Our view on cDNA chip analysis from engineering informatics standpoint

Chonghun Han, **Sungwoo Kwon**

Intelligent Process System Lab

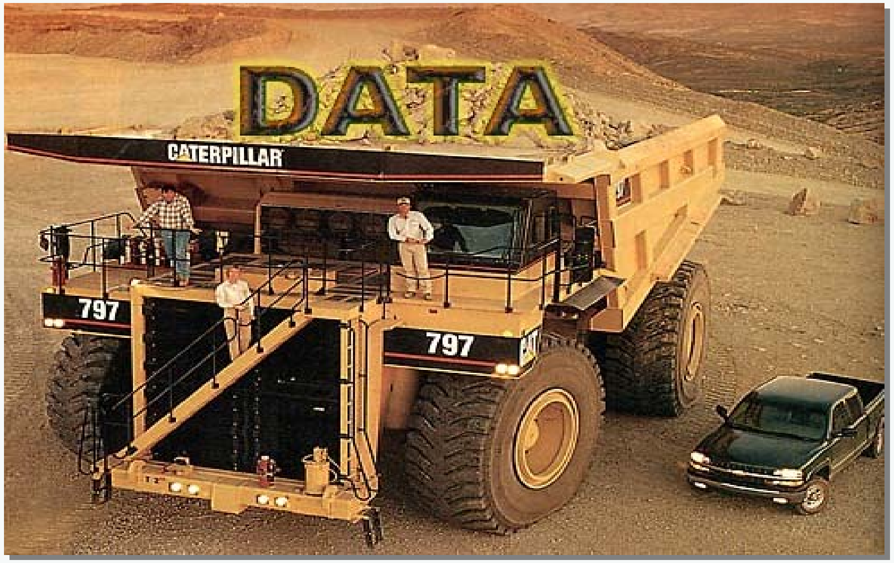
Department of Chemical Engineering

Pohang University of Science and Technology

Outline

- Introduction to Bioinformatics
- Introduction to cDNA chip
- Classification of Tumor Classes
- Identification of Marker Genes
- Conclusions and Future Works

The Information Revolution



I need tools to extract important information from mountains of data

Data Mining

- Data mining = 'exploration and analysis by automatic and semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules'

Applications of data mining

- ✓ Search database
- ✓ Structural pattern recognition
- ✓ Medline abstract analysis
- ✓ DNA chip data analysis

Who has information and uses it wins

(Watterson, K)

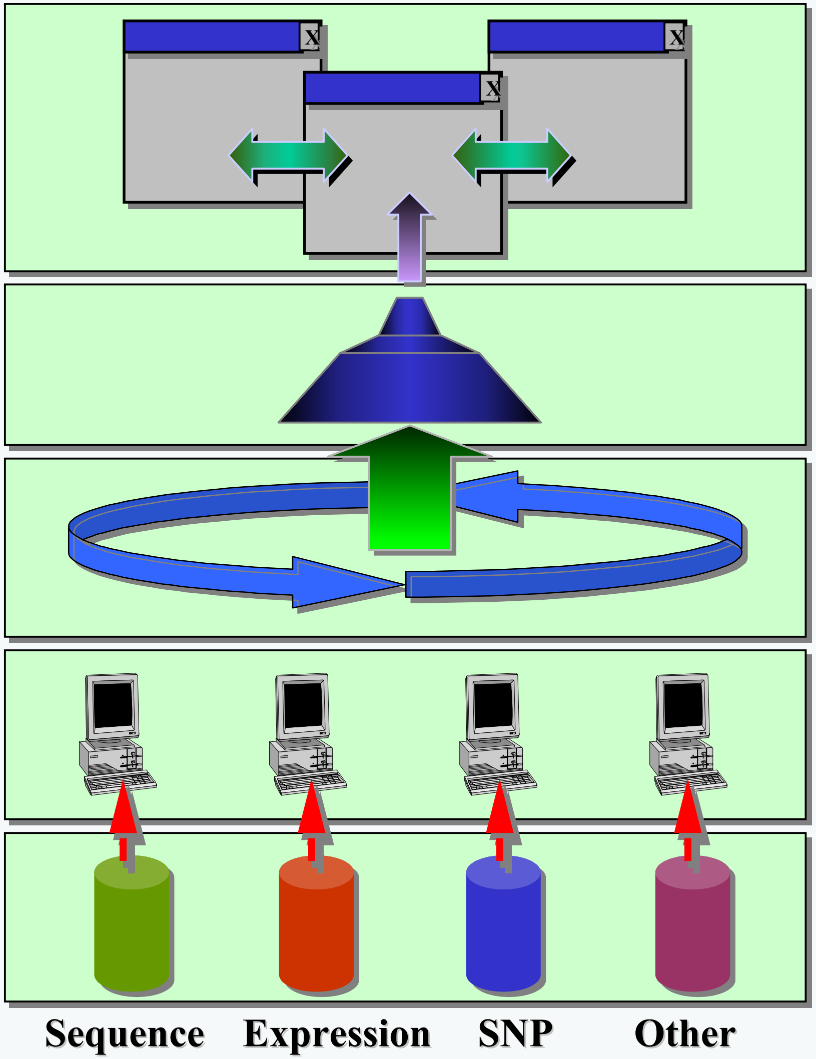
Data Mining + Domain Knowledge = Engineering Informatics

- **Informatics (정보 과학):** *the study of the structure, behavior, and interactions of natural and artificial computational systems*
- **Informatics = Information + Mathematics**
- **Application Areas**
  - ✓ **Bioinformatics (or Biomolecular Informatics)**
  - ✓ **Cheminformatics**
  - ✓ **Environmental Informatics**
  - ✓ **Medical Informatics**
  - ✓ **Neuro Informatics**
  - ✓ **Process Informatics**
  - ✓ **Many More …**

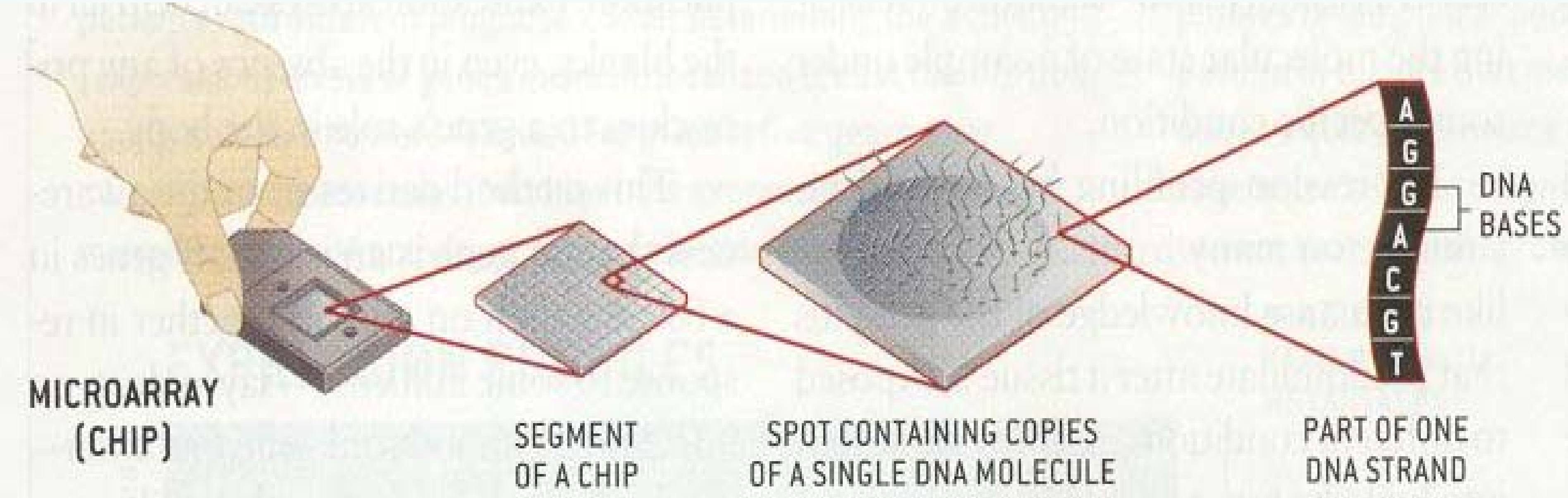
Bioinformatics

Bioinformatics =
Biology + Informatics

- Artificial Intelligence
- Combinatorial Optimization
- Data Mining
- Digital Signal Processing
- Machine Learning
- Mathematical Modeling
- Multivariate Statistics
- Pattern Recognition
- System identification
- …



