Review III: DNA Microarrays

Rajan Munshi

BBSI @ Pitt 2006

Department of Computational Biology
University of Pittsburgh School of Medicine

May 25, 2006

Also called

- DNA chips
- biochips
- gene chips
- gene arrays
- genome chips
- genome arrays

What is a microarray?

- An arrangement of DNA sequences on a solid support
- Each microarray contains thousands of genes
- Able to simultaneously monitor gene expression levels in all these genes
- Used for:
    - gene expression studies
    - disease diagnosis
    - pharmacogenetics (drug discovery)
    - toxicogenomics

Types

- Two basic microarray technologies
- cDNA arrays (Stanford)
- High-density oligonucleotide arrays (Affymetrix)
- Each technology has its merits and demerits

Definition

Solid support:
glass slides,
plastic base

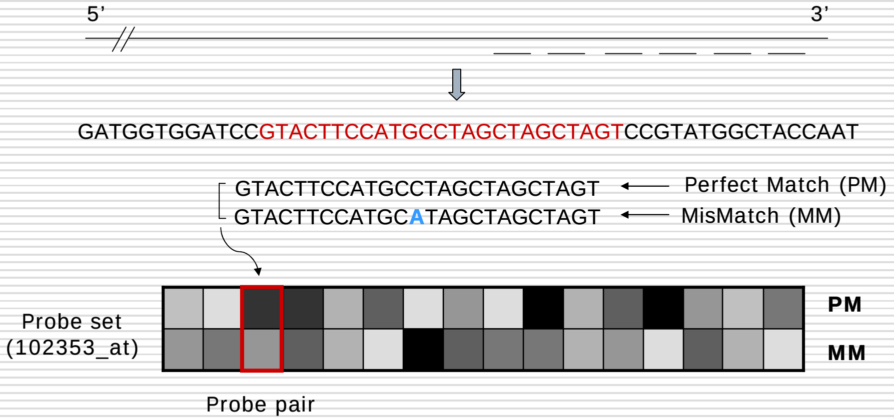
High-density oligonucleotide arrays (1)

- Pioneered by Affymetrix (GeneChip®)
- DNA probe sequences are 25-mer fragments
- Built *in situ* (“on-chip”) by photolithography
- Uses 1 flourescent dye

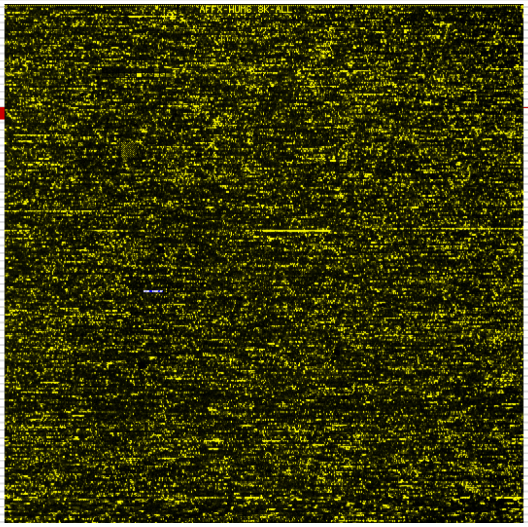
High-density oligonucleotide arrays (2)

- Each sequence is represented by a probe set
- 1 probe set = 16 probe pairs
- Each probe pair = 1 Perfect Match (PM) probe cell and 1 MisMatch (MM) probe cell
- PM = perfectly complementary to target
- MM = central base is mismatched to target

