Microarray - Data Analysis

Introduction to bioinformatics - Bio2304

Hilde Hansen

Topics

Gene expression

Different microarrays

Raw data from microarrays

Data quality

Gene expression matrices

Grouping of expression data

Clustering methods

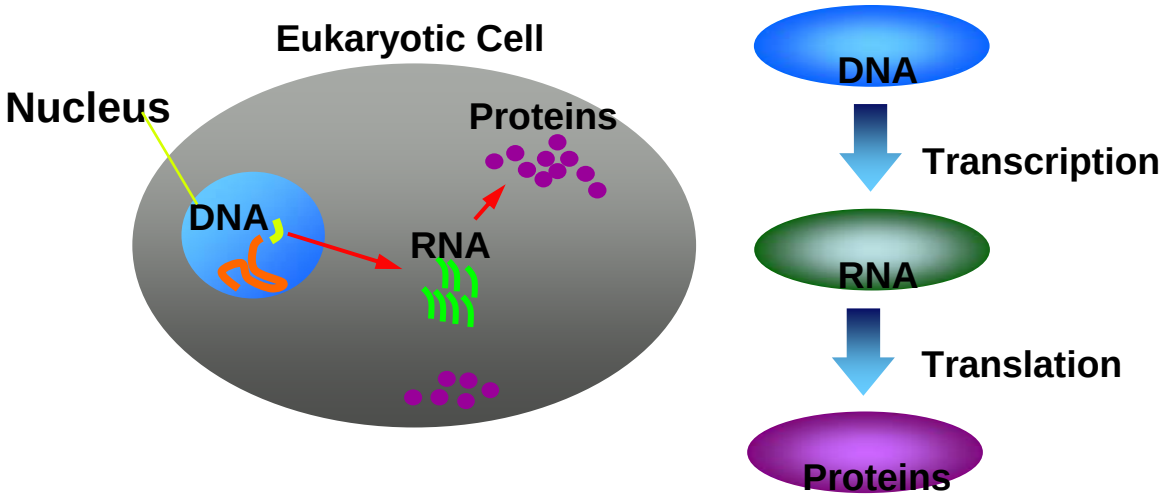
Feature reduction

Data tools

Applications of microarrays

- Study gene expression of cells and microorganisms (basic research)
- Identification of genetic diseases
- Differentiation
- Responses to environmental factors
- Drug management

Gene expression



- Only ~1.5% of genome (genetic material) encodes genes
- 20,000 to 30,000 genes
- Physiological / pathophysiological responses are linked to changes in gene expression
- Knowledge of gene expression under different conditions can create hypothesis about gene functions and new diagnostic tools.

Genomewide expression analysis

- To measure RNA levels of all genes in the genome (high throughput analysis)

- RNA levels vary with following:
  - Cell type
  - Developmental stage
  - External stimuli

- Measure differential gene expression between samples ie. healthy tissue versus diseased tissue

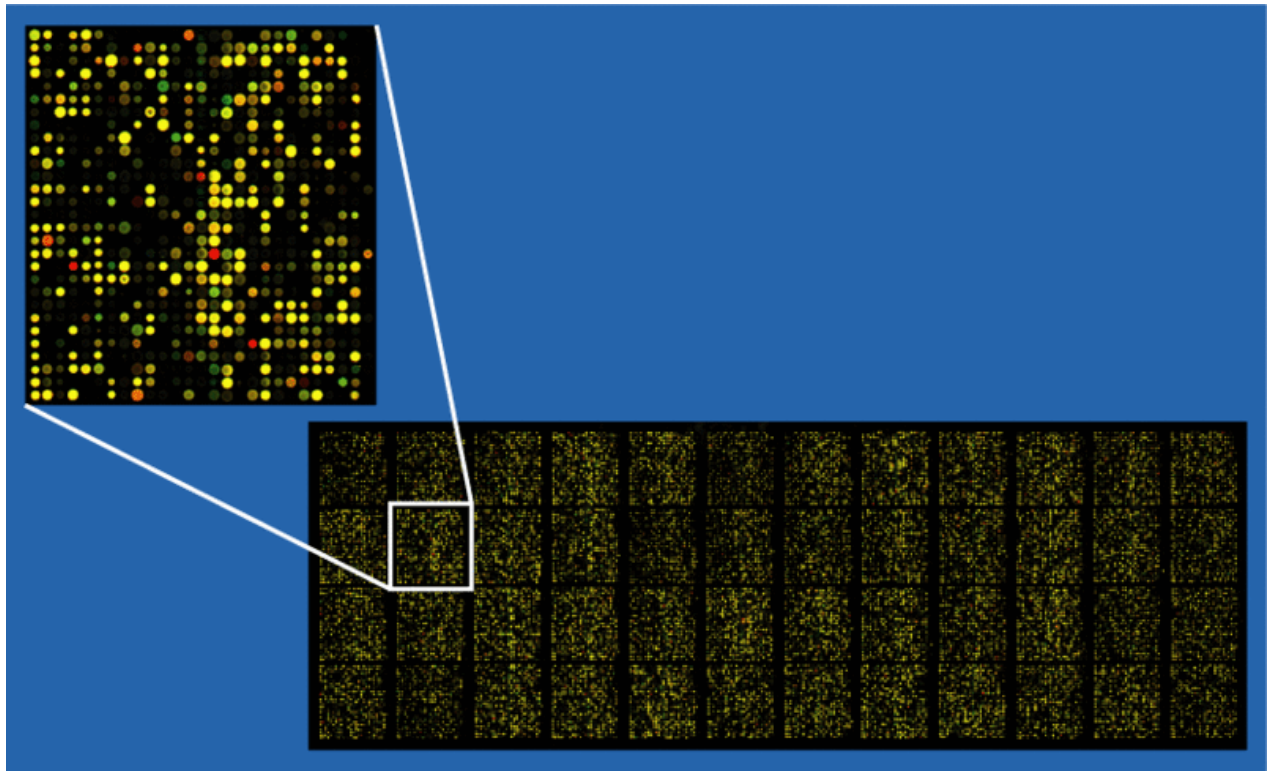
- Differences in cell state or type are correlated with changes in mRNA levels of many genes

- Time and location of expression provide useful information on gene function

What are microarrays?