DNA chip


MICROARRAY
(CHIP)
SEGMENT
OF A CHIP
SPOT CONTAINING COPIES
OF A SINGLE DNA MOLECULE
A
G
G
A
C
G
T
DNA
BASES
PART OF ONE
DNA STRAND

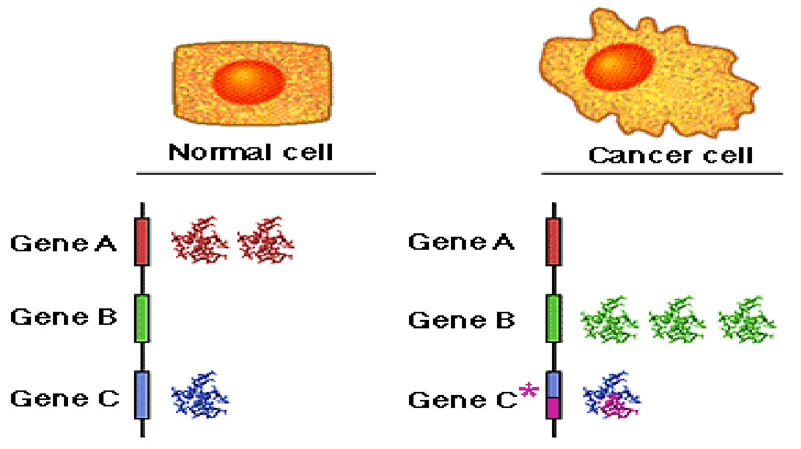
The biological meaning of DNA chip

- Genome map is completed
  - ✓ Need to study functional genomic
- Know who, when, where, why, how much gene expressed
  - ✓ To classify different types of diseases (ex. Cancer types)
  - ✓ To understand the behavior of a biological system
  - ✓ To understand cell dynamics
- Can systematically disturb cell
- DNA chip experiment and data analysis are different matter
  - ✓ Methods of data analysis variant result of DNA chip experiment
  - ✓ Require suitable method of data analysis for DNA chip experiment objective

Application of DNA chip

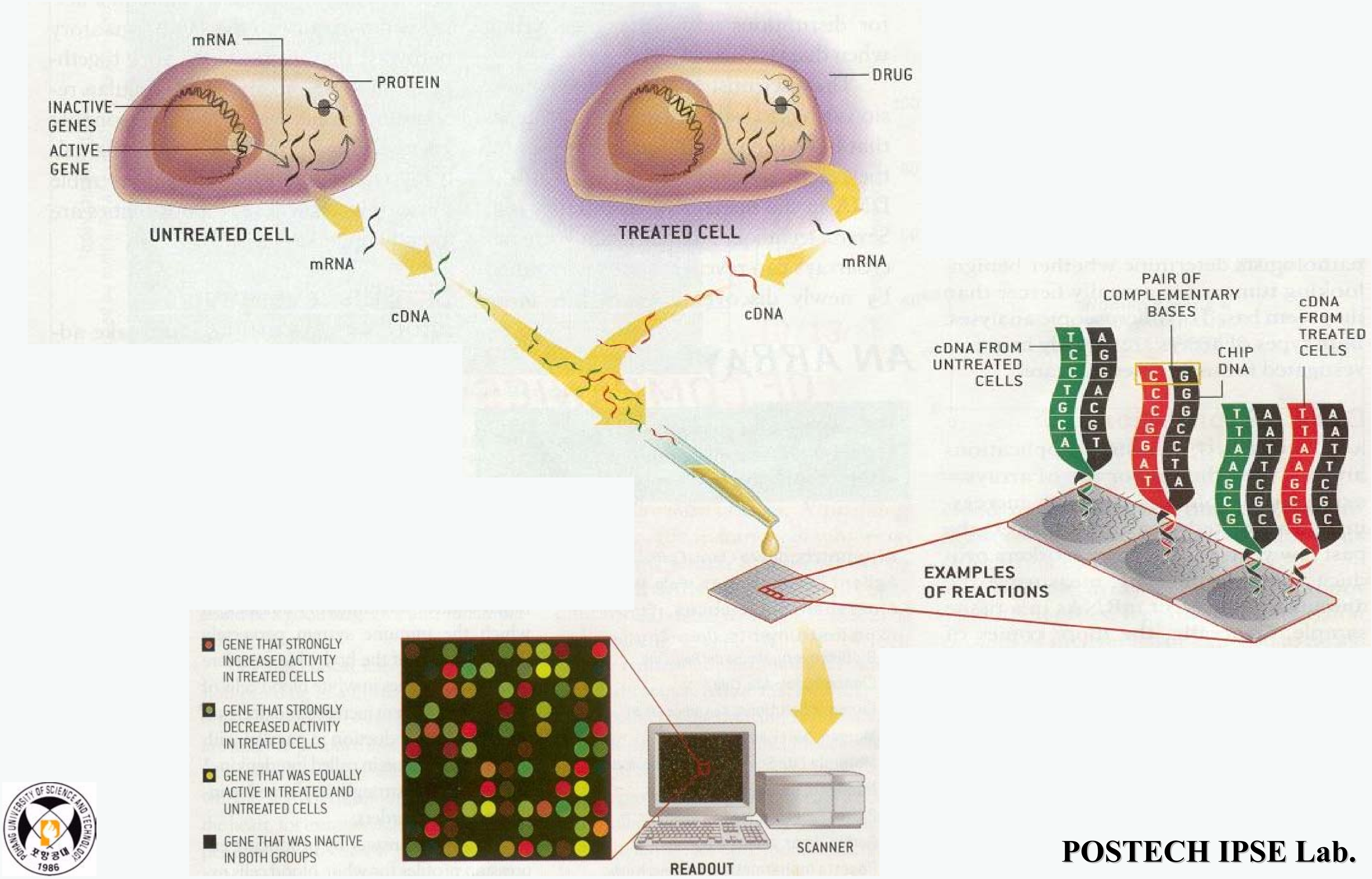
- Analysis of gene expression and regulation
  - ✓ Genetic network, pathway analysis, metabolic engineering
- Disease diagnosis
  - ✓ Molecular cancer classification, the discovery of disease subtype, The marker gene discovery
- And many more…

- Cancer diagnosis



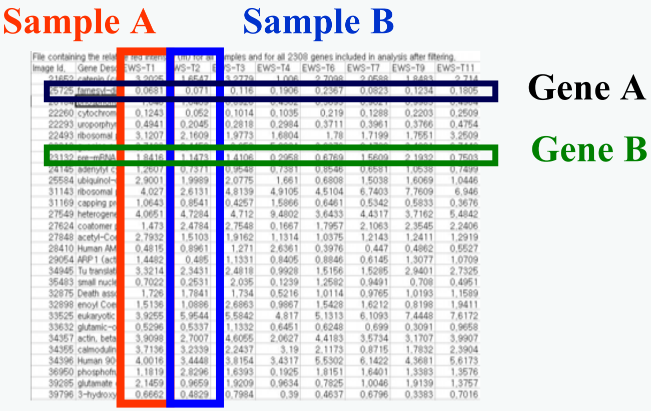
✓ Because significantly different groups of genes are expressed by many type of cell, we can fingerprint characteristic cell

cDNA chip: Lab Experiment


mRNA
PROTEIN
INACTIVE GENES
ACTIVE GENE
UNTREATED CELL
mRNA
cDNA
DRUG
TREATED CELL
mRNA
cDNA
PAIR OF COMPLEMENTARY BASES
cDNA FROM TREATED CELLS
cDNA FROM UNTREATED CELLS
CHIP DNA
TCCTGCA
AGGACGT
CCCGGAT
GGGCCTA
TTAAGCG
AATTCGC
TTAAGCG
AATTCGC
EXAMPLES OF REACTIONS
GENE THAT STRONGLY INCREASED ACTIVITY IN TREATED CELLS
GENE THAT STRONGLY DECREASED ACTIVITY IN TREATED CELLS
GENE THAT WAS EQUALLY ACTIVE IN TREATED AND UNTREATED CELLS
GENE THAT WAS INACTIVE IN BOTH GROUPS
SCANNER
READOUT

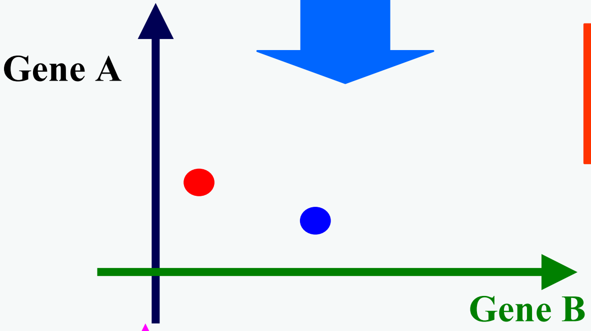
POSTECH IPSE Lab.