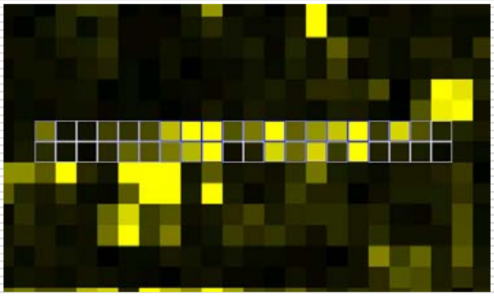
Affymetrix Probe Sets



Affymetrix chip



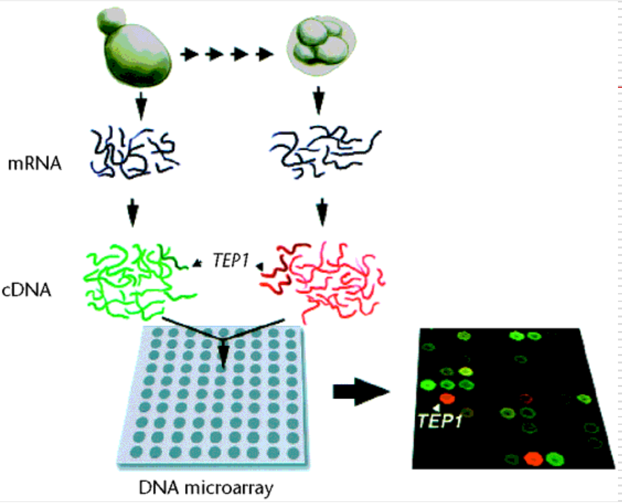
A single probe set



cDNA arrays

- Also known as spotted arrays
- Support can be glass or membrane
- DNA sequences are robotically “imprinted”
- Sequences can range from 30 bp to 2 kb
- Sequences are cDNA clones
- Uses 2 fluorescent dyes (cy3, cy5)

cDNA arrays overview



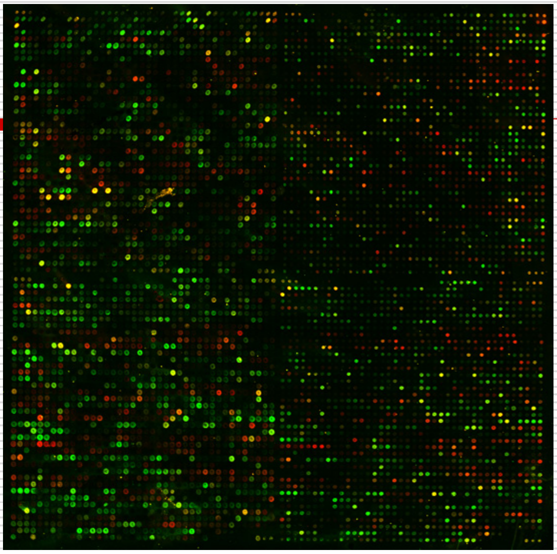
cDNA arrays

Animation

(Courtesy: Dr. A. Malcolm Campbell, Davidson College, NC)

(www.bio.davidson.edu/courses/genomics/chip/chip.html)

Genome-on-a-chip (yeast)



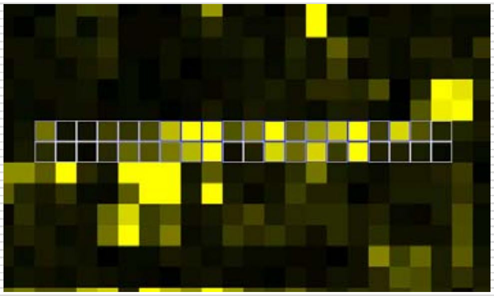
General Steps

Probe	**Chip Fabrication**	**Target**	**Assay**	**Readout**	**Informatics**
DNA or cDNA with known identity	Putting probes on chip (robotic imprinting, photolithogr-aphy)	Fluorescently labeled cDNA (single channel, dual channel)	Hybridization (Southern Blot)	Fluorescence intensities, fold-change ratios (up- or down-regulated)	Visualization, data mining **What do the results mean?**

Analysis

- Low-level analysis
  - Extraction of signal intensities
  - Normalization of samples
- High-level analysis
  - Unsupervised learning (clustering)
    - Aggregation of a collection of data into clusters based on different features in a data set (e.g. heirarchical clustering, SOM)
  - Supervised learning (class discovery)
    - Incorporates knowledge of class label information to make distinctions of interest by using a training set.

