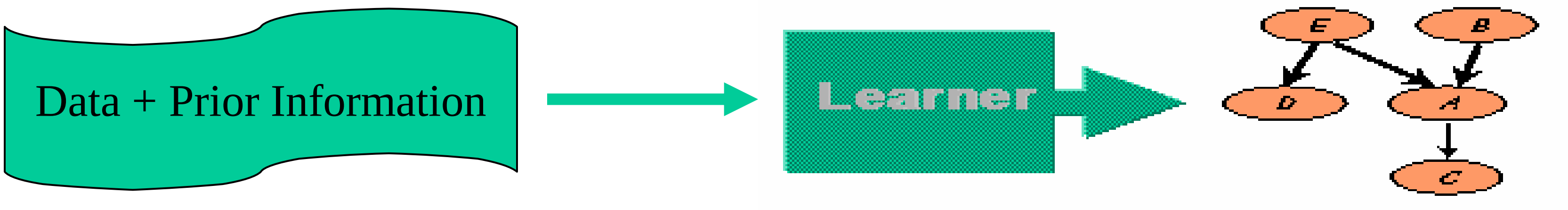
Advantages of BN

- Compact & intuitive representation
- Captures causal relationships
- Efficient model learning (parameters and structure)
- Deals with noisy data
- Integration of prior knowledge
- Effective inference algorithms

N. Friedman, 2005

Learning BN from Gene Expression Data

Measured expression level of each gene (discretized) → Random variables Affecting on another



Data + Prior Information

Learn parameters (conditional probabilities) from data
Learn structure (casual relation) from data
Make inference given a learned BN model

