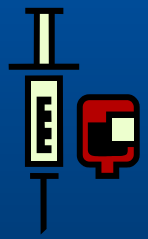
Advanced Artificial Intelligence : Gene Expression Dataset

October 7, 2004

Study

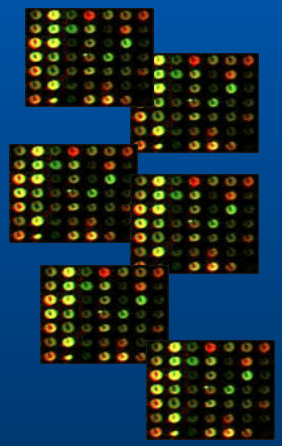
- Treatment-specific changes in gene expression discriminate *in vivo* drug response in human leukemia cells, MH Cheok *et al.*, *Nature Genetics* 35, 2003.

60 leukemia patients



Bone marrow samples

Affymetrix GeneChip arrays



Gene expression data

Gene Expression Data

- # of data examples
  - 120 (60: before treatment, 60: after treatment)
- # of genes measured
  - 12600 (Affymetrix HG-U95A array)
- Task
  - Classification between “before treatment” and “after treatment” based on gene expression pattern

Affymetrix GeneChip Arrays

- Use short oligos to detect gene expression level.
- Each gene is probed by a set of short oligos.
- Each gene expression level is summarized by
  - Signal: numerical value describing the abundance of mRNA
  - A/P call: denotes the statistical significance of signal

post	HDMTX-1	HDMTX-1	HDMTX-1	HDMTX-1	HDMTX-1	HDMTX-1
AFFX-MurIL	131	A	0.814869	114.8	A	0.99156
AFFX-MurIL	195.5	A	0.897835	330.8	A	0.945787
AFFX-MurIL	37.7	A	0.986189	113.4	A	0.993129
AFFX-MurF/	368.7	A	0.737173	401.4	A	0.979978
AFFX-BioB-	6191.1	P	0.000662	27842.1	P	0.00034
AFFX-BioB-	15308.8	P	0.00007	49224.5	P	0.00006
AFFX-BioB-	6877.9	P	0.000044	22224.5	P	0.00006
AFFX-BioC-	22128.7	P	0.00006	71208.3	P	0.000044
AFFX-BioC-	18752.3	P	0.000044	56108.3	P	0.000052
AFFX-BioDı	27926.1	P	0.000044	79416.9	P	0.000044
AFFX-BioDı	110043.7	P	0.000044	292185.6	P	0.000044

Preprocessing

- Remove the genes having more than 60 ‘A’ calls
  - # of genes: 12600 → 3190
- Discretization of gene expression level
  - Criterion: median gene expression value of each sample
  - 0 (low) and 1 (high)

ProbeSetI	HDMTX.1	HDMTX.11	HDMTX.12	HDMTX.2	HDMTX.13	HDMTX.14	HDMTX.3	HDMTX.4	HDMTX.5	HDMTX.6
AFFX-BioB	1	1	0	1	1	1	1	1	1	0
AFFX-BioB	1	1	1	1	1	1	1	1	1	1
AFFX-BioB	1	1	0	1	1	1	1	1	1	0
AFFX-BioC	1	1	1	1	1	1	1	1	1	1
AFFX-BioC	1	1	1	1	1	1	1	1	1	1
AFFX-BioDı	1	1	1	1	1	1	1	1	1	1
AFFX-BioDı	1	1	1	1	1	1	1	1	1	1
AFFX-CreX	1	1	1	1	1	1	1	1	1	1
AFFX-CreX	1	1	1	1	1	1	1	1	1	1
AFFX-hum_	1	1	1	1	1	1	1	1	1	1
AFFX-HUMI	0	0	0	0	0	0	0	0	0	0
AFFX-HUMI	0	1	1	1	1	0	1	1	1	0
AFFX-HUM	1	1	1	1	1	1	1	1	1	1
AFFX-HUM	1	1	1	1	1	1	1	1	1	1
AFFX-HUM	1	1	1	1	1	1	1	1	1	1
AFFX-HSAC	1	1	1	1	1	1	1	1	1	1
AFFX-HSAC	1	1	1	1	1	1	1	1	1	1
AFFX-HSAC	1	1	1	1	1	1	1	1	1	1
AFFX-HSAC	0	1	0	0	0	1	0	0	0	0
31330_at	1	1	1	1	1	1	1	1	1	1

Gene Filtering

- Using mutual information

$$I(G;C) = \sum_{G,C} P(G,C) \frac{\log P(G,C)}{\log P(G)P(C)}$$

  - Estimated probabilities were used.
  - # of genes: 3190 → 1000
- Final dataset
  - # of attributes: 1001 (one for the class)
    - Class: 0 (after treatment), 1 (before treatment)
  - # of data examples: 120