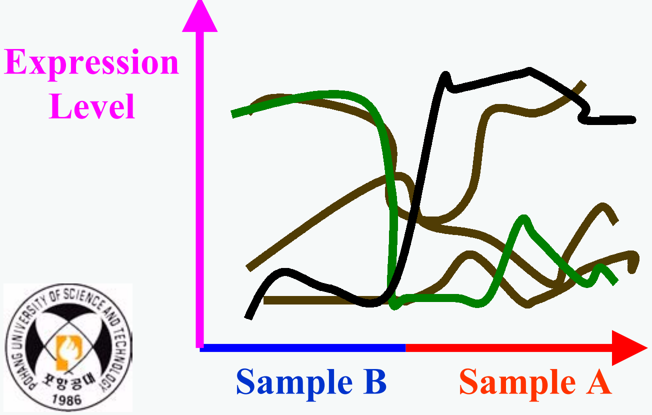
cDNA chip: Data Analysis


Sample A
Sample B
Gene A
Gene B

File containing the rela[illegible] red intens[illegible] (m) for all [illegible]mples and for all 2308 genes included in analysis after filtering.

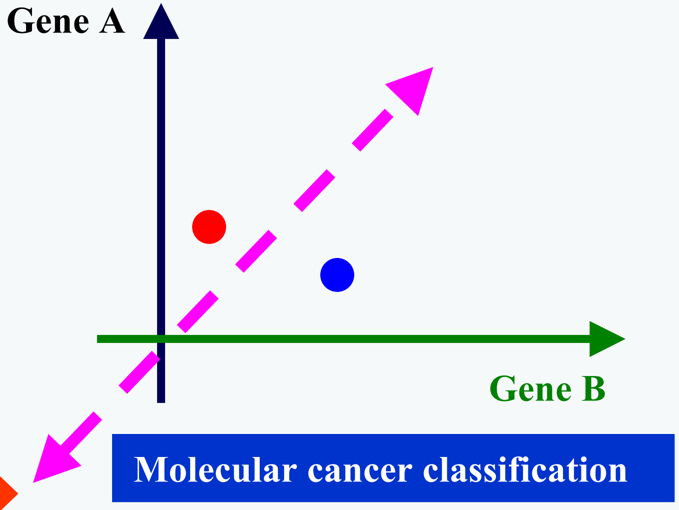
Image Id.	Gene Desc	EWS-T1	EWS-T2	EWS-T3	EWS-T4	EWS-T6	EWS-T7	EWS-T9	EWS-T11
21652	catenin (c	3,2025	1,6547	3,2779	1,006	2,7098	2,0588	1,8483	2,714
25725	farnesyl-d	0,0681	0,071	0,116	0,1906	0,2367	0,0823	0,1234	0,1805
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
22260	cytochrom	0,1243	0,052	0,1014	0,1035	0,219	0,1288	0,2203	0,2509
22293	uroporphyr	0,4941	0,2045	0,2818	0,2984	0,3711	0,3961	0,3766	0,4754
22493	ribosomal	3,1207	2,1609	1,9773	1,6804	1,78	1,7199	1,7551	3,2509
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
23132	pre-mRNA	1,8416	1,1473	1,4106	0,2958	0,6769	1,5609	2,1932	0,7503
24145	adenylyl c	1,2607	0,7371	0,9548	0,7381	0,8546	0,6581	1,0538	0,7499
25584	ubiquinol-	2,9001	1,9989	2,0775	1,661	0,6808	1,5038	1,6069	1,0446
31143	ribosomal	4,027	2,6131	4,8139	4,9105	4,5104	6,7403	7,7609	6,946
31169	capping pr	1,0643	0,8541	0,4257	1,5866	0,6461	0,5342	0,5833	0,3676
27549	heterogene	4,0651	4,7284	4,712	9,4802	3,6433	4,4317	3,7162	5,4842
27624	coatomer	1,473	2,4784	2,7548	0,1667	1,7957	2,1063	2,3545	2,2406
27848	acetyl-Co	2,7932	1,5103	1,9162	1,1314	1,0375	1,2143	1,2411	1,2919
28410	Human AM	0,4815	0,8961	1,271	2,6361	0,3976	0,447	0,4862	0,5527
29054	ARP1 (act	1,4482	0,485	1,1331	0,8405	0,8846	0,6145	1,3077	1,0709
34945	Tu translat	3,3214	2,3431	2,4818	0,9928	1,5156	1,5285	2,9401	2,7325
35483	small nucle	0,7022	0,2531	2,035	0,1239	1,2582	0,9491	0,708	0,4951
32875	Death ass	1,726	1,7841	1,734	0,5216	1,0114	0,9765	1,0193	1,1589
32898	enoyl Coe	1,5136	1,0886	2,6863	0,9867	1,5428	1,6212	0,8198	1,9411
33525	eukaryotic	3,9255	5,9544	5,5842	4,817	5,1313	6,1093	7,4448	7,6172
33632	glutamic-o	0,5296	0,5337	1,1332	0,6451	0,6248	0,699	0,3091	0,9658
34357	actin, beta	3,9098	2,7007	4,6055	2,0627	4,4183	3,5734	3,1707	3,9907
34355	calmodulin	3,7136	3,2339	2,2437	3,19	2,1173	0,8715	1,7832	2,3904
34396	Human 90	4,0016	3,4448	3,8154	3,4317	5,5302	6,1422	4,3681	5,6173
36950	phosphofru	1,1819	2,8296	1,6393	0,1925	1,8151	1,6401	1,3383	1,3576
39285	glutamate	2,1459	0,9659	1,9209	0,9634	0,7825	1,0046	1,9139	1,3757
39796	3-hydroxy	0,6662	0,4829	0,7984	0,39	0,4637	0,6796	0,3383	0,7016



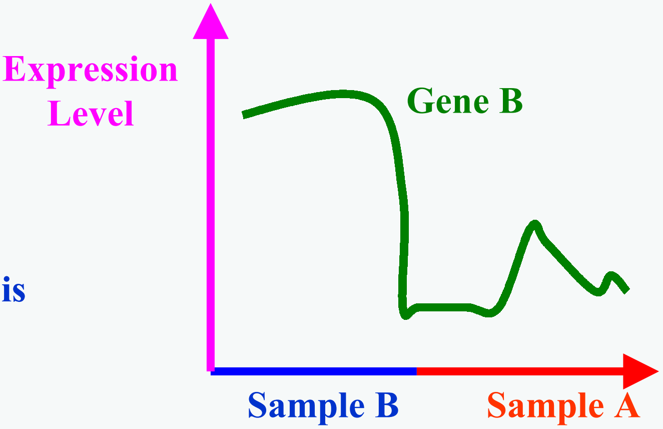


Statistical analysis (data mining)

- Hierarchal clustering
- K-means clustering
- Self-organizing map
- Neural network
- Bayesian decision theory
- Principal component analysis
- And many more…