N. Friedman, 2005

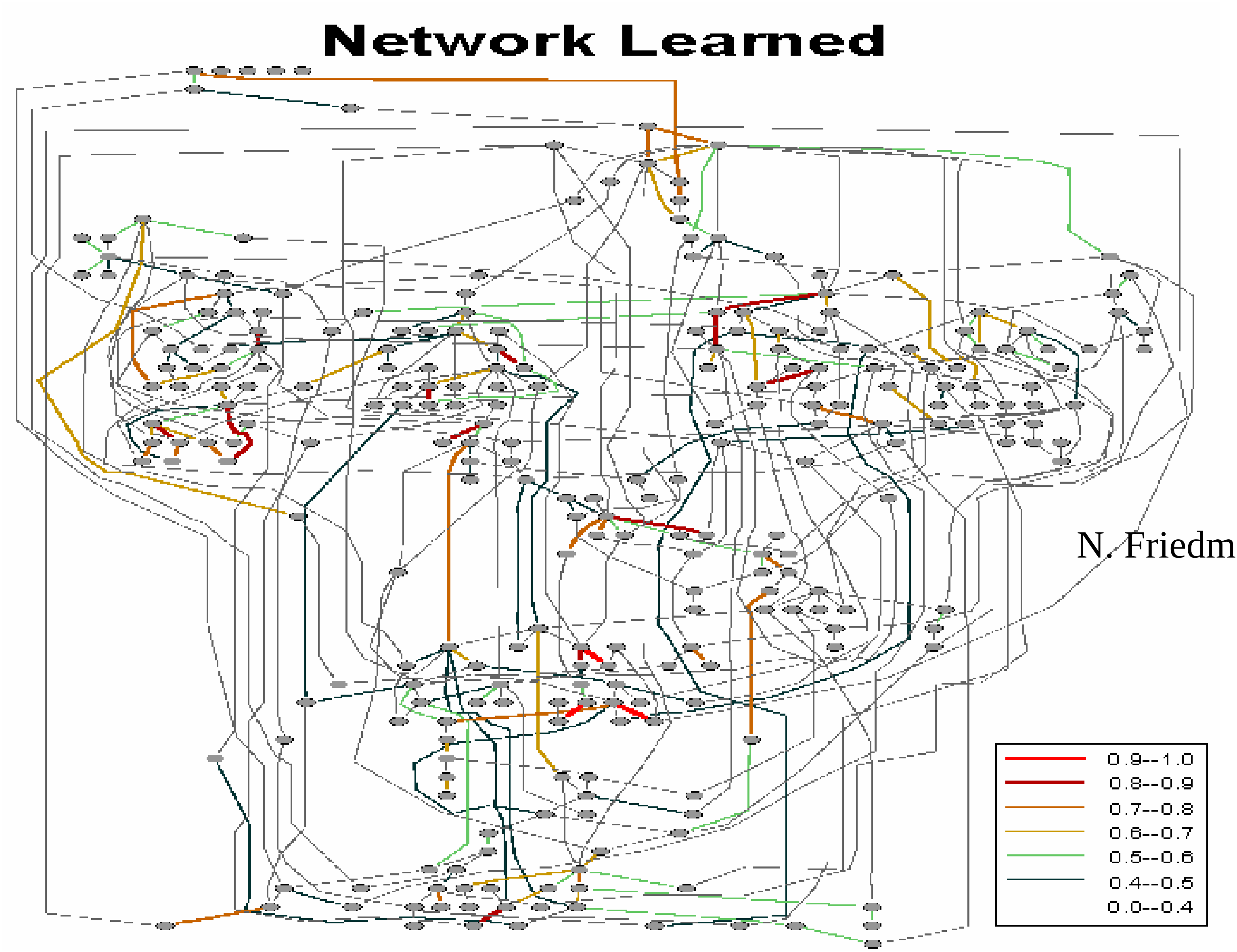
Challenges of Gene Bayesian Network

- Massive number of variables (genes)
- Small number of samples (dozens)
- Sparse networks (only a small number of genes directly affect one another)
- Two crucial aspects: computational complexity and statistical significance of relations in learned models

N. Friedman, 2005

Solutions

- Sparse candidate algorithm (by Nir Friedman): Choose a small candidate set for direct influence for each gene. Find optimal BN constrained on candidates. Iteratively improve candidate set.
- Bootstrap confidence estimate: use re-sampling to generate perturbations of training data. Use the number of times a relation (or feature) is repeated among networks learned from these datasets to estimate confidence of Bayesian network features.



N. Friedman, 2005

Data: 76 samples of 250 cell-cycle related genes in yeast genome
Discretized into 3 expression levels. Run 100 bootstrap using sparse learning algorithm.
Compute the confidence of features (relations). Most high confident relations make bio-senses.