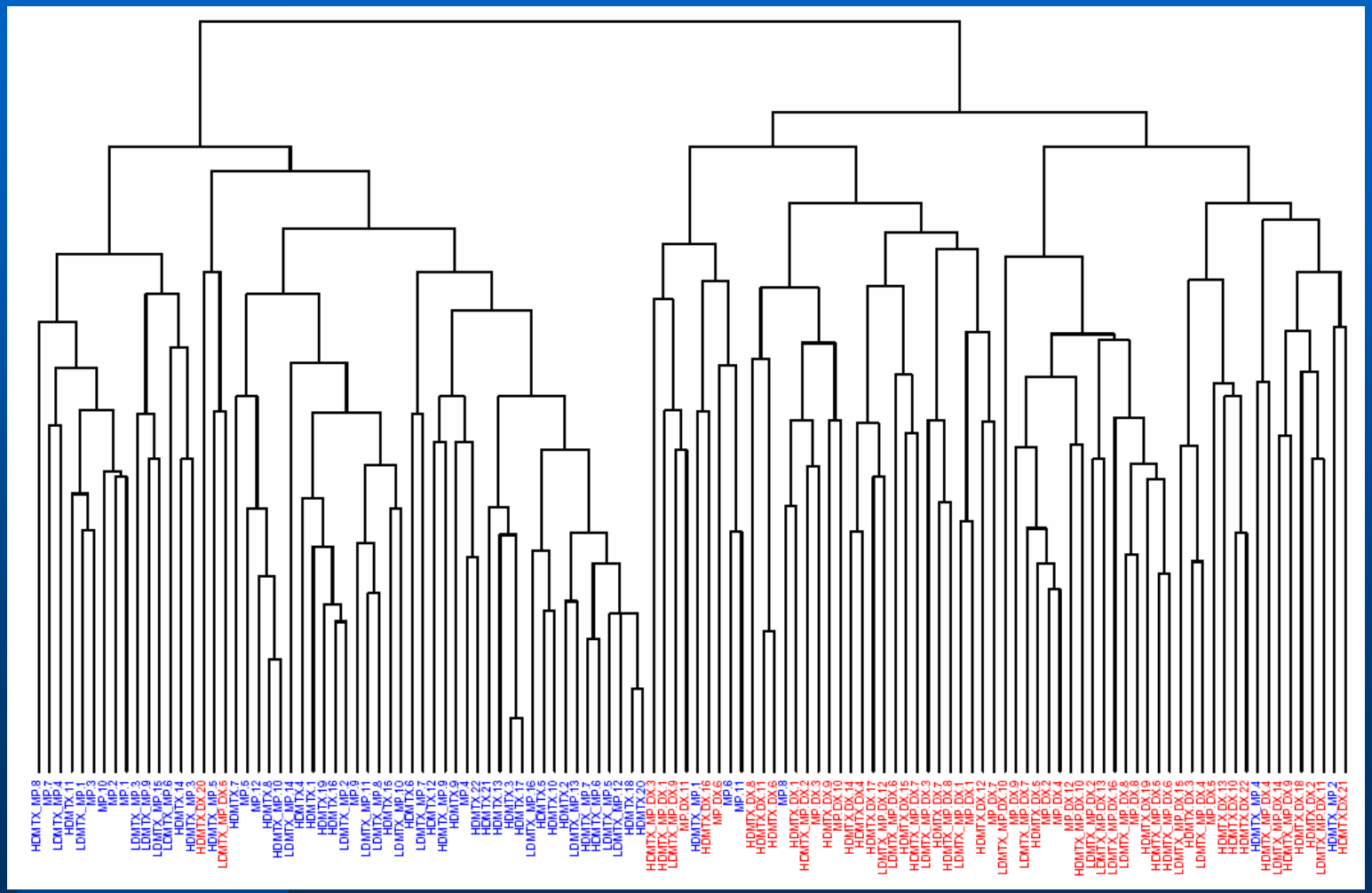
Final Dataset

ProbeSetID	HDMTX.1	HDMTX.11	HDMTX.12	HDMTX.2	HDMTX.13	HDMTX.14	HDMTX.3	MI
36822_at	1	1	1	1	0	1	0	2.79E-01
1894_f_at	0	1	0	1	1	1	0	2.53E-01
34836_at	0	0	0	1	0	1	0	2.34E-01
1915_s_at	0	0	0	0	0	0	0	2.16E-01
38982_at	0	0	0	0	0	0	0	2.14E-01
32815_at	0	0	0	0	0	0	0	2.03E-01
38072_at	0	1	0	1	1	0	1	1.96E-01
39332_at	0	0	0	0	0	0	0	1.91E-01
1916_s_at	0	0	0	0	0	0	0	1.68E-01
36950_at	1	1	1	1	1	1	1	1.53E-01
33016_at	0	1	0	0	1	0	0	1.53E-01
32833_at	1	1	1	1	1	1	1	1.53E-01
292_s_at	1	1	0	1	1	1	1	1.51E-01
34231_at	0	0	0	0	0	0	0	1.48E-01
632_at	0	1	0	0	0	0	0	1.47E-01
31936_s_at	1	1	1	1	1	1	1	1.47E-01
37028_at	1	0	0	0	0	0	0	1.47E-01
37645_at	0	0	0	0	0	0	0	1.42E-01
1937_at	0	1	0	0	0	0	0	1.41E-01
33870_at	0	0	0	0	0	0	0	1.38E-01

1000

120

Hierarchical Clustering of Samples



after treatment

before treatment

Classification of Samples by SVM