

Introduction to MicroArrays and Gene Expression Profiling

Haibe-Kains B^{1,2} Bontempi G² Sotiriou C¹

¹Unité Microarray, Institut Jules Bordet

²Machine Learning Group, Université Libre de Bruxelles

March 23, 2007



Research Groups

Functional Genomic Unit

- 7 researchers (1 prof, 3 postDocs, 6 PhD students), 3 technicians).
- Research topics : Genomic analyses and clinic studies.
- Website : <http://www.bordet.be/en/services/medical/array/practical.htm>.
- National scientific collaborations : ULB, Erasme, ULg, Gembloux
- International scientific collaborations : Genome Institute of Singapore, John Radcliffe Hospital, Karolinska Institute and Hospital, MD Anderson Cancer Center, Netherlands Cancer Institute, Swiss Institute for Experimental Cancer Research, NCI/NIH, Gustave-Roussy Institute, IDDI

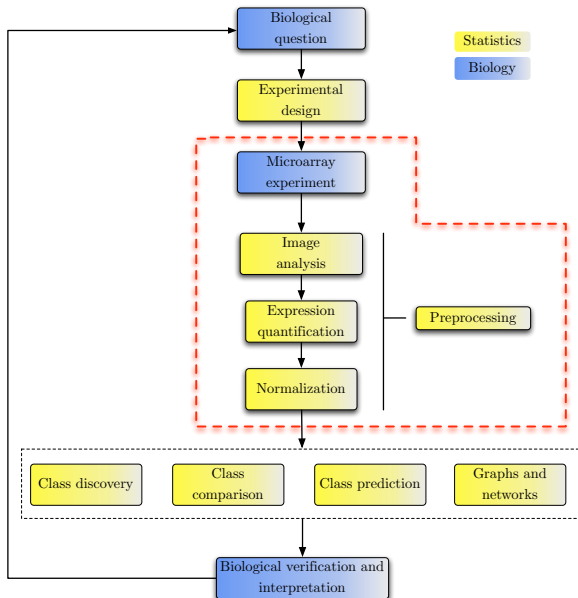
Research Groups

Machine Learning Group

- 7 researchers (1 prof, 6 PhD students), 4 graduate students).
- Research topics : Bioinformatics, Classification, Regression, Time series prediction, Sensor networks.
- Website : <http://www.ulb.ac.be/di/mlg>.
- Scientific collaborations in ULB : IRIDIA (Sciences Appliquées), Physiologie Molculaire de la Cellule (IBMM), Conformation des Macromolcules Biologiques et Bioinformatique (IBMM), CENOLI (Sciences), Microarray Unit (Hopital Jules Bordet), Service d'Anesthesie (Erasme).
- Scientific collaborations outside ULB : UCL Machine Learning Group (B), Politecnico di Milano (I), Università del Sannio (I), George Mason University (US).
- The MLG is part to the "Groupe de Contact FNRS" on Machine Learning and to CINBIOS (<http://babylone.ulb.ac.be/Joomla/>)

- 1 MicroArray Analysis Design
- 2 MicroArray Technologies
 - 1 Introduction
 - 2 Differences
 - 3 Comparison
 - 4 Affymetrix
- 3 Preprocessing
 - 1 Visualization
 - 2 Missing Data
- 4 Biological Annotations
 - 1 NCBI Entrez
- 5 MicroArray Databases

Microarray Analysis Design



Part I

MicroArray Technologies

- Enormous advances in genomics and molecular biology during last decade
- High-throughput techniques like microarrays carry the promise to give new insights into functionalities of whole genomes in a systematic manner
- Moreover, the analysis of such a vast amount of data leads to new challenges in statistics and bioinformatics