Low-level analysis



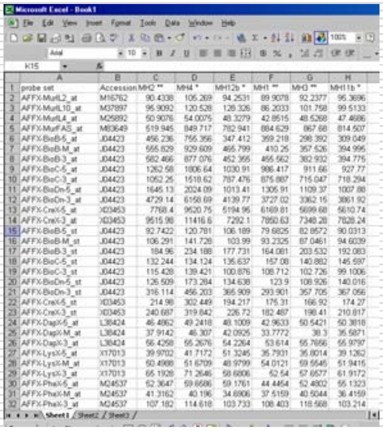
Gene Expression Intensity (Signal)

In other words, a numerical value is obtained

Now, these values can be compared because fluorescense intensity is directly proportional to gene expression

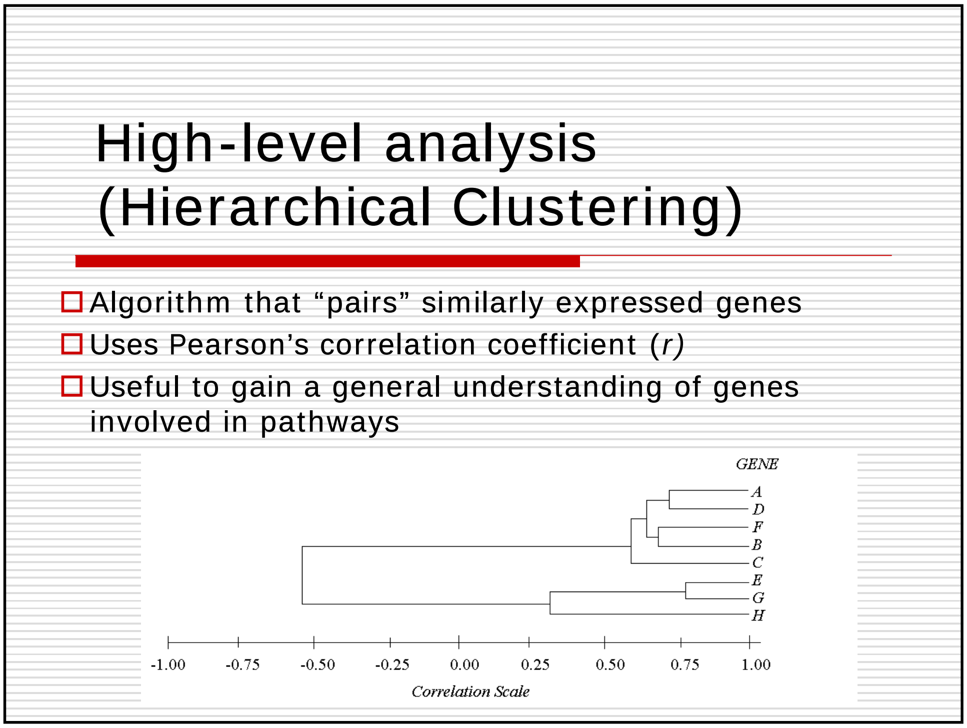
---

High-level analysis



	A	B	C	D	E	F	G	H
1	probe set	Accession	MH2 **	MH4 *	MH12b *	MH1 **	MH3 **	MH11b *
2	AFFX-MurIL2_at	M16762	90.4338	105.269	94.2531	89.9078	92.2377	95.3696
3	AFFX-MurIL10_at	M37897	95.9092	120.528	128.326	86.2033	101.758	99.5133
4	AFFX-MurIL4_at	M25892	50.9076	54.0075	48.3279	42.8515	48.5268	47.4686
5	AFFX-MurFAS_at	M83649	519.945	849.717	782.941	884.629	867.68	814.507
6	AFFX-BioB-5_at	J04423	456.236	755.356	347.412	359.218	298.392	309.049
7	AFFX-BioB-M_at	J04423	555.829	929.609	465.799	410.25	357.526	394.995
8	AFFX-BioB-3_at	J04423	582.466	877.076	452.355	455.562	382.932	394.775
9	AFFX-BioC-5_at	J04423	1262.58	1806.64	1030.91	986.417	911.66	927.77
10	AFFX-BioC-3_at	J04423	1052.25	1518.62	787.476	875.887	715.047	718.294
11	AFFX-BioDn-5_at	J04423	1645.13	2024.09	1013.41	1305.91	1109.37	1007.88
12	AFFX-BioDn-3_at	J04423	4729.14	6158.69	4139.77	3727.02	3362.15	3861.92
13	AFFX-CreX-5_at	X03453	7768.4	9520.75	5194.95	6169.81	5699.68	5610.74
14	AFFX-CreX-3_at	X03453	9515.98	11416.6	7292.1	7850.63	7348.28	7828.24