Outline

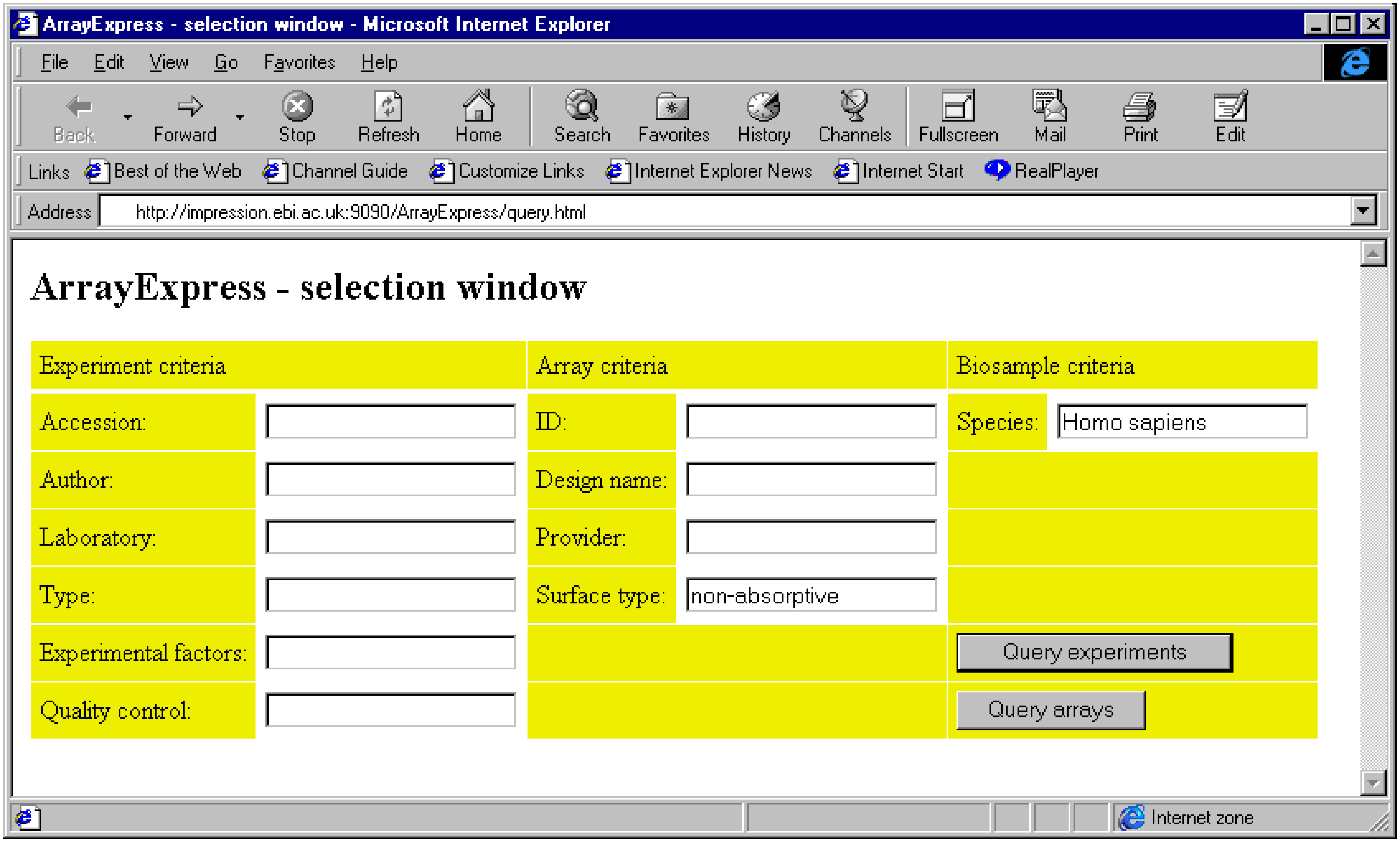
- Introduction to gene expression and DNA microarray
- Data normalization
- Analysis of differential gene expression
- Clustering
- Classification
- Inference of gene regulatory networks
- Databases and software

Major Public Gene Expression Databases

- 3D-GeneExpression Database
- ArrayExpress
- BodyMap
- ChipDB
- ExpressDB
- Gene Expression Omnibus (GEO)
- Gene Expression Database (GXD)
- Gene Resource Locator
- GeneX
- Human Gene Expression Index (HuGE Index)
- RIKEN cDNA Expression Array Database (READ)
- RNA Abundance Database (RAD)
- Saccharomyces Genome Database (SGD)
- Standford Microarray Database (SMD)
- TissueInfo
- yeast Microarray Global Viewer (yMGV)

Y. F. Leung, 2005

ArrayExpress - queries



Address: http://impression.ebi.ac.uk:9090/ArrayExpress/query.html

ArrayExpress - selection window

Experiment criteria		Array criteria		Biosample criteria	
Accession:		ID:		Species:	Homo sapiens
Author:		Design name:			
Laboratory:		Provider:			
Type:		Surface type:	non-absorptive		
Experimental factors:				Query experiments	
Quality control:				Query arrays	

H. Parkinson, 2002

Major Image Analysis Software

- AIDA array
- ArrayPro
- ArrayVision
- Dapple
- F-scan
- GenePix Pro 3.0.5
- ImaGene 4.0
- Iconoclust
- Iplab
- Lucidea Automated Spotfinder
- Phoretix Array3
- P-scan
- QuantArray 3.0
- ScanAlyze 2
- Spot
- TIGR Spotfinder
- UCSF Spot

Some Common Image Analysis Software

- ScanAlyze 2 (Mike Eisen, LBNL)
- GenePix Pro 3.0.5 (Axon Instruments)
- QuantArray 3.0 (Packard Instrument)
- ImaGene 4.0 (Biodiscovery)