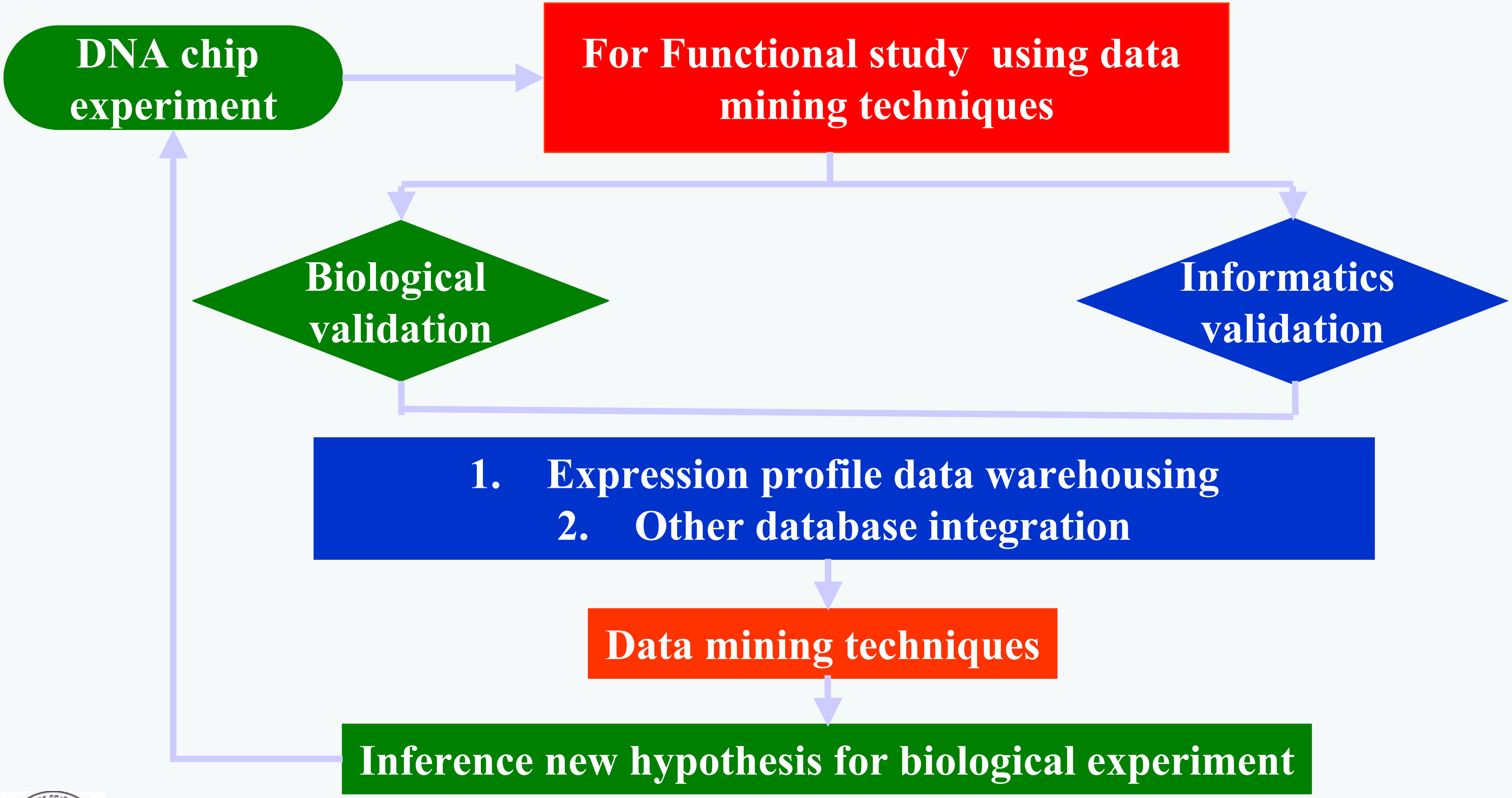
Molecular cancer classification



Identification of the potential marker gene

POSTECH IPSE Lab.

cDNA chip: Procedure



DNA chip experiment

For Functional study using data mining techniques

Biological validation

Informatics validation

1. Expression profile data warehousing
2. Other database integration

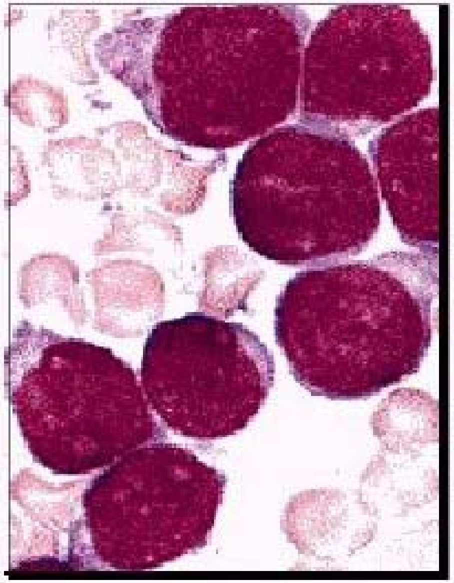
Data mining techniques

Inference new hypothesis for biological experiment

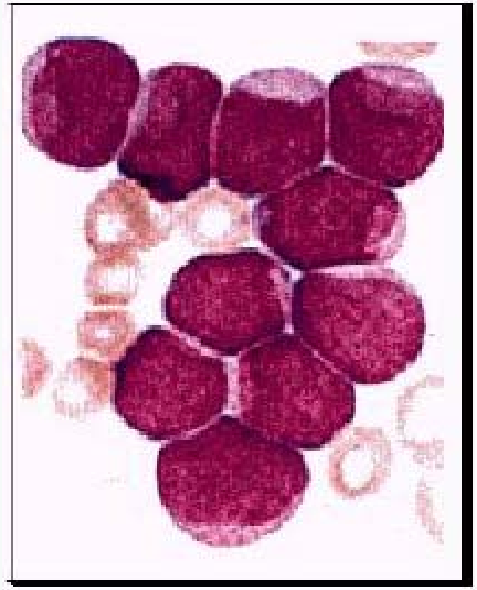
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

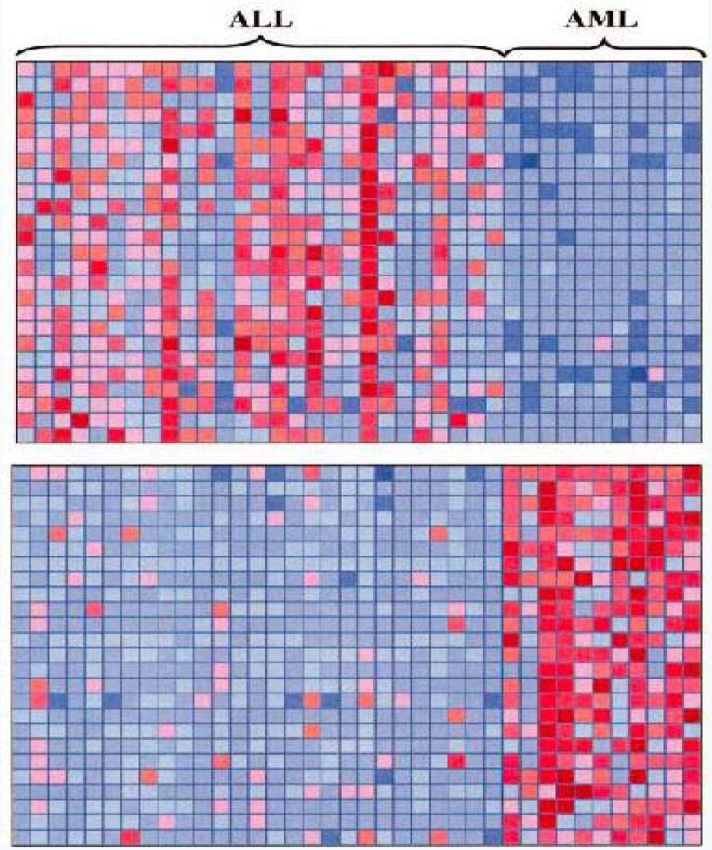
Moving from morphological to molecular classification

Acute lymphoblastic leukemia (ALL)



Acute myelogenous leukemia (AML)