15	AFFX-BioB-5_st	J04423	92.7422	120.781	106.189	79.6825	82.8572	90.0313
16	AFFX-BioB-M_st	J04423	106.291	141.728	103.99	93.2325	87.0461	94.6039
17	AFFX-BioB-3_st	J04423	184.96	234.188	177.731	164.081	203.532	192.083
18	AFFX-BioC-5_st	J04423	132.244	134.124	135.637	157.08	140.882	145.597
19	AFFX-BioC-3_st	J04423	115.428	139.421	100.876	108.712	102.726	99.1006
20	AFFX-BioDn-5_st	J04423	126.509	173.284	134.638	123.9	108.926	140.016
21	AFFX-BioDn-3_st	J04423	316.114	456.203	365.909	293.901	357.705	367.056
22	AFFX-CreX-5_st	X03453	214.98	302.449	194.217	175.31	166.92	174.27
23	AFFX-CreX-3_st	X03453	240.687	319.842	226.72	182.487	198.41	210.817
24	AFFX-DapX-5_at	L38424	46.4862	49.2418	48.1009	42.9633	50.5421	50.3818
25	AFFX-DapX-M_at	L38424	37.9142	46.307	42.0925	33.7772	38.3	35.5871
26	AFFX-DapX-3_at	L38424	56.4258	55.2676	54.2264	53.614	55.7656	55.9797
27	AFFX-LysX-5_at	X17013	39.9702	41.7172	51.3245	35.7931	35.8014	39.1262
28	AFFX-LysX-M_at	X17013	50.4988	51.6709	48.9799	54.0121	59.5545	51.9415
29	AFFX-LysX-3_at	X17013	65.1928	71.2646	58.6806	52.54	57.6577	61.9172
30	AFFX-PheX-5_at	M24537	52.3647	59.6586	59.1761	44.4454	52.4802	55.1323
31	AFFX-PheX-M_at	M24537	41.3162	40.196	34.6906	37.5159	40.5044	36.4159
32	AFFX-PheX-3_at	M24537	107.182	114.618	103.733	108.403	118.568	103.214

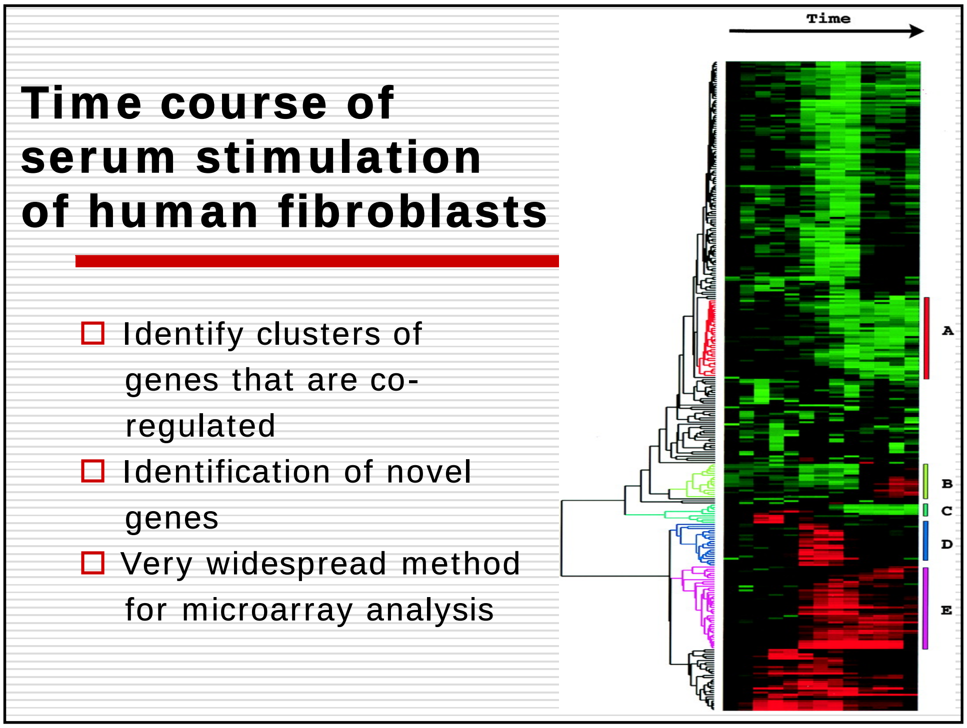
Now what??