MicroArray

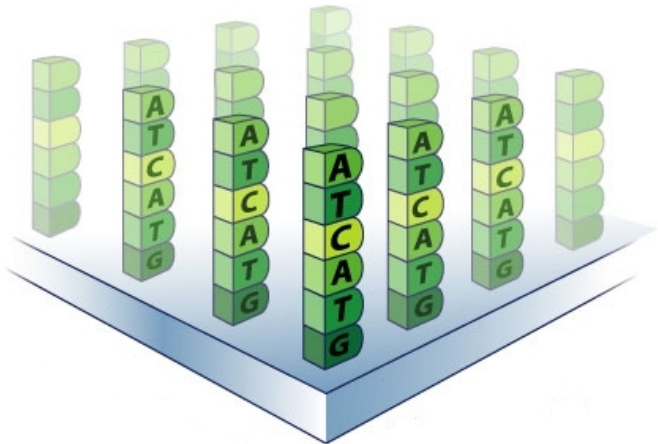
Definition

A *microarray* is composed of

- DNA fragments (*probes*) fixed on a solid support
- ordered position of probes
- principle of hybridization to a specific probe of complementary sequence
- molecular labeling

→ Simultaneous detection of thousands of sequences in parallel

Probes



There exist several high-throughput methods to simultaneously measure the expression of a large number of genes :

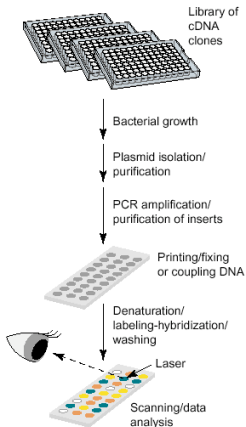
- cDNA microarray
- oligonucleotide microarray
 - ▶ short oligonucleotide (AFFYMETRIX[©])
 - ▶ long oligonucleotide (AGILENT[©], CODELINK[©])
- multiplex quantitative RT-PCR

We could classify microarray technologies by the building process and the gene expression measurement

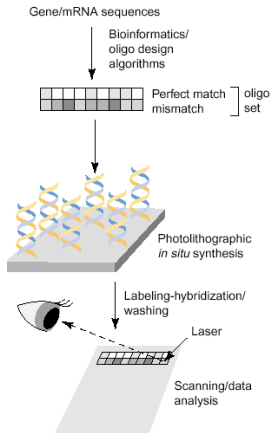
- spotted vs synthesized : some microarrays are built using a robot that puts a small amount of RNA (probes) on a glass. For others, the probes are synthesized directly on the glass using a technology similar to integrated circuits manufacturing.
- dual-channel vs one-channel : some microarrays uses 2 types of labeling to detect hybridization. The first label is for the target tissue and the second one for the reference tissue. For other microarray technologies, only one labeling/tissue is used.