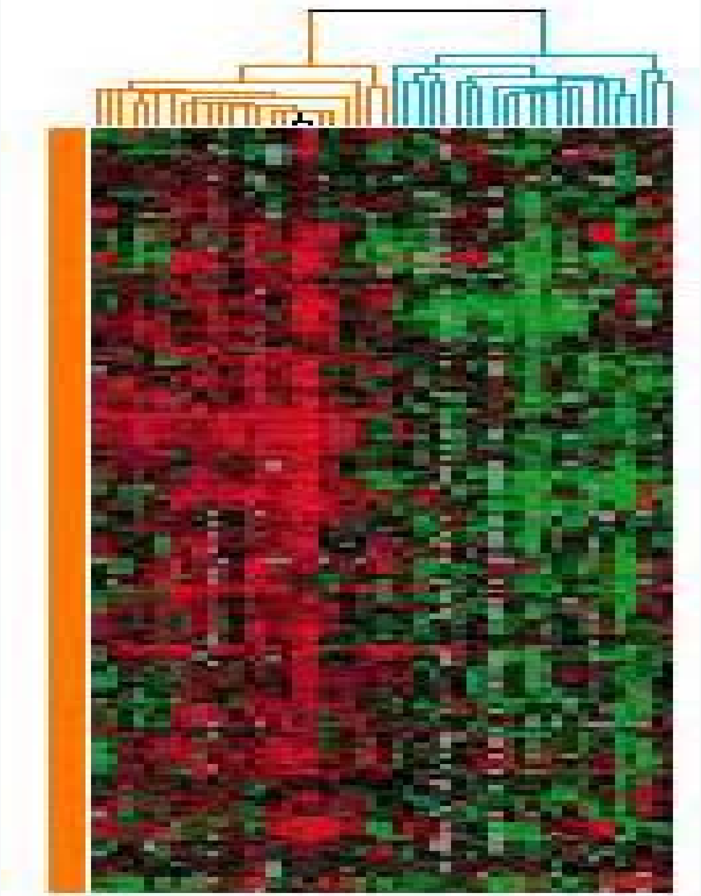


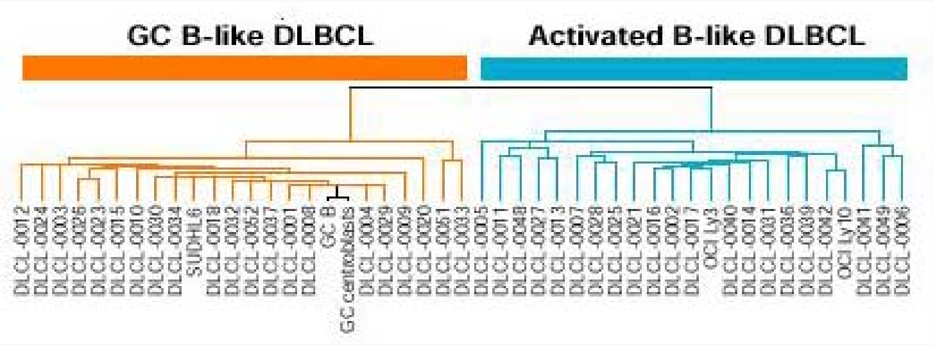
Golub.T.R.,et al. 1999. *Science*, 286, 531-537.

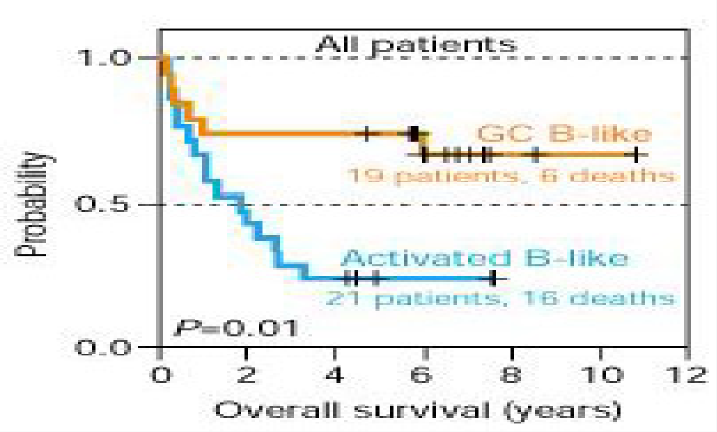
POSTECH IPSE Lab.

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

- There are new cancer class discovery, two molecularly distinct forms of B-cell lymphoma (DLBCL) that are composed of GC B-like and Activated B-like DLBCL





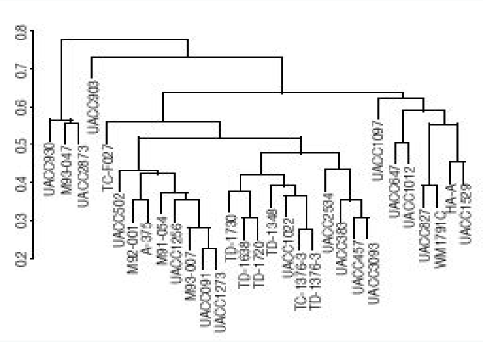


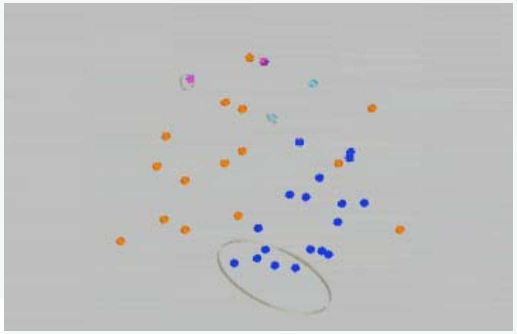
Ash A. Alizadeh.,et al. 2000. *Nature,* 403, 503-511

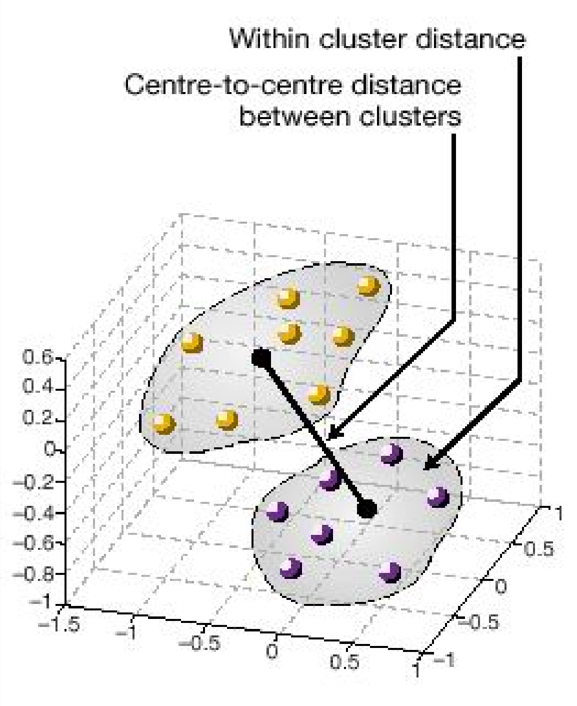
Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profile

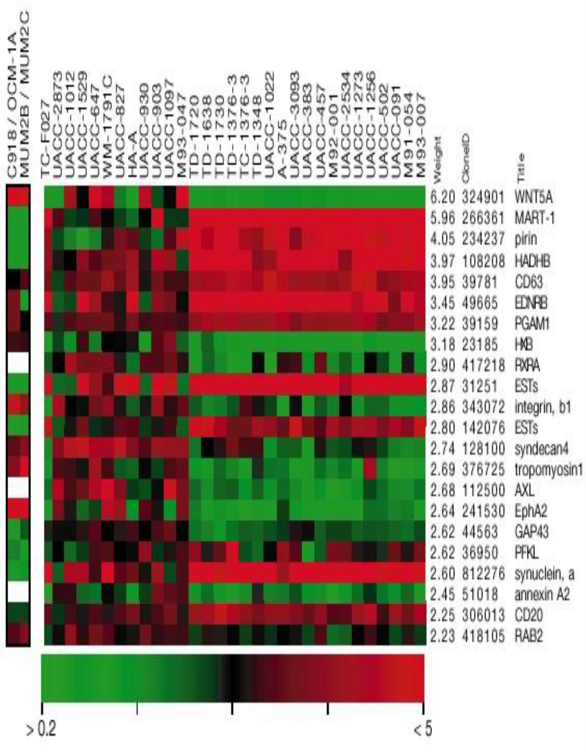
Limitation,

- ✓ Because weighting method based on univariative or bivariative statistical analysis, we can not capture correlated structure in the data
- ✓ When multi-class cancer classify, it is hard to know whether highly express or not









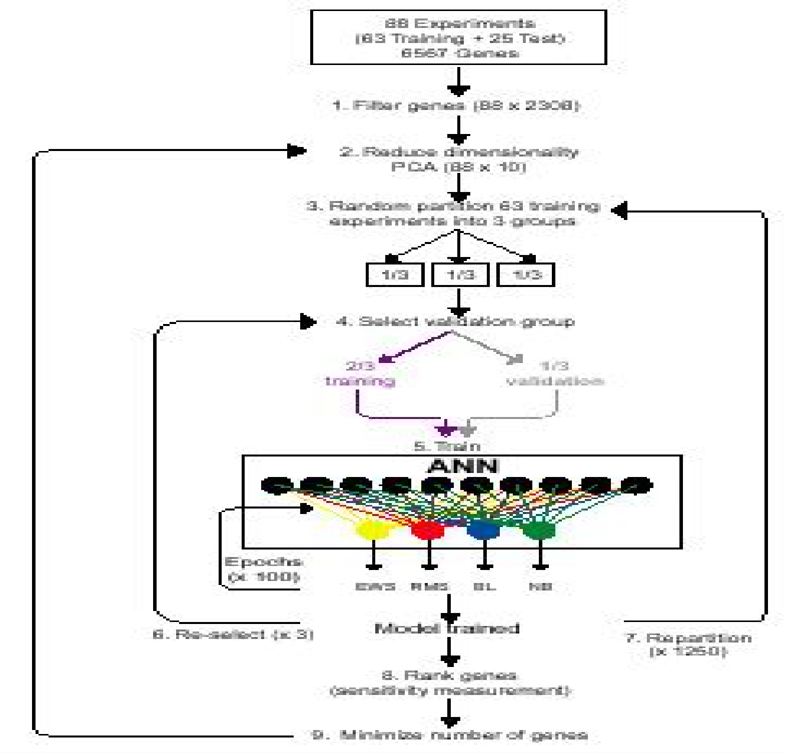
M. Bittner.,et al. 2000. *Nature*, 406, 536-540.