- Microarrays are miniature devices comprising a large number of sequences (features) immobilized on a slide.
- The sequences/features are arranged as a grid.
- Hybridized with a complex probe (labeled DNA or RNA)
- The intensity of the hybridization signal for each feature corresponds to the amount of that particular molecule in the probe. This is directly proportional to the level of gene expression in the sample.
- Thus, the relative expression levels of thousands genes can be determined at the same time.

Microarray slide



Steps in transcriptional analysis with microarrays

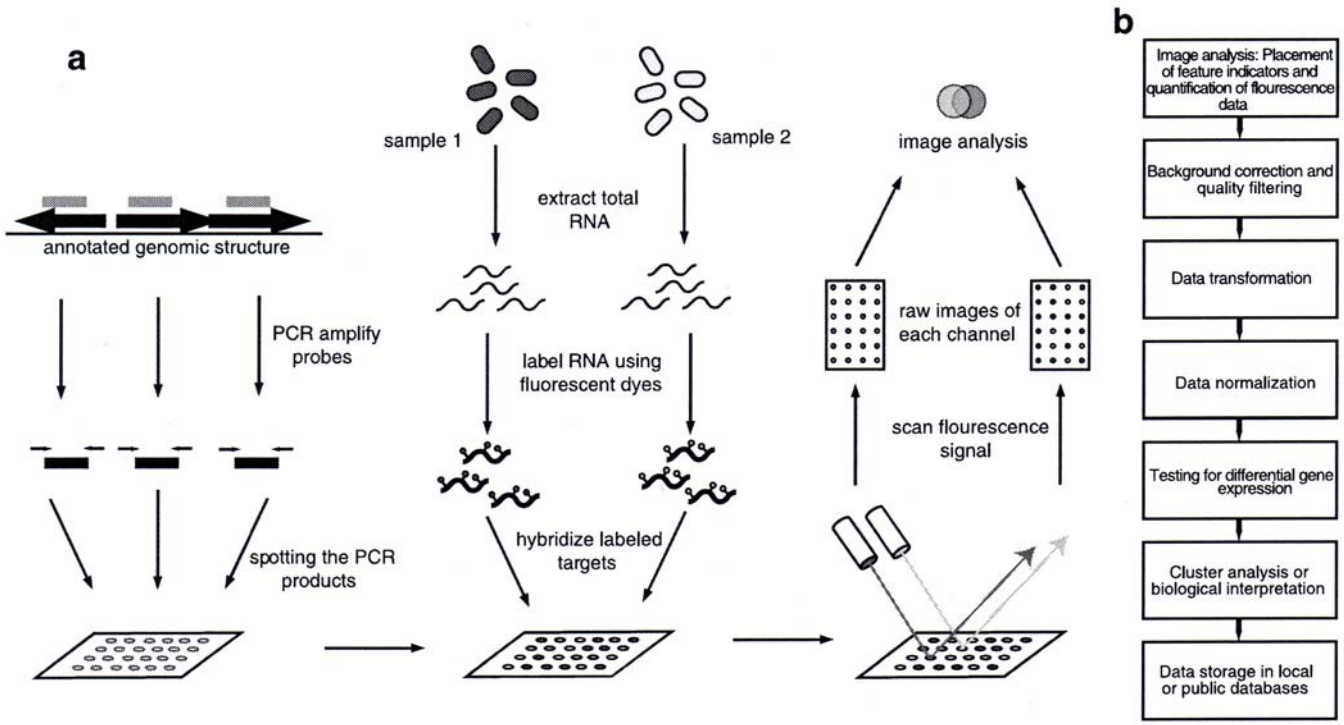


Fig. 1 Main steps in transcription analysis with microarrays. **a** Probes are generated from an annotated genome sequence and spotted on a microarray slide. For target preparation, RNA is extracted from two experimental conditions and labeled with fluorescent dyes by reverse transcription. The labeled target is then hybridized with the array, and the fluorescence of the features is determined using an array scanner. **b** After image analysis, quality filtering, data transformation, and normalization are done. The remaining steps are dependent on the experiment, but in most cases, the data are tested for different gene expression, clustered, and finally stored in a database

Three basic types of arrays

Macroarrays

Robot spotted DNA microarrays

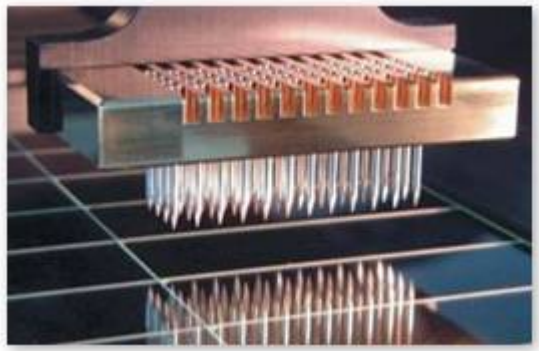
GeneChips

Macroarray

- Spotting DNA molecules onto nylon paper (macroarray).
- Hybridization with radioactivity or colorimetric labeled probes.
- Resolution of signals are low, and the features on the array cannot be tightly packed.
- First generation