Major Data Mining Software

- AIDA Array
- AMADA
- ANOVA program for microarray data
- ArrayMiner
- arraySCOUT
- ArrayStat
- BRB ArrayTools
- CHIPSpace
- Cleaver
- CIT
- CLUSFAVOR
- Cluster
- Cyber T
- DNA-arrays analysis tools
- dchip
- Expression Profiler
- Expressionist
- Freeview & FreeOView
- Gene Cluster
- GeneLinker Gold
- GeneMaths
- GeneSight
- GeneSpring
- Genesis
- Genetraffic
- J-Express
- MAExplorer
- Partek
- R cluster
- Rosetta Resolver
- SAM
- SpotFire Decision Site
- SNOMAD
- TIGR ArrayViewer
- TIGR Multiple Experiment Viewer
- TreeView
- Xcluster
- Xpression NTI

Y. F. Leung, 2005

Comprehensive Software

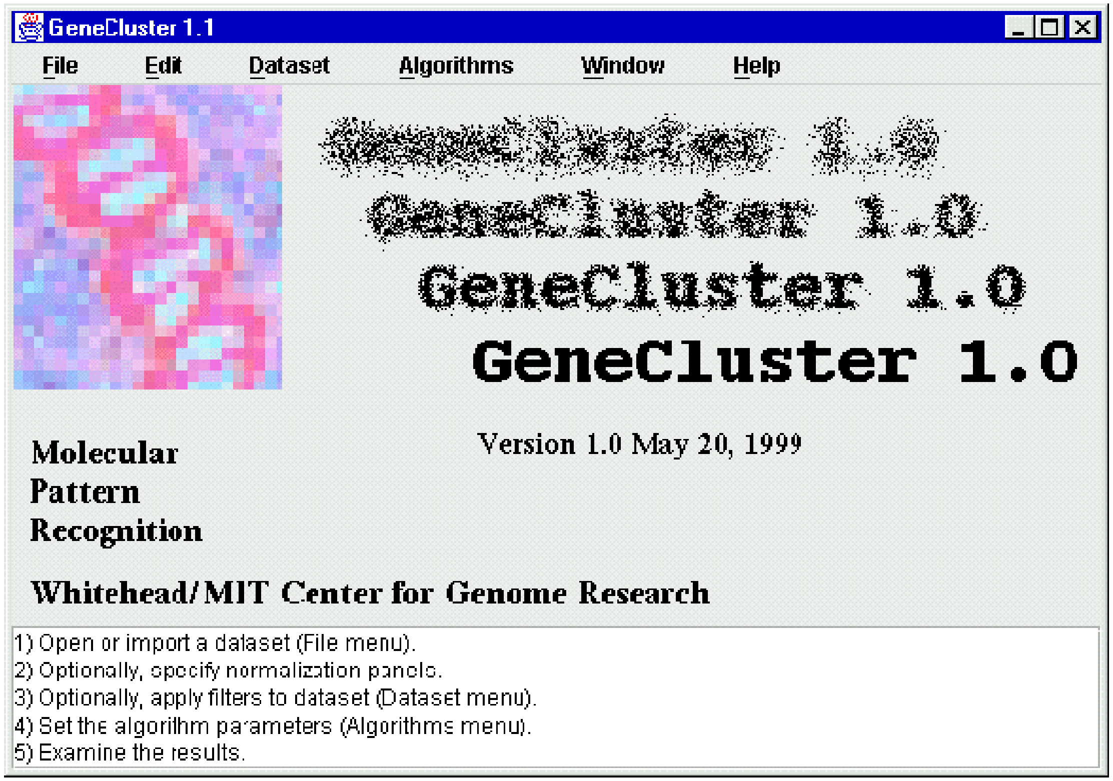
- Definition: Software incorporate many different analyses for different stage in a single package.
- Examples
  - **Cluster (Mike Eisen, LBNL)**
  - GeneMaths (Applied Maths)
  - GeneSight (Biodiscovery)
  - GeneSpring (Silicon Genetics)