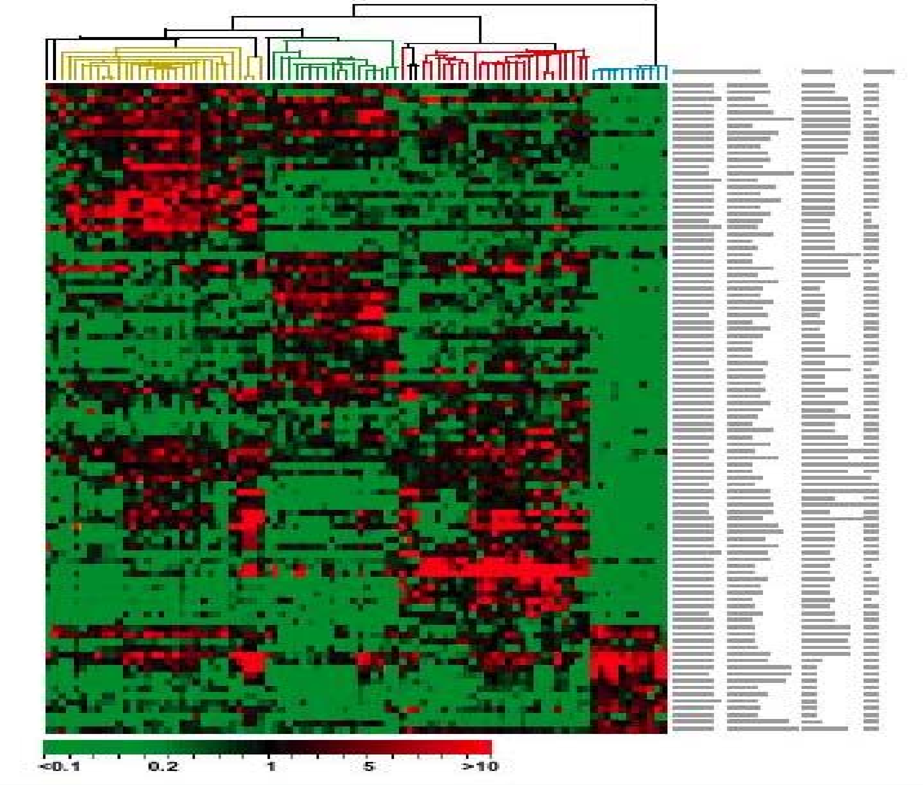
POSTECH IPSE Lab.

Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Network

Limitation,

- ✓ Because relevant gene extraction method based on univariative statistical analysis, we can not capture correlated structure in the data

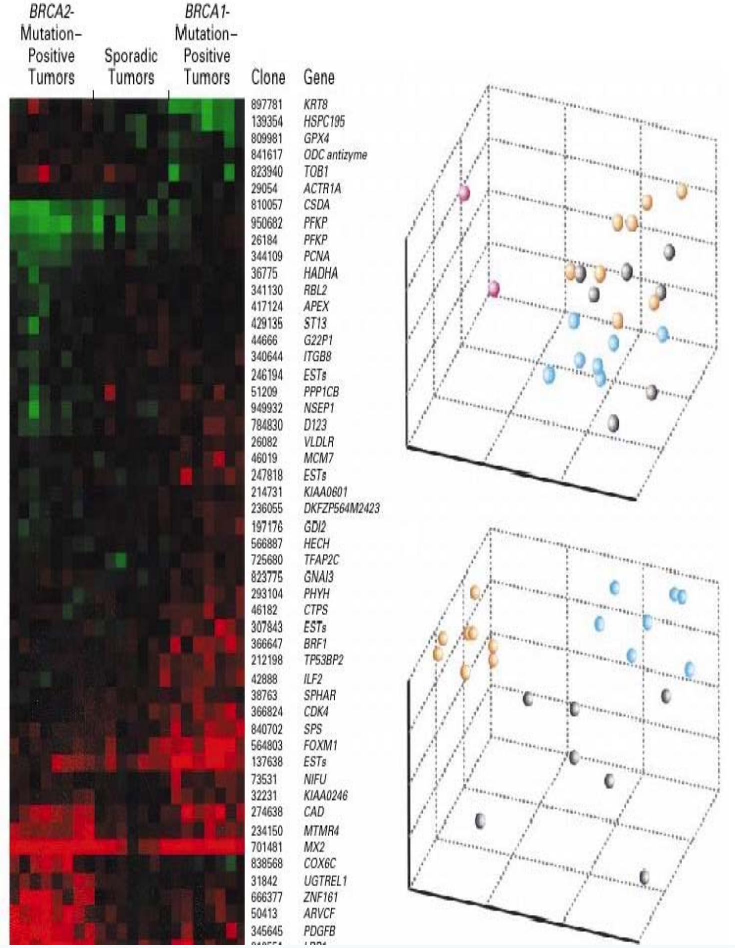


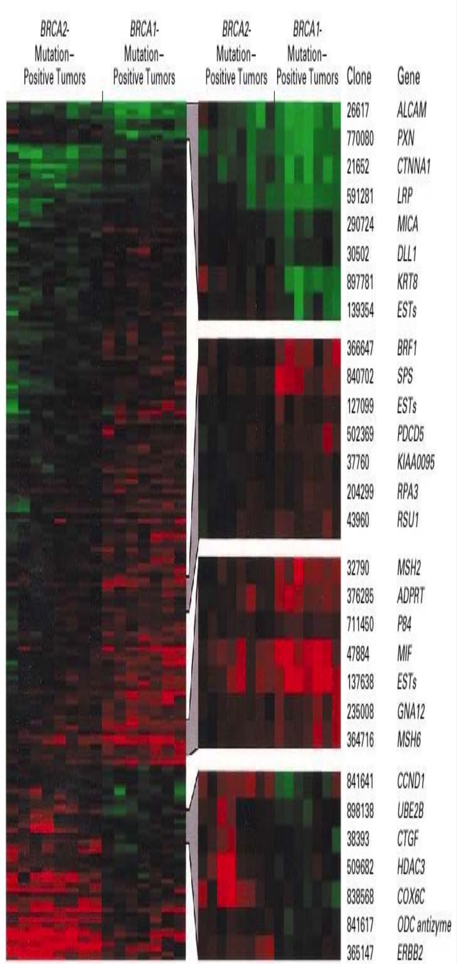


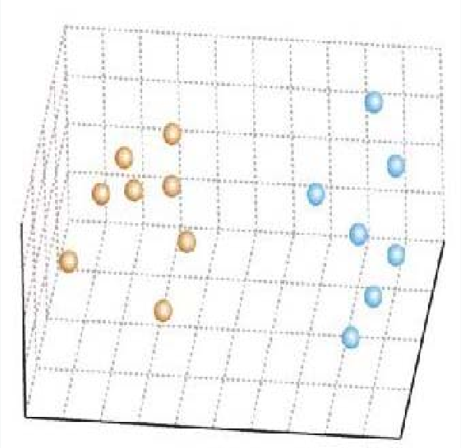
Javed. Kahn .,et al. 2001. *Nature*, 17, 1131-1142.