MicroArray Comparison

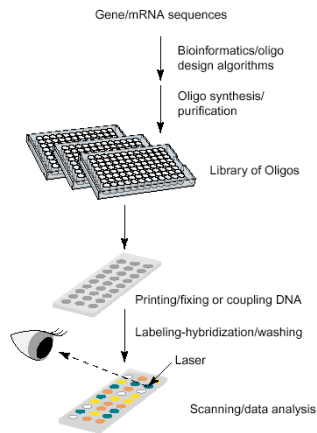
cDNA



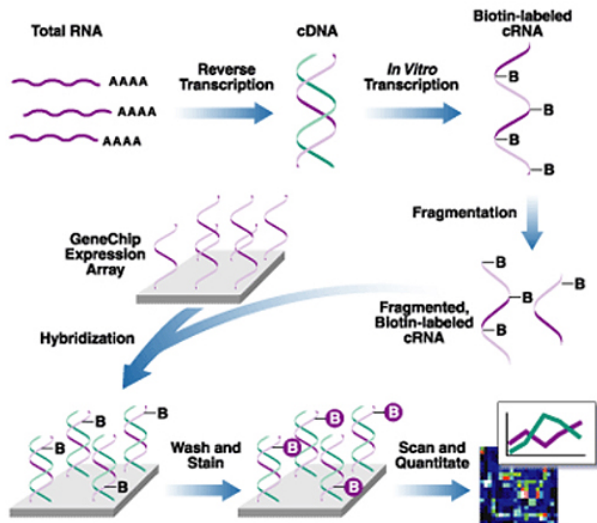
short oligonucleotide

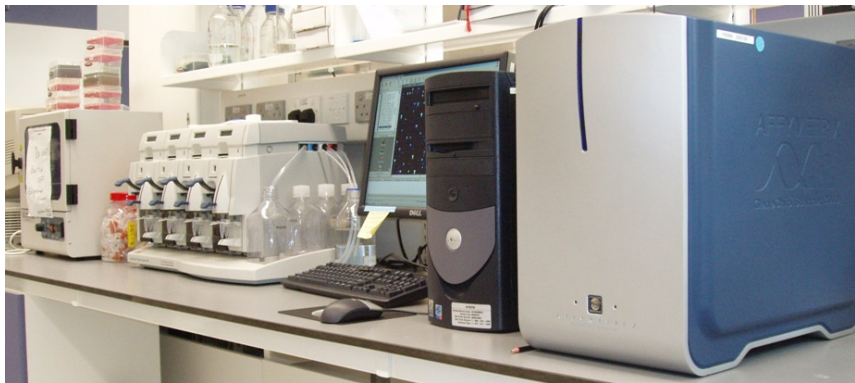


long oligonucleotide



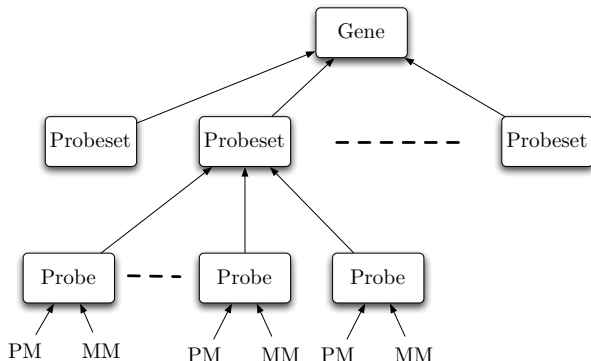
- AFFYMETRIX[©] is a short oligonucleotide (synthesized, one-channel) microarray technology
- We will focus on this platform in order to overview the main characteristics of microarray technology

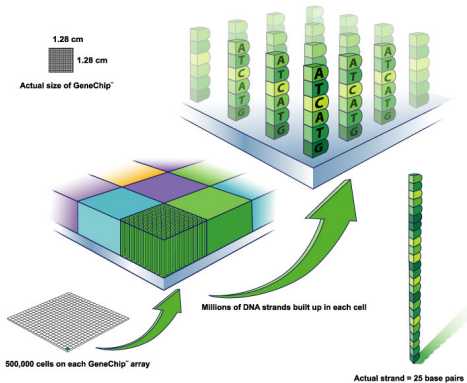




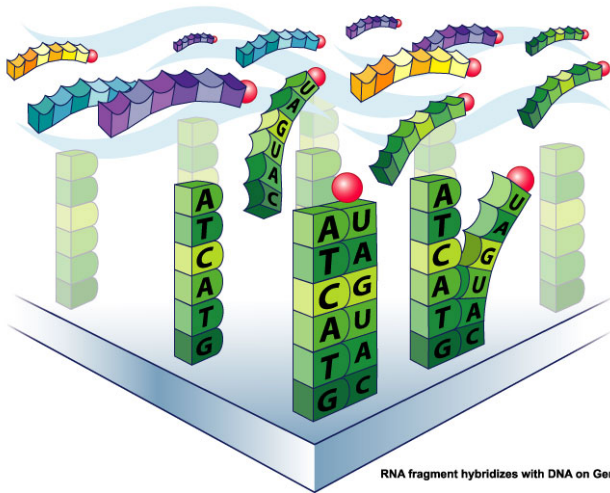


- 1 gene is represented by 1 or more probe sets
- 1 probe set includes 11 to 20 probe pairs
- 1 probe pair includes a perfect match (PM) value and a mismatch value (MM)