Specific Analysis Software

- Definition: Software performing a few/ one specific analysis
- Examples
  - GeneCluster (Whitehead Institute Centre for genome research)
  - INCLUSive - INtegrated CLustering, Upstream Sequence retrieval and motif Sampler (Katholieke Universiteit Leuven)
  - SAM – Significance Analysis of Microarrays (Stanford University)

GeneCluster

- GeneCluster – performing normalization, filter and SOM

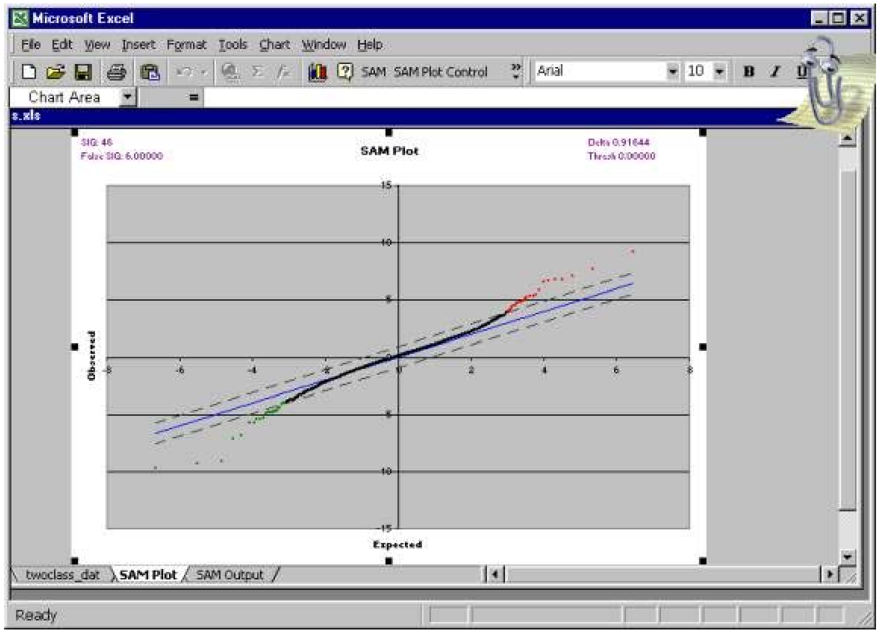


1) Open or import a dataset (File menu).
2) Optionally, specify normalization panels.
3) Optionally, apply filters to dataset (Dataset menu).
4) Set the algorithm parameters (Algorithms menu).
5) Examine the results.

Y. F. Leung, 2005

Inclusive

- INCLUSive - INtegrated CLustering, Upstream Sequence retrieval and motif Sampler
- SAM – finding statistical significant differentially expressed gene



SIG: 46
False SIG: 6.00000

Delta 0.91644
Thresh 0.00000

Y. F. Leung, 2005

Free, Useful Software

- **Michael Eisen's Cluster (Windows only)** (http://rana.lbl.gov/EisenSoftware.htm)
- M. de Hoon's Cluster 3.0 (all OS) (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/)
- Tree viewing (links on same site)
  - Java Treeview
  - Maple Tree (also Michael Eisen's lab)
  - Free View

General Statistics software

- Excel
- MATLAB
- Octave
- SAS
- SPSS
- S-PLUS
- Statistica
- R

R-packages

- A language and environment for statistical computing and graphics.
- Highly compatible to S/ S-plus
- Open source under GNU General Public License
- Runs on many UNIX/ Linux/ windows family and MacOS platform
- There are growing number of microarray analysis software (packages) written in R

R-packages

- Dedicated for microarray analysis
  - affy
  - Bioconductor
  - SMA extension
  - Cyber T
  - GeneSOM
  - Permax
  - OOMAL (S-Plus)
  - SMA
  - YASMA
- General packages
  - cclust
  - cluster
  - mclust
  - multiv
  - mva
  - …etc!

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene Identification
- 6. Phylogenetic Analysis
- 7. Protein Structure Analysis and Prediction
- 8. RNA Secondary Structure Prediction
- **9. Clustering and Classification of Gene Expression Data**
- 10. Search and Mining of Biological Databases, Databanks, and Literature