High-level analysis (Hierarchical Clustering)

- ☐ Algorithm that “pairs” similarly expressed genes
- ☐ Uses Pearson’s correlation coefficient (*r*)
- ☐ Useful to gain a general understanding of genes involved in pathways

---

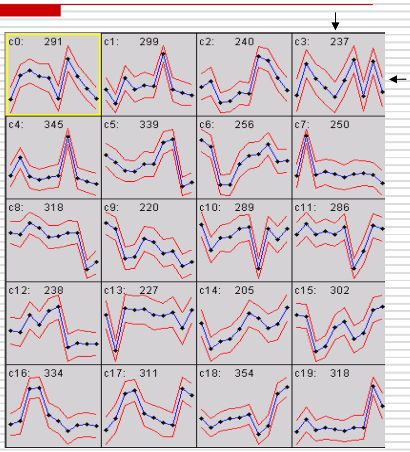


Time course of serum stimulation of human fibroblasts

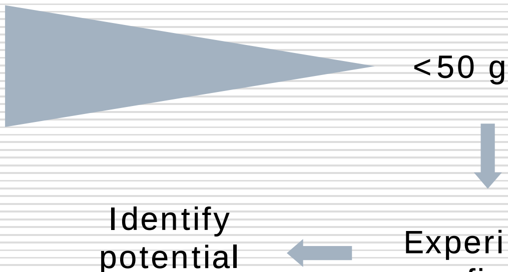
- ☐ Identify clusters of genes that are co-regulated
- ☐ Identification of novel genes
- ☐ Very widespread method for microarray analysis

High-level analysis (self-organizing maps)

- Algorithm that clusters genes based on similar expression values
- Useful for finding patterns in biological data
- Cocaine study
- 5 regions of the rat brain under treated and untreated conditions
- e.g. cluster 3



Overall Goal



>10,000 genes

<50 genes

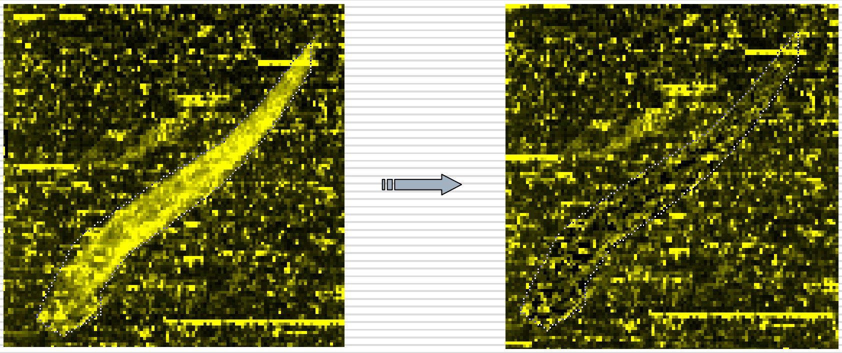
Experimental confirmation

Identify potential therapeutic targets

Potential Problems

- Local contamination

Array Contamination



Potential Problems

- Local contamination
- Normalization
- Statistical significance of difference in expression
- cDNA arrays
  - must have the genes cloned
  - need relatively pure product
- Affymetrix arrays
  - need sequence information

Additional Reading

- Affymetrix website: www.affymetrix.com
- Stanford University: genome-www.stanford.edu
- Nature Genetics, vol. 21 supplement, “The Chipping Forecast”
- www.microarray.org
- www.gene-chips.com/
- ihome.cuhk.edu.hk/~b400559/array.html
- www.stat.wisc.edu/~yandell/statgen/reference/array.html