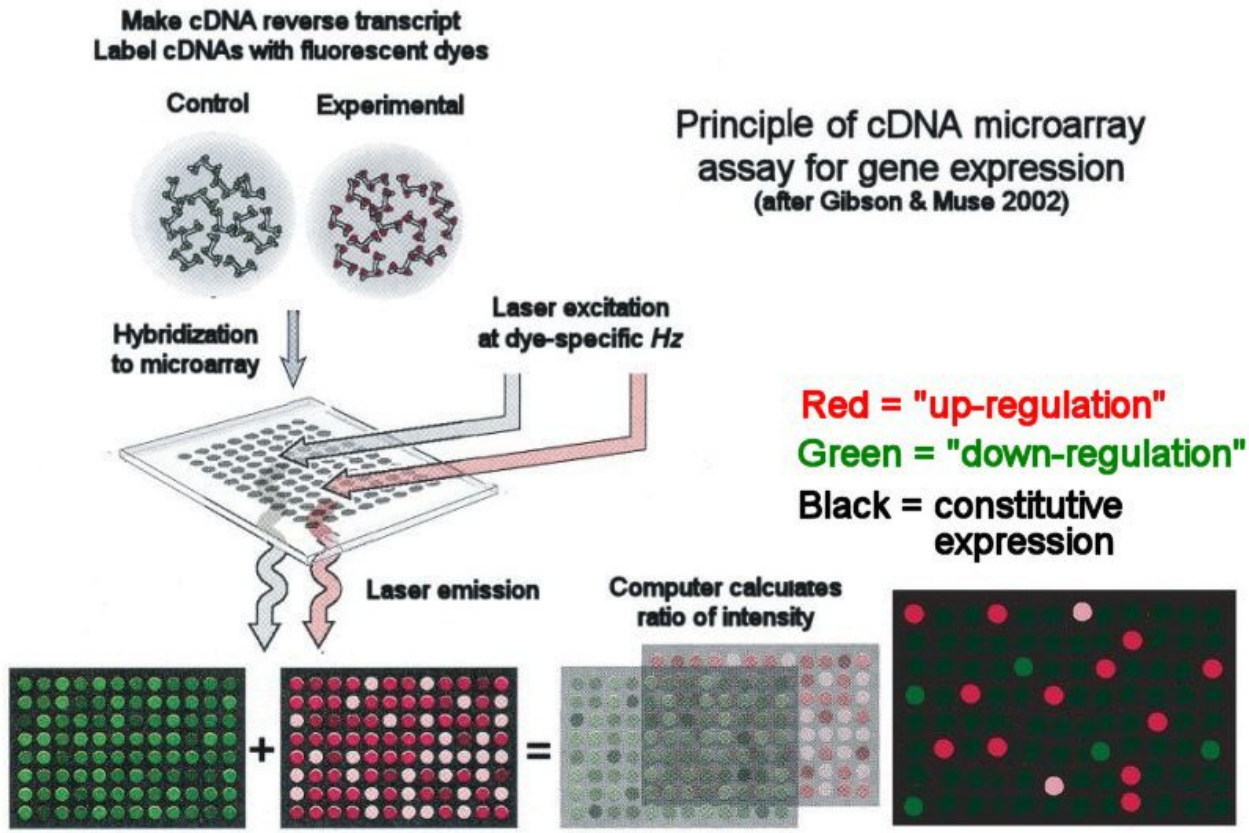
Robot spotted DNA microarrays

- Robot spotted DNA microarrays with up to 5000 features per square cm. The features comprise dsDNA up to 400 bp and must be denatures prior to hybridization.
- Oligos (50-70 mers for consistent TMs) can also be spotted
- Successful arraying include highly accurate array printing (reproducibility)



- Dual labeled fluorescent probe/sample. Different fluorophores (ex. Cy3 / Cy5) can be used to label the different RNA populations, and hybridized simultanously to the same array allowing differential expression to be detected directly.

Dual label

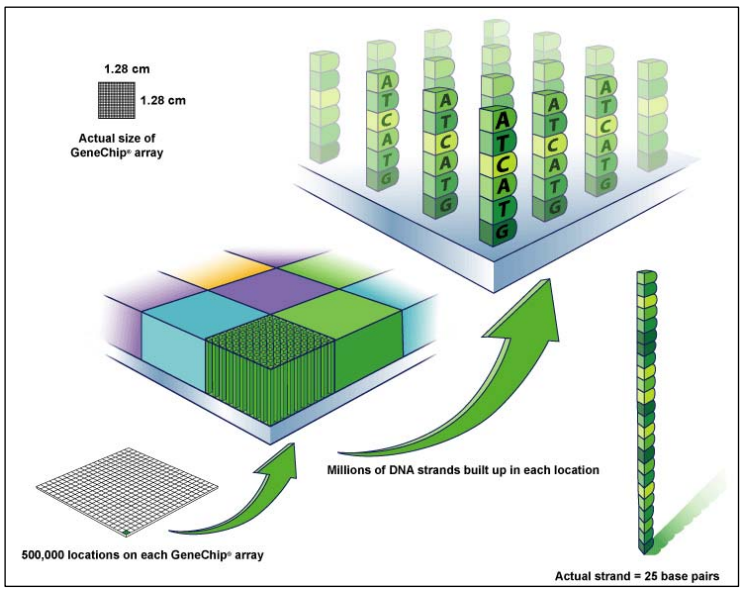


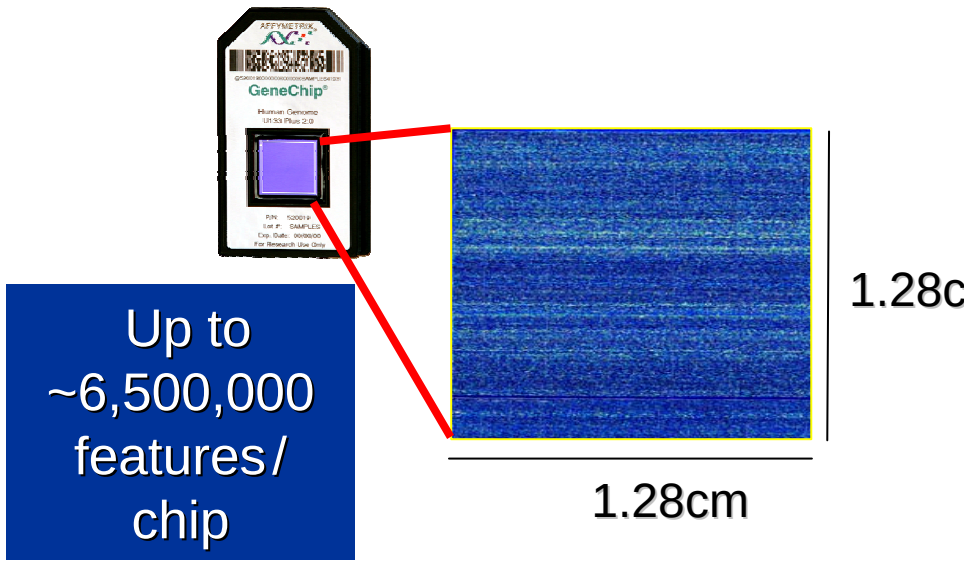
Fluorescent signal are acquired by confocal laser scanning of the hybridized array at the appropriate excitation wavelength and recording at the appropriate emission wavelength (or channel)