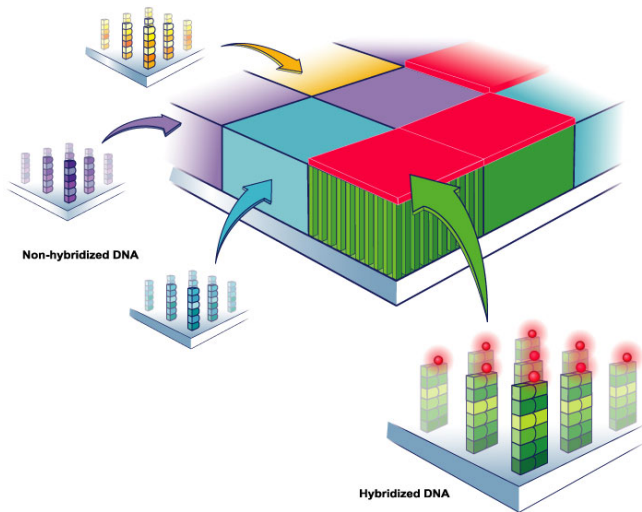




RNA fragments with fluorescent tags from sample to be tested



Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow



- Advantages :

- ▶ Commercially available for several years (strong manufacturing)
- ▶ Large number of published studies (generally accepted method)
- ▶ No reference sample → possible comparison between studies

- Disadvantages :

- ▶ Cost of the devices and the chips (but easy use)
- ▶ Changes in probe design is hard (but new program permits to create his own design)
- ▶ Short oligos → several oligos per gene, specificity/sensitivity trade-off (complex methods to get gene expression)

Part II

Preprocessing

We will focus on **preprocessing** methods for AFFYMETRIX[©] data

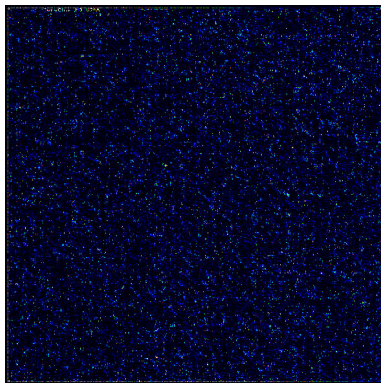
- image analysis : get raw probe intensities from chip image
- expression quantification : get gene expressions from raw probe intensities
- normalization : remove systematic bias to compare gene expressions

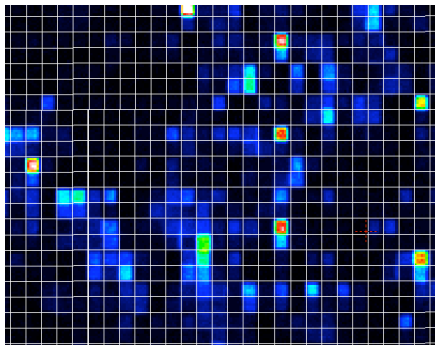
There exist several methods but we will focus on **robust multi-array analysis** (RMA) methods [Irizarry et al., 2003]

AFFYMETRIX[©] Preprocessing

Main software from AFFYMETRIX[©] was MicroArray cont'd 5 (MAS5), now called GeneChip Operating Software 1.4 (GCOS)

- **DAT** file is the image file





Each probe cell is composed by 10x10 pixels

- remove outer 36 pixels → 8x8 pixels
- probe cell signal, PM or MM, is the 75th percentile of the 8x8 pixel values

- **CEL** file is the cell intensity file including probe level PM and MM values

The last file provided by AFFYMETRIX[©] concerns the annotations of the chip

- **CDF** file is the chip description file describing which probes go in which probesets (genes, gene fragments, ESTs)