Gene-expression profiles in hereditary breast cancer

- Limitation,
  - ✓ Because relevant gene extraction method based on univariative statistical analysis, we can not capture correlated structure in the data
  - ✓ When multi-class cancer classify, it is hard to know whether highly express or not



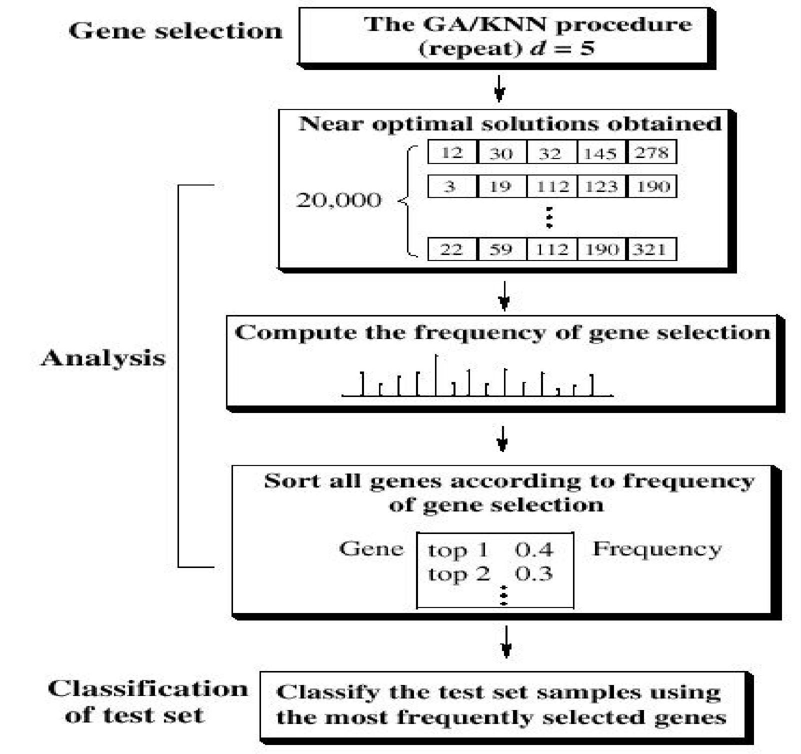




Ingrid Hedenfalk ., et al. 2001. *N Eng j Med*, 344, 539-548.

Gene Selection for Sample Classification based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method

- Limitation,
  - ✓ It is difficult to determine parameter value
  - ✓ When multi-class cancer classify, it is hard to know whether highly express or not
  - ✓ Computing time take a long time



Gene selection

The GA/KNN procedure (repeat) $d = 5$

Analysis

Near optimal solutions obtained

20,000

12	30	32	145	278
3	19	112	123	190
⋮				
22	59	112	190	321

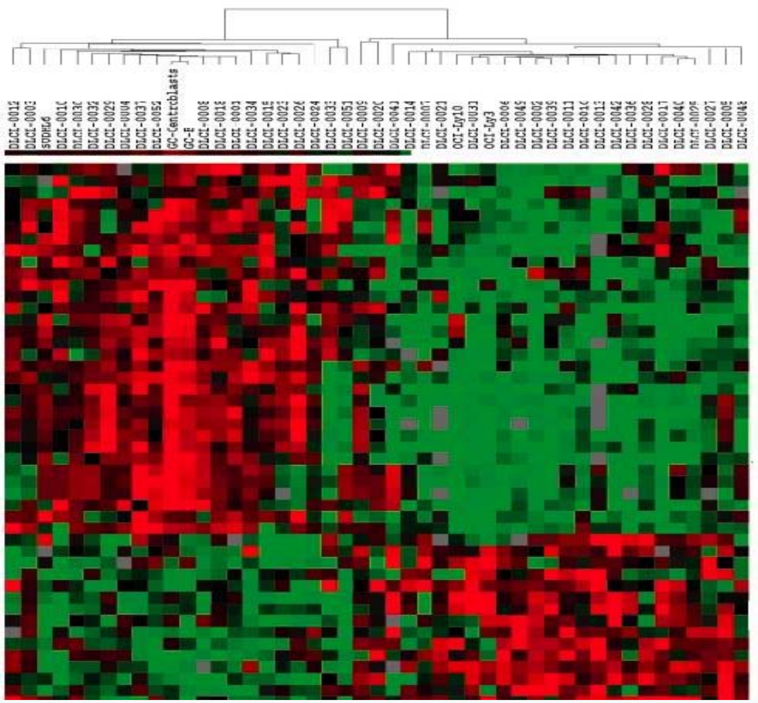
Compute the frequency of gene selection

Sort all genes according to frequency of gene selection

Gene		Frequency
top 1	0.4	
top 2	0.3	
⋮		

Classification of test set

Classify the test set samples using the most frequently selected genes



Leping. L .,et al. 2001. *bioinformatics,* 17, 1131-1142.

Limitations and Improvements

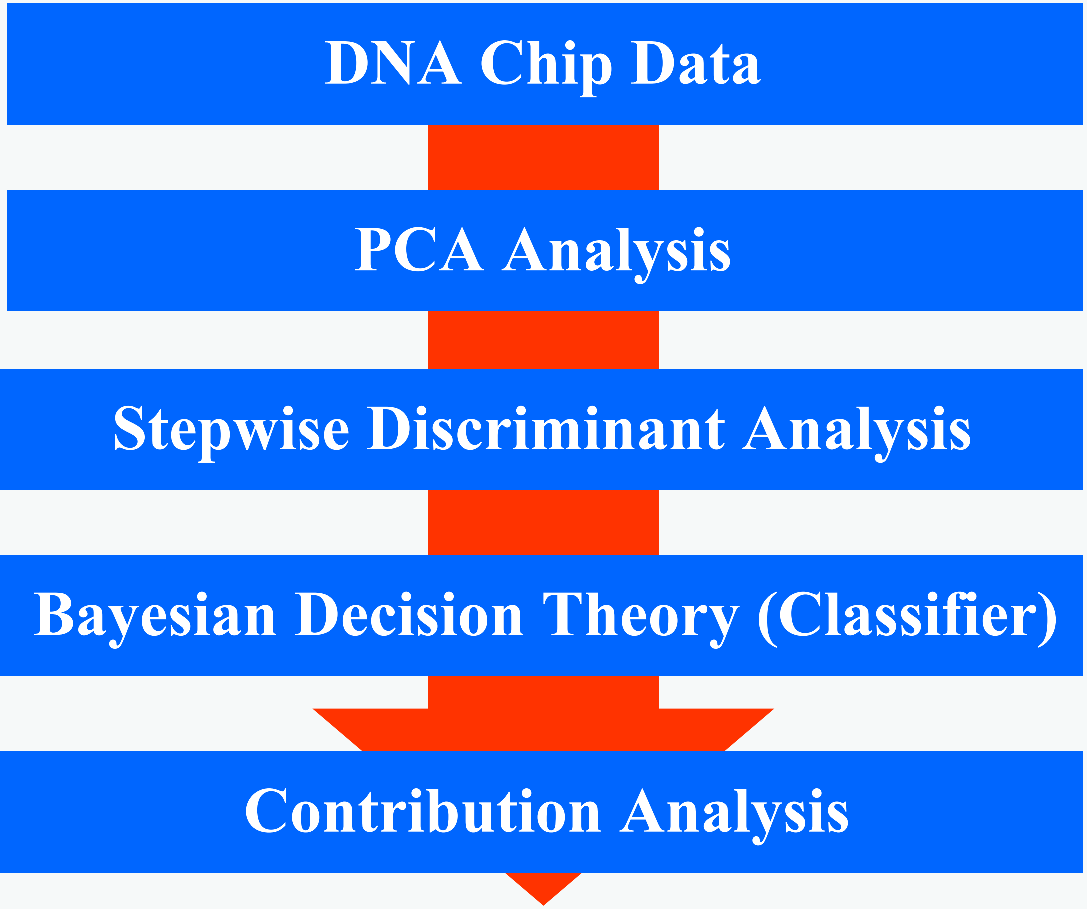
Limitations of Previous Approaches

- ✓ Small number of samples vs many variables
- ✓ Strong variable interaction
- ✓ Lack of interpretation based on biological meanings
- ✓ Limitation in the identification of marker genes due to the black box model
- ✓ Limitations due to univariate approaches
- ✓ Procedure of analysis are very complex and take a long time

Improvements

- ✓ Overcome interaction of many variables
- ✓ Develop to a method to select potential marker genes
- ✓ Develop multivariate approach
- ✓ Develop simple and ease procedure of data analysis

Proposed Procedure



DNA Chip Data

PCA Analysis

Stepwise Discriminant Analysis

Bayesian Decision Theory (Classifier)

Contribution Analysis

- High dimensional data
- Highly correlated variables

- Data preprocessing
  1. Dimension reduction
  2. Modeling of correlation structure

- Feature selection

(Select highly discriminant PC)

- Classification of tumor classes

- Select the potential marker genes

Major Steps

Stepwise Discriminant Analysis

- ✓ Select subset where Wilks' lambda value is minimum
- ✓ Maximize the discriminant power

$$\Lambda = \frac{SS_w}{SS_t}$$

SS_t : Class heterogeneity

SS_w : Class homogeneity

Contribution Analysis

- ✓ Discover potential marker genes to discriminate cancer classes

$$C_j = \sum_{n=1}^{k} w_n \times p_{n,j} \times \left(t_{i,n} - t_{r,n}\right)$$