GeneChips

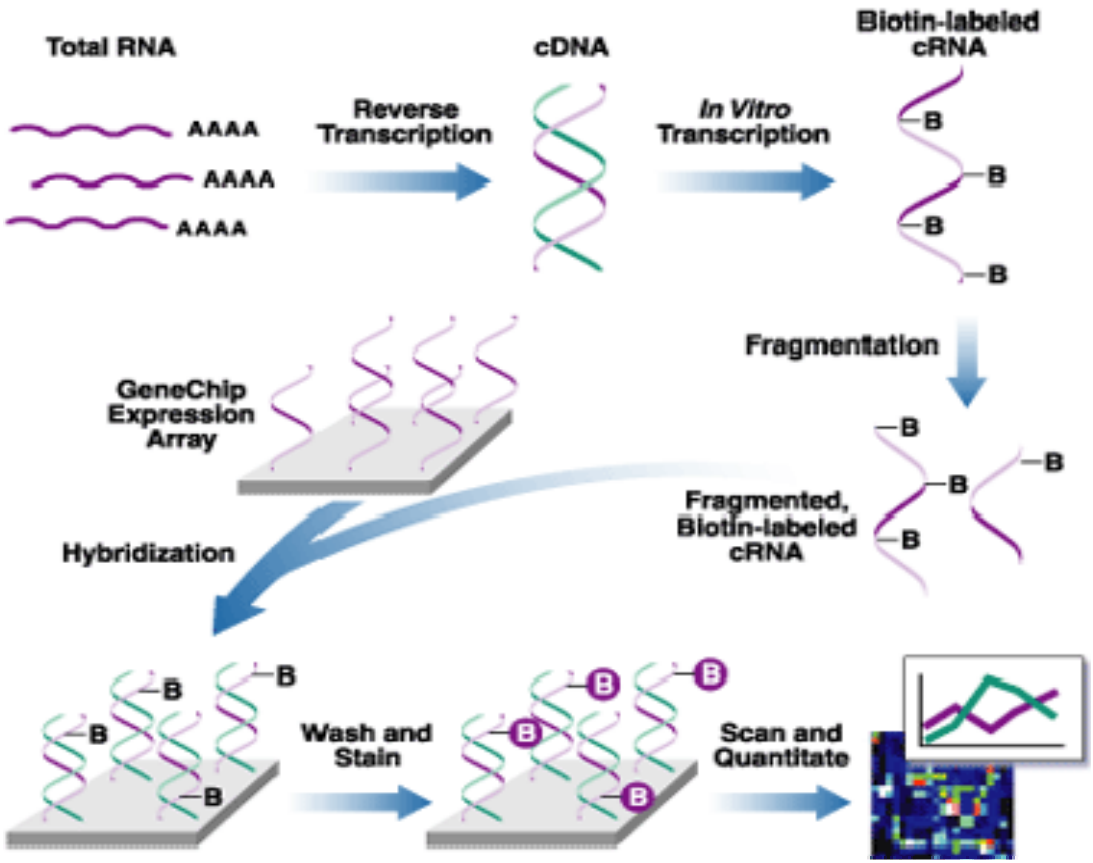
- GeneChips made by “on-chip photolithographic syntesis”, in which short oligonucleotides (25 nt) are printed onto the chips using solid state chemical synthesis during chip manufacturing (Affymetrix Inc). Each gene is represented by 20 non-overlapping oligos.

- Hybridization is carried out with separate probes on two identical chips and the signal intensities are measured and compared by accompanying analysis software.

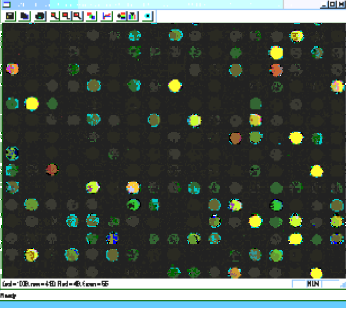




GeneChip



Affymetrix



Data quality

- Arrays contain thousands of features which require automated data acquisition and analysis
- The scanner determines the boundaries of individual spots and measure the total signal intensity over the whole spot
- Signal intensities can vary across individual spots and the shape and size of different spots may not be uniform.
- It is therefore important to locate the spots precisely and it may be necessary to align the grid manually
- Background problems due to non-specific hybridization, autofluorescence, dust and other contaminants or poor hybridization techniques.

Data quality

- Background may vary over the array surface so that signal intensities must be normalized for local background values. This can be difficult when the signal intensity is low for a particular spot.

- Control features are needed to measure non-specific hybridization and variable hybridization from array to array (variation in array manufacture or experimental protocol when using duplicate arrays are used to study differential gene expression).

- Experimental design

- Sample preparation before hybridization and datacollection
  - good sample preparation
  - high quality RNA

Data quality and processing

- Replicates
  - The number of replicates increases the quality of microarray data
  - Two types of replicates
    - In-Array replicates are replicate spots on the same array
    - Between array replicates are replicated microarrays (same biological material)

- Filtering
  - Remove unwanted spots
    - Spots below a given diameter
    - Spots with foreground below 2x background
    - Empty spots and controls

- Normalization
  - Remove unwanted dye bias
  - Remove the systematic effects, and bring the data from different microarrays onto a common scale

Gene expression matrices

- Microarrays generate large datasets.
- Hybridization signals/raw data must be converted into numbers which can be used to build a gene expression matrix
- Rows represent genes and columns represent conditions (time, treatment)

Example

3 different conditions (arrays)

3 different genes

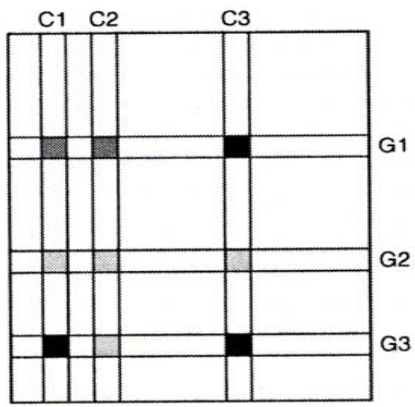
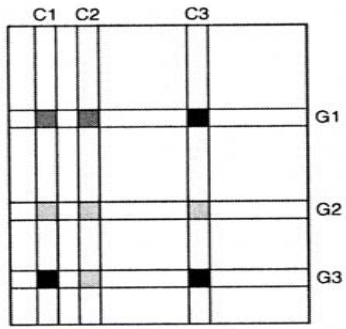


Fig. 1. Schematic of an idealized expression array, in which the results from three experiments are combined. Three genes (G1, G2 and G3) are labeled on the vertical axis and three experimental conditions (C1, C2 and C3) are labeled on the horizontal axis, giving a total of nine data points. The shading of each data point represents the level of gene expression, with darker colors representing higher expression levels.