Most of microarray analyses start from CEL files

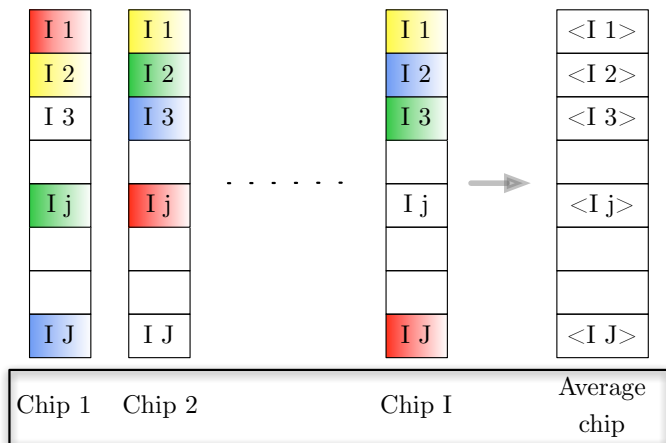
For each probe set, **summarization** of the probe level data (11-20 PM and MM pairs) into a single expression measure

RMA procedure

- use only PM and ignore MM
- adjust for background on the raw intensity scale
- carry out **quantile** normalization [Bolstad et al., 2003] of $PM - \hat{BG}$ and call the result $n(PM - \hat{BG})$
- Take \log_2 of normalized background adjusted PM
- Carry out a **medianpolish** of the quantities $\log_2 n(PM - \hat{BG})$ to summarize the probe level values in one probeset value

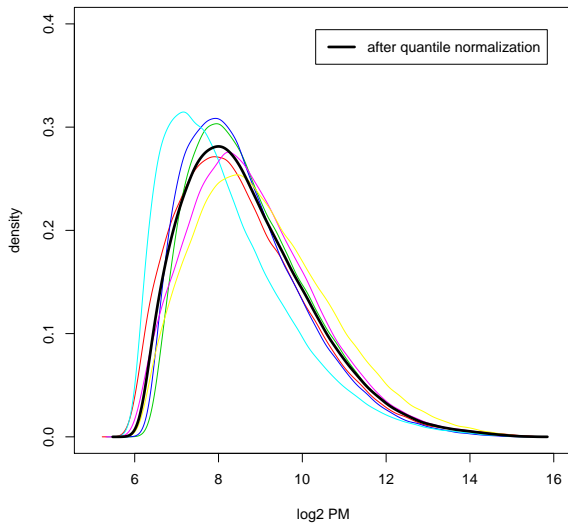
AFFYMETRIX[©] Preprocessing

Quantile Normalization



AFFYMETRIX[©] Preprocessing

Quantile Normalization



Visualizing MicroArray Data

- In order to visualize microarray data, we draw a *heatmap*
- it is an image where the gene expressions are represented by colors
 - ▶ red for high gene expression
 - ▶ green for low gene expression
- Patients are in columns and genes in rows
- Microarray data are usually standardized

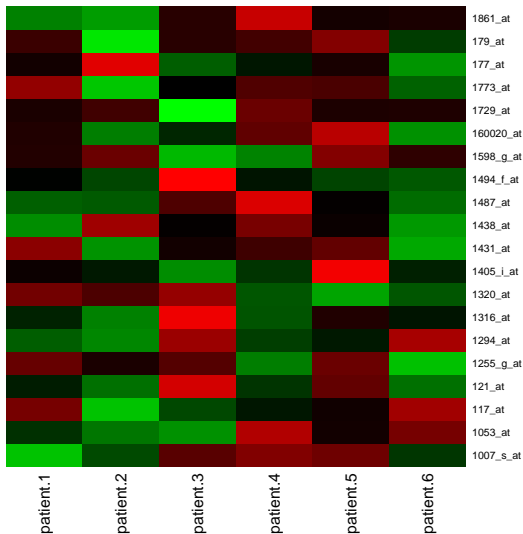
$$\frac{x - center(x)}{scale(x)}$$

may improve the comparison between microarray experiments but may also remove some interesting features in the data.

- We usually apply gene centering in one-channel microarray data, ie $x_{ij} = x_{ij} - center(x_{.j})$ with *center* being the median.