Grouping of expression data

- Grouping expression data according to similar expression profiles.

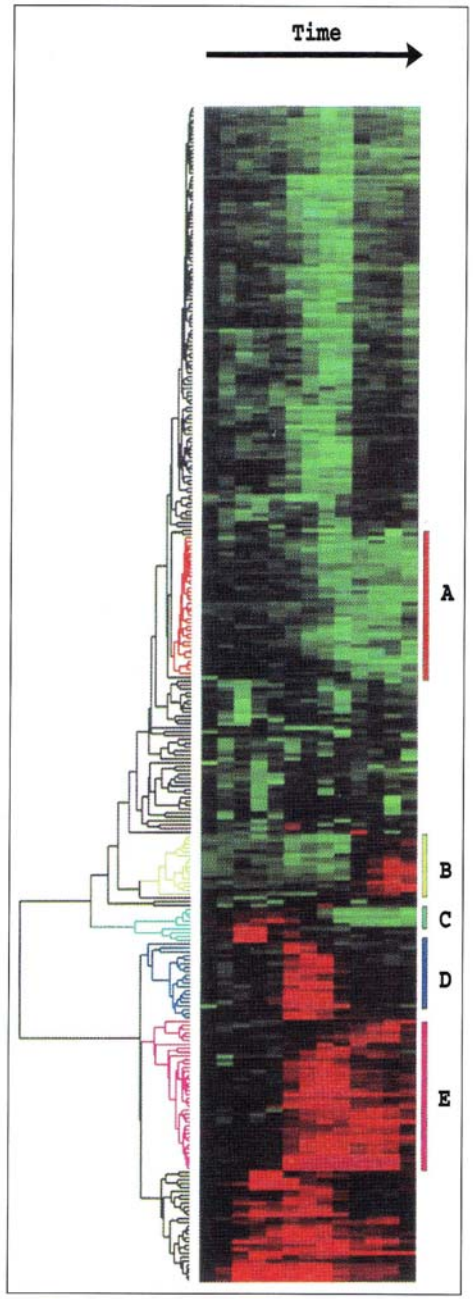
- An expression profile can be defined as the expression measurements of a given gene over a set of conditions (reading along a row of data in the gene expression matrix)



- Supervised methods — ex. different cancer diagnosis (classification)
- Unsupervised methods – no inbuilt classifiers (clustering)

Clustering analysis of data from multiple microarray expression experiments can identify coregulated genes

- Each row represents a different gene at times after addition of serum
- Expression of 8600 mammalian genes was detected by microarray analysis over a time periode of 24 hours
- The computer algorithm groups genes that show similar changes in expression compared to a serum starved control sample
- Red indicates an increase in expression compared to the control
- Green: decreased expresssion
- Black: no significant changes in expression



Clustering methods

- Large datasets: thousands of genes and 50-100 samples
- Simplify the datasets by grouping similar data into clusters
- Clustering generates groups or hierarchies based on similarity. Expression similarity is often caused by co-regulation or other interesting biological processes.
- Clustering requires measurements of ***the distance*** between the genes in terms of their expression profiles. Thus, the gene expression matrix must be converted into a distance matrix so that similarities and differences between data points can be determined.

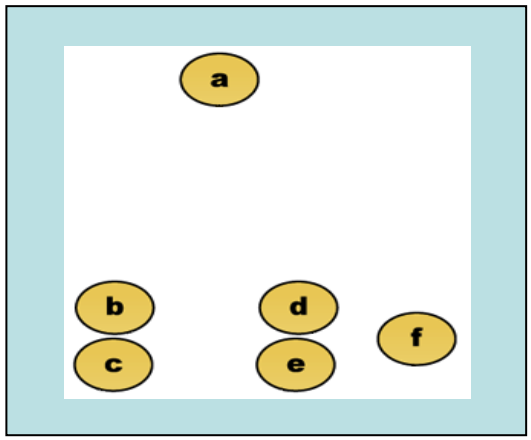
Euclidean distance

$$D_{Euclid} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Is the distance is the square root of the sum of the squared differences between any two data points ((x1,y1) and (x2,y2))

- Can be misleading when two genes have similar expression profiles but different amplitudes, or if two genes have different but very low expression profiles

Euclidean distance matrix



	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

Pearson linear correlation-based distance

$$R(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}, \qquad (2)$$

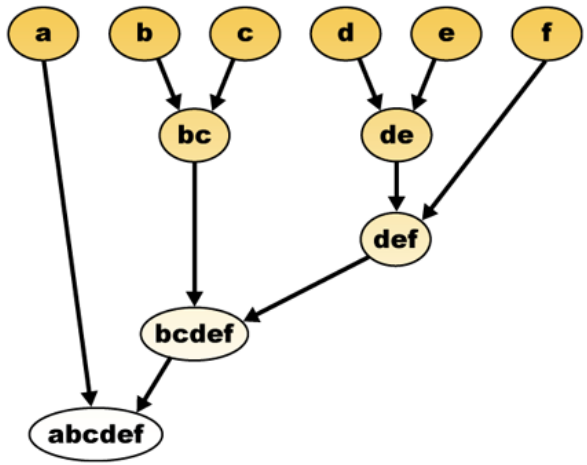
- The Pearson correlation reflects the degree of linear relationship between two patterns.
- It ranges between -1 to +1, reflecting respectively a perfect negative (positive) linear relationship between the patterns.
- A zero correlation value implies that there is no linear relationship between the two patterns, yet it gives no indication regarding nonlinear relationships that might exist between the patterns.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1858704

Dendrograms

- The distance matrix is used to make a dendrogram showing the clustring of expression profiles.

- Hierarchical clustering dendogram using Euclidean distance



$$D_{Euclid} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Dendrograms

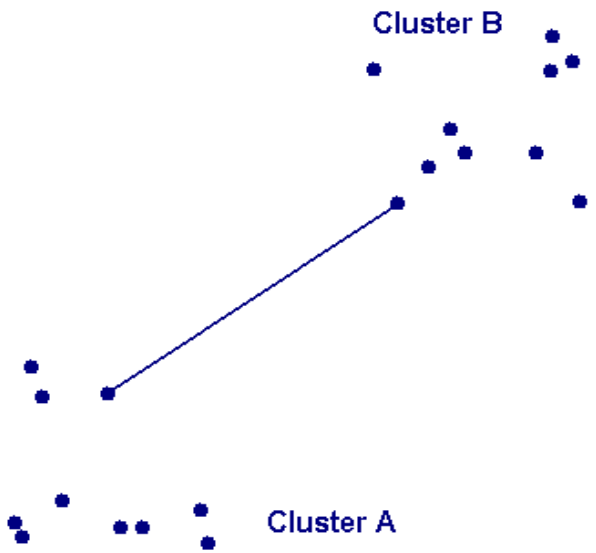
- Essentially the same process that is used to build phylogenetic trees, but the data sets are larger and a rooted tree is not necessary (no ancestor)

- Hierarchial clustring most widely used in which each gene defines its own singelton cluster and the algorithm searches for the two most similar clusters. These neighbors are then merged into a single cluster and the process is repeated.

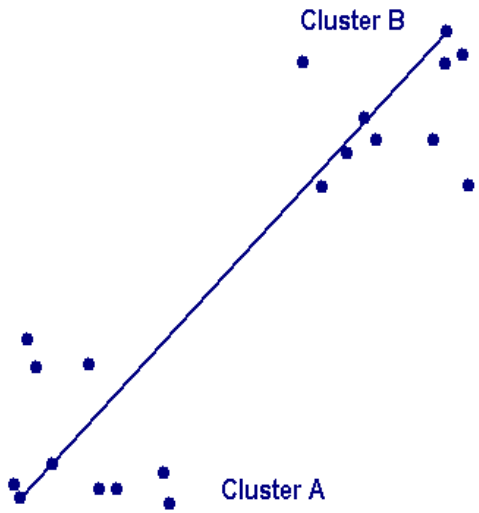
- Single- (nearest neighbors), complete- (farthest neighbors) or average linking generates dendrograms with different topologies

Hierarchical clustering methods

- **Single linkage** or **nearest neighbor**. The distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group.



- The **complete linkage**, or **farthest neighbor**, clustering method. Distance between groups is defined as the distance between the most distant pair of objects, one from each group



http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

Some examples…

- Red experimental mRNA (yeast grown without O_2)
- Green control mRNA (yeast grown with O_2)
- The colored spots are converted into numbers that represent the light intensities of the red dye, green dye and the ratios of red to green
- Ratios below 1 indicate repression of a gene
- Ratios above 1 indicate induction of a gene

Table 6.1 Summary table of data for one yeast DNA microarray.
The fluorescence intensity for each color was determined, and the ratio of red signal to green signal was calculated. This table shows data for only 14 of the 6,200 genes on the full microarray. Also contained in the data is the location for each spot. When a microarray is fabricated, it is important to know where each gene is located on it.

Block	Column	Row	Gene Name	Red	Green	Red : Green Ratio
1	1	1	*Tub1*	2,345	2,467	0.95
1	1	2	*Tub2*	3,589	2,158	1.66
1	1	3	*Sec1*	4,109	1,469	2.80
1	1	4	*Sec2*	1,500	3,589	0.42
1	1	5	*Sec3*	1,246	1,258	0.99
1	1	6	*Act1*	1,937	2,104	0.92
1	1	7	*Act2*	2,561	1,562	1.64
1	1	8	*Fus1*	2,962	3,012	0.98
1	1	9	*Idp2*	3,585	1,209	2.97
1	1	10	*Idp1*	2,796	1,005	2.78
1	1	11	*Idh1*	2,170	4,245	0.51
1	1	12	*Idh2*	1,896	2,996	0.63
1	1	13	*Erd1*	1,023	3,354	0.31
1	1	14	*Erd2*	1,698	2,896	0.59

Campbell and Heyer (2007). Discovering genomics, proteomics and bioinformatics

400 red LU / 1200 green LU = 0.33

→ 3x repression the gene

1800 red LU / 180 green LU = 10

→ 10x induction of the gene

Table 6.2 Data showing fold change (experimental ÷ control) in mRNA production for 12 hypothetical genes (C – N).

Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene C	1	8	12	16	12	8
gene D	1	3	4	4	3	2
gene E	1	4	8	8	8	8
gene F	1	1	1	0.25	0.25	0.1
gene G	1	2	3	4	3	2
gene H	1	0.5	0.33	0.25	0.33	0.5
gene I	1	4	8	4	1	0.5
gene J	1	2	1	2	1	2
gene K	1	1	1	1	3	3
gene L	1	2	3	4	3	2
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125

Campbell and Heyer (2007). Discovering genomics, proteomics and bioinformatics

400 red LU / 1200 green LU = 0.33

→ 3x repression the gene

1800 red LU / 180 green LU = 10

→ 10x induction of the gene

Table 6.3 Reorganization, or clustering (see Math Minute 6.3) of gene order from Table 6.2 based on similarity of expression patterns or profiles.

Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125
gene H	1	0.5	0.33	0.25	0.33	0.5
gene K	1	1	1	1	3	3
gene J	1	2	1	2	1	2
gene E	1	4	8	8	8	8
gene C	1	8	12	16	12	8
gene L	1	2	3	4	3	2
gene G	1	2	3	4	3	2
gene D	1	3	4	4	3	2
gene I	1	4	8	4	1	0.5
gene F	1	1	1	0.25	0.25	0.1

Campbell and Heyer (2007). Discovering genomics, proteomics and bioinformatics

Why log2 transform microarray data?