Visualizing MicroArray Data

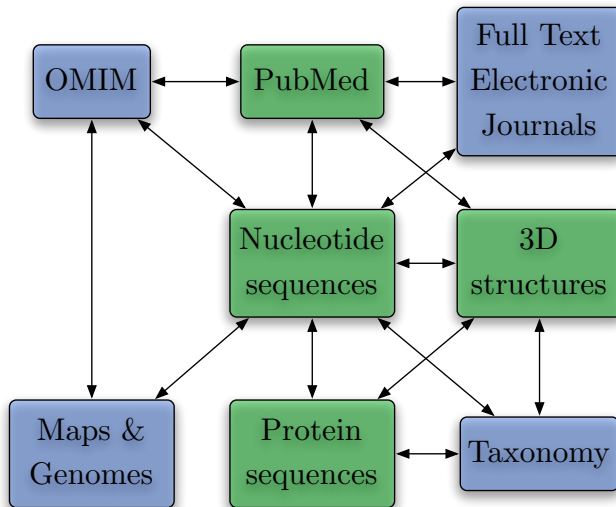
Heatmap



- Depending of the microarray technology and the preprocessing methods, some gene expressions may be missing
- If further analyses require all gene expressions, you can
 - ▶ Replace a missing value by mean or median of row/column
 - ▶ Replace a missing value by a prediction using
 - ★ KNN [Troyanskaya et al., 2001]
 - ★ regression [Kim et al., 2005]
 - ★ bayesian estimation [Oba et al., 2004]
- Such methods have an impact on the overall analysis [Scheel et al., 2005]

Part III

Biological Annotations



Help is available at <http://www.ncbi.nlm.nih.gov/Entrez>

Part IV

MicroArray Databases

Human Cancer Microarray Datasets

Differences between datasets due to :

- Different types of cancer : breast, lymphoma, lung, brain ...
- Different types of microarray technology
- Be careful when comparing different datasets
 - ▶ mapping of the probes through annotations (eg gene ids, unigene clusters)
 - ▶ meta-analysis to consider several datasets in one study as in [Shen et al., 2004, Rhodes et al., 2004, Sotiriou et al., 2006].




- Databases of microarray datasets
 - ▶ gene expression omnibus (GEO) from NCBI : <http://www.ncbi.nlm.nih.gov/geo/>
 - ▶ array express (AE) from EBI : <http://www.ebi.ac.uk/arrayexpress/>
 - ▶ oncomine : <http://www.oncomine.org>

- Databases for mapping
 - ▶ Cleanex from SIB : <http://www.cleanex.isb-sib.ch>
 - ▶ Adapt from Paterson Institute for Cancer Research : <http://bioinformatics.picr.man.ac.uk/adapt>

- Master in Bioinformatics at ULB and other belgian universities :
<http://www.bioinfomaster.ulb.ac.be/>
- Personal homepage : <http://www.ulb.ac.be/di/map/bhaibeka/>
- This presentation : <http://www.ulb.ac.be/di/map/bhaibeka/papers/haibekains2007iddimarray.pdf>

Thank you for your attention.

For Further Reading I

-  Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.
-  Irizarry, R. A., Boldstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. R. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4).
-  Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198.

For Further Reading II



Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S. (2004).

A bayesian missing value estimation method for gene expression profile data.

Bioinformatics, 19(16).



Rhodes, D. R., Yu, J., Shanker, K., Desphande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004).

Latge-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.

PNAS, 101(25):9309–9314.

For Further Reading III



Scheel, I., Aldrin, M., Glad, I. K., Serum, R., Lyng, H., and Frigessi, A. (2005).

The influence of missing value imputation on detection of differentially expressed genes from microarray data.

Bioinformatics, 21(23):4272–4279.





Shen, R., Ghosh, D., and Chinnaiyan, A. M. (2004).

Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data.

BMC Genomics, 5(94).

For Further Reading IV

-  Sotiriou, C., Wirapati, P., Loi, S. M., Desmedt, C., Durbecq, V., Harris, A., Bergh, J., Smeds, J., Haibe-Kains, B., Larsimont, D., Cardoso, F., Buyse, M., Delorenzi, M., and Piccart, M. (2006). Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *Journal of National Cancer Institute*, 98:262–272.
-  Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bolstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.