You have seen how red-to-green intensity ratios are computed from the intensity levels determined in the scanning process, and how they represent repression or induction of the genes. For example, a four-fold repression in gene expression results in a ratio of approximately 0.25. Similarly, a 16-fold repression results in a ratio of 0.0625, a 4-fold induction 4.0, and a 16-fold induction 16.0. But there is a much greater numerical difference between 4.0 and 16.0 than there is between 0.25 and 0.0625. Graph the expression patterns for genes C and N from Table 6.1 on a single set of axes, and notice how the induction stands out much more clearly than the repression, even though the repression is of the same magnitude. In addition to biasing our visual interpretation of gene expression patterns, the compression of ratios between 0 and 1 causes problems with mathematical techniques for analyzing and comparing gene expression patterns (Math Minute 6.2). Furthermore, to interpret a ratio less than 1 as a fold repression, you must take its reciprocal (e.g., $1/0.0625 = 16.0$), an operation that most people find difficult to do quickly and accurately in their heads.

To avoid these problems, investigators often use a **log transformation** of the ratio data. Log transformation is illustrated in the following example. Suppose the ratio is 0.0625. Take the base 2 logarithm of this number:

$$\log_2(0.0625) = \log_2(1/16) = \log_2(1) - \log_2(16) = -\log_2(16) = -4.$$

Because the $\log_2$ of 1/16 is the negative of the $\log_2$ of 16, a 16-fold induction and a 16-fold repression have the same magnitude (one positive and one negative) in the $\log_2$-transformed data. The magnitude of the number is the power of 2 needed to get the induction/repression number (e.g., $16 = 2^4$). The $\log_2$ transformations of the ratios in Table 6.1 are given in Table MM6.1.

Sometimes $\log_{10}$ is used instead of $\log_2$; in this case, the magnitude of the transformed data is the power of 10 needed to get the induction/repression number. Transforming the same ratios as above leads to the following:

$$\log_{10}(4) \approx 0.6 \qquad \log_{10}(.25) \approx -0.6,$$
$$\log_{10}(16) \approx 1.2 \qquad \log_{10}(0.0625) \approx -1.2.$$

Table MM6.1 Log_2 transformation of gene expression data in Table 6.1.

Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene C	0	3	3.58	4	3.58	3
gene D	0	1.58	2	2	1.58	1
gene E	0	2	3	3	3	3
gene F	0	0	0	−2	−2	−3.32
gene G	0	1	1.58	2	1.58	1
gene H	0	−1	−1.60	−2	−1.60	−1
gene I	0	2	3	2	0	−1
gene J	0	1	0	1	0	1
gene K	0	0	0	0	1.58	1.58
gene L	0	1	1.58	2	1.58	1
gene M	0	−1.60	−2	−2	−1.60	−1
gene N	0	−3	−3.59	−4	−3.59	−3

Campbell and Heyer (2007). Discovering genomics, proteomics and bioinformatics

Pearson correlation-based distance matrix

Table MM6.2 Correlation coefficient between each pair of genes, based on $\log_2$-transformed gene expression data in Table MM6.1.

	gene C	gene D	gene E	gene F	gene G	gene H	gene I	gene J	gene K	gene L	gene M	gene N
gene C	1	0.94	0.96	−0.40	0.95	−0.95	0.41	0.36	0.23	0.95	−0.94	−1
gene D	0.94	1	0.84	−0.10	0.94	−0.94	0.68	0.24	−0.07	0.94	−1	−0.94
gene E	0.96	0.84	1	−0.57	0.89	−0.89	0.21	0.30	0.43	0.89	−0.84	−0.96
gene F	−0.40	−0.10	−0.57	1	−0.35	0.35	0.60	−0.43	−0.79	−0.35	0.10	0.40
gene G	0.95	0.94	0.89	−0.35	1	−1	0.48	0.22	0.11	1	−0.94	−0.95
gene H	−0.95	−0.94	−0.89	0.35	−1	1	−0.48	−0.21	−0.11	−1	0.94	0.95
gene I	0.41	0.68	0.21	0.60	0.48	−0.48	1	0	−0.75	0.48	−0.68	−0.41
gene J	0.36	0.24	0.30	−0.43	0.22	−0.21	0	1	0	0.22	−0.24	−0.36
gene K	0.23	−0.07	0.43	−0.79	0.11	−0.11	−0.75	0	1	0.11	0.07	−0.23
gene L	0.95	0.94	0.89	−0.35	1	−1	0.48	0.22	0.11	1	−0.94	−0.95
gene M	−0.94	−1	−0.84	0.10	−0.94	0.94	−0.68	−0.24	0.07	−0.94	1	0.94
gene N	−1	−0.94	−0.96	0.40	−0.95	0.95	−0.41	−0.36	−0.23	−0.95	0.94	1

Table MM6.3 Correlations between [EC] and all other objects.

D	F	H	I	J	K	M	N	[LG]
0.90	−0.48	−0.93	0.32	0.33	0.32	−0.90	−0.99	0.93

Campbell and Heyer (2007). Discovering genomics, proteomics and bioinformatics

Table MM6.4 Summary of the hierarchical clustering algorithm applied to the 12 genes in Table 6.2.

	Two Most Similar Objects			
Iteration	Object 1	Object 2	Correlation	New Object
1	L	G	1.00	[LG]
2	E	C	0.96	[EC]
3	N	H	0.95	[NH]
4	M	[NH]	0.95	[MNH]
5	[LG]	D	0.94	[LGD]
6	[EC]	[LGD]	0.94	[ECLGD]
7	I	F	0.60	[IF]
8	J	[ECLGD]	0.29	[JECLGD]
9	K	[JECLGD]	0.19	[KJECLGD]
10	[KJECLGD]	[IF]	−0.12	[KJECLGDIF]
11	[MNH]	[KJECLGDIF]	−0.96	[MNHKJECLGDIF]

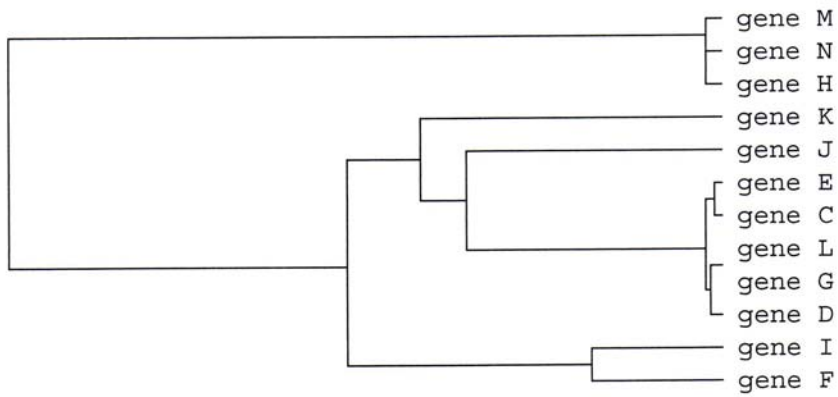


Figure MM6.1 Dendrogram of clustered genes from Table MM6.3 and Figure 6.8.

Campbell and Heyer (2007). Discovering genomics, proteomics and bioinformatics

Feature reduction

- Microarray datasets are large, making classification and clustering laborious.
- Remove noninformative data points
  - When the expression a gene is the same over a range of conditions it does not give useful information when analyzing differential gene expression
- Remove redundant data points
  - When two conditions have the same effect one of the columns of the matrix can be removed
- Result: The algorithms run faster

Microarray data format

- There is no international standard convention yet
- Reflects different platforms (nylon macroarrays, spotted glass microarrays, oligonucleotide chips) and large variation in experimental design, hybridization protocols and data gathering techniques

MIAME - Minimum Information About a Microarray Experiment

- Is a standard for reporting microarray experiments.
- Intends to specify the information necessary to interpret the results unambiguously and to reproduce the experiment
- Documentation / data sharing

- Experimental design
- Array design (identification of each spot on the array)
- Probe source and labeling methods
- Hybridization procedures and parameters
- Measurements procedures (includes normalization methods)
- Controls , values and specification

- The standard defines the content required for compliant reports, but does not specify the format in which this data should be presented

Tools

Microarray databases

National Center for Biotechnology Information (NCBI)

http://www.ncbi.nlm.nih.gov/geo/

GEO (Gene expression Omnibus)

Gene expression and array database

Web-based microarray data analysis

Expression profiler

http://ep.ebi.ac.uk/EP/

Programs for analysis and clustering of expression data

Summary

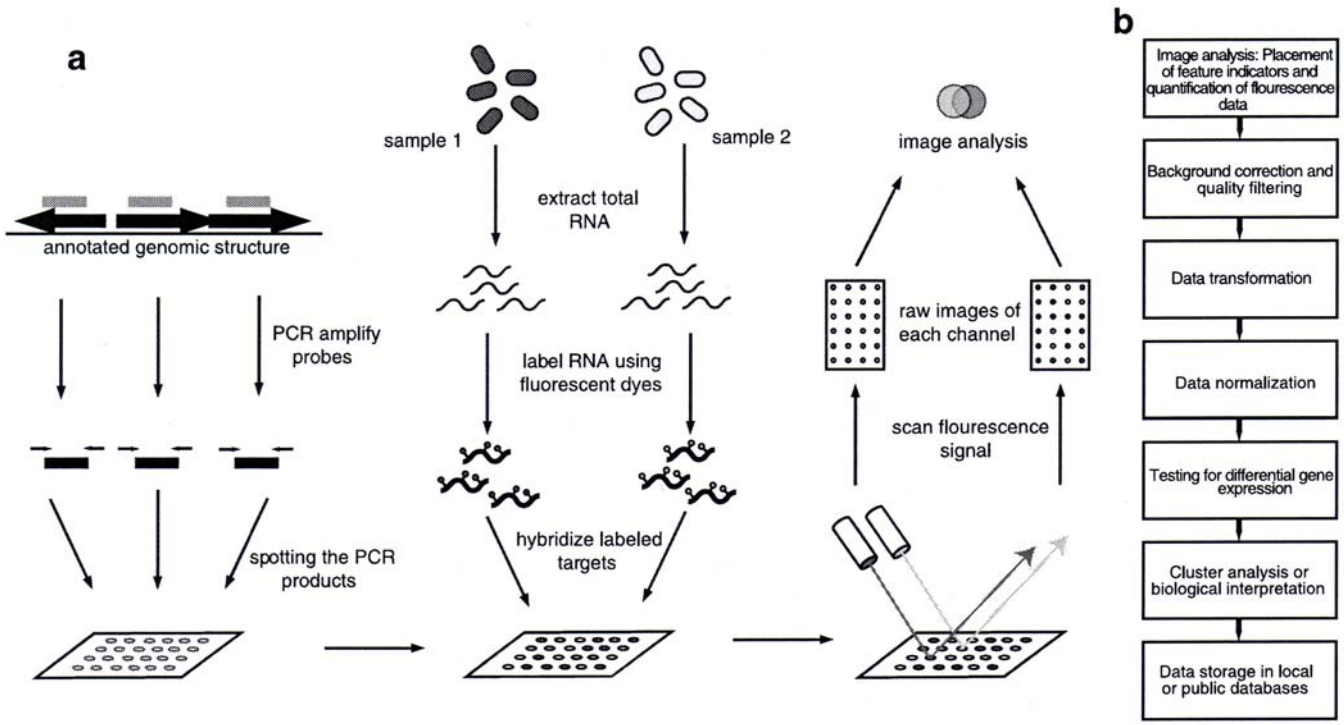


Fig. 1 Main steps in transcription analysis with microarrays. **a** Probes are generated from an annotated genome sequence and spotted on a microarray slide. For target preparation, RNA is extracted from two experimental conditions and labeled with fluorescent dyes by reverse transcription. The labeled target is then hybridized with the array, and the fluorescence of the features is determined using an array scanner. **b** After image analysis, quality filtering, data transformation, and normalization are done. The remaining steps are dependent on the experiment, but in most cases, the data are tested for different gene expression, clustered, and finally stored in a database