C_j : the contribution of gene j

$p_{n,i}$: the loading of the j-th gene on the n-th PC

$t_{i,n}$: the average score of cancer class i

$t_{r,n}$: the average score of reference cancer class

w_n : the weight factor (eigenvalue of n-th PC)

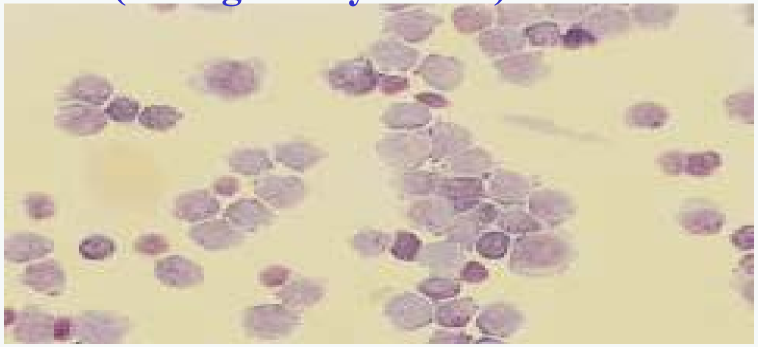
Case Study: Classification of Small Round Blue Cell Tumor

Cancer DNA Chip data

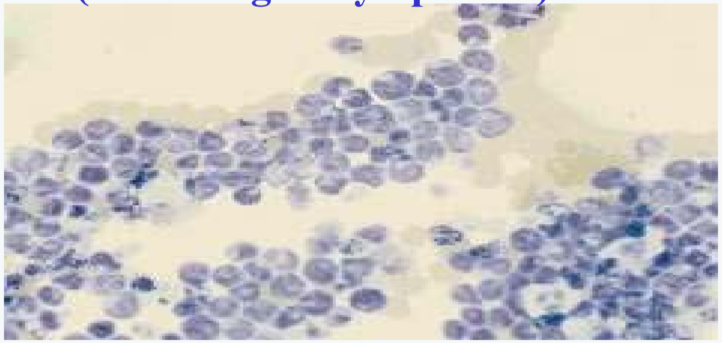
- ✓ Total samples : 88-by-2308 (samples-by-variables)
- ✓ Training samples (63), Testing samples (20), Noise samples (5)

Small round blue cell tumor

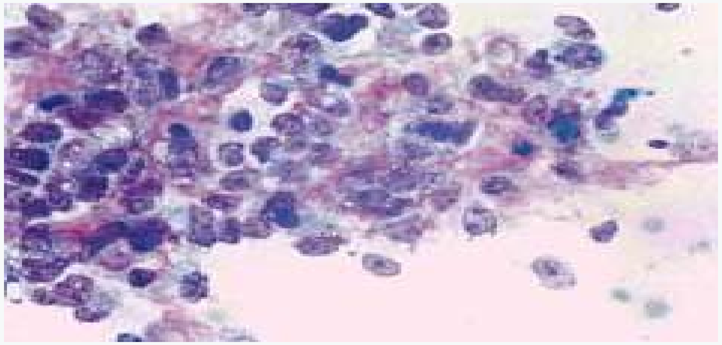
EWS (Ewing family tumor)



BL (non-Hodgkin lymphoma)



RMS (rhabdomyosarcoma)



NB (neuroblastoma)

Classification Results with SDA

Classification power 100%

Bayesian decision theory		Parametric method		Nonparametric method		
		Linear discriminant function	Quadratic discriminant function	K- nearest neighbor	Kernel density	
					normal	biweight
Without SDA	Cross-validation of training set	0.8359	0.75	0	0	0.45
	Classification of the test set	0.7833	0.75	0.7833	0.7833	0.3475
Using SDA	Cross-validation of training set	0.0238	0.5575	0	0	0.1051
	Classification of the test set	0	0.5417	0	0	0.1667

Training sample: 63, test sample: 25

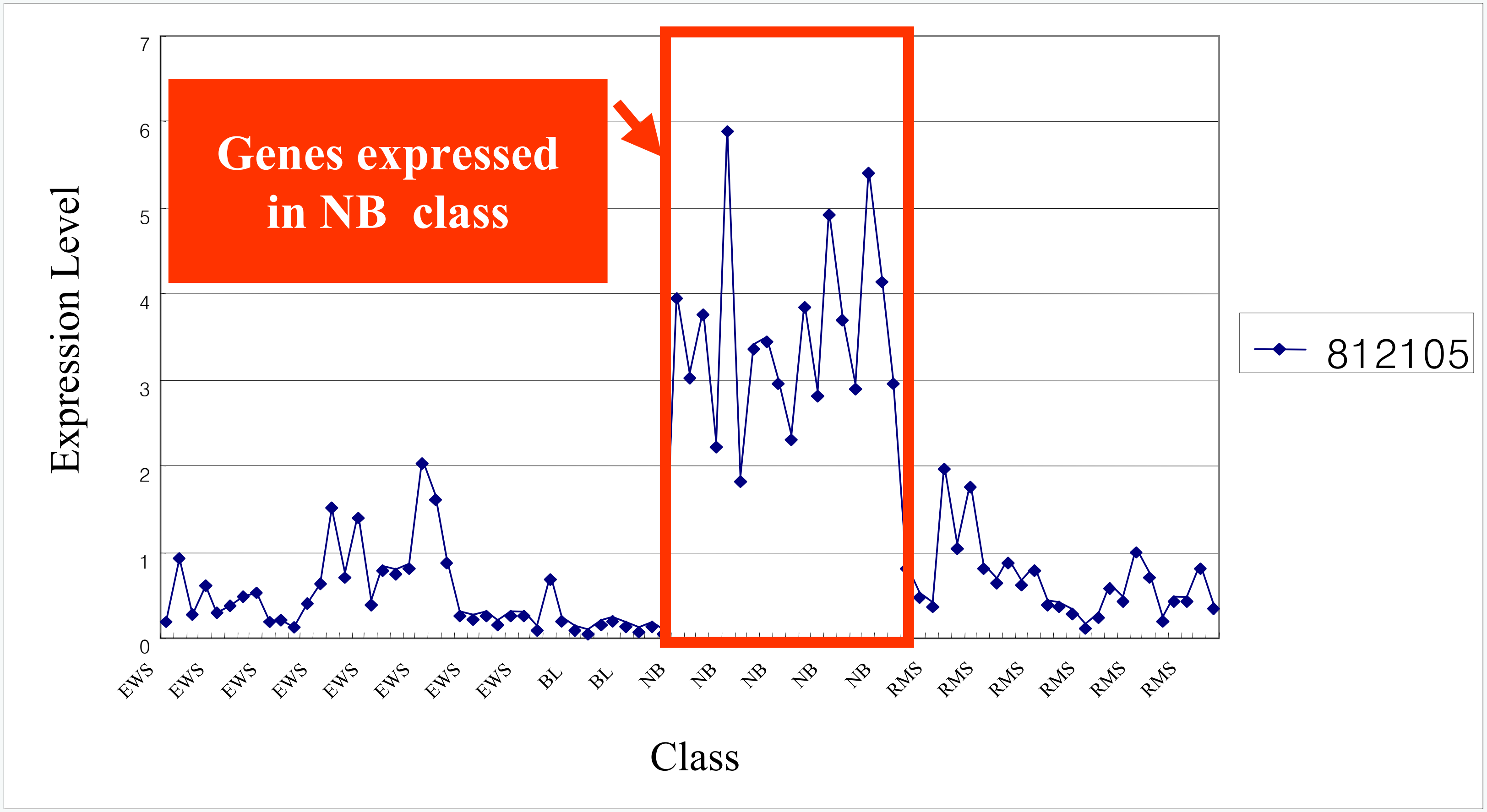
The number of identified potential marker genes

Class	Number of genes identified using the proposed method	Number of genes identified (Khan et al., 2001)	Number of matched genes	Number of mismatched genes	Image ID number
EWS	54	16	16	0	
BL	45	10	9	1	200814
NB	95	15	13	2	82225, 813266
RMS	68	20	14	6	788107,809901,122159 245330,246377,1409509
Not BL	61	12	8	4	45291, 204545 233721, 563673
Not EWS	12	1	1	0	
Overlap	24	0			
Total	311	74	61	13	

Khan et al. misjudgment 5 : image ID 82225, 813266, 233721, 245330, 122159
Redefine 2: image ID 45291, 563673
Overall trend agree 6 : image ID 204545, 788107, 1409509, 809901, 246377, 200814

The expression profile of potential marker gene (1)

Results are consistent with that of Khan et al., (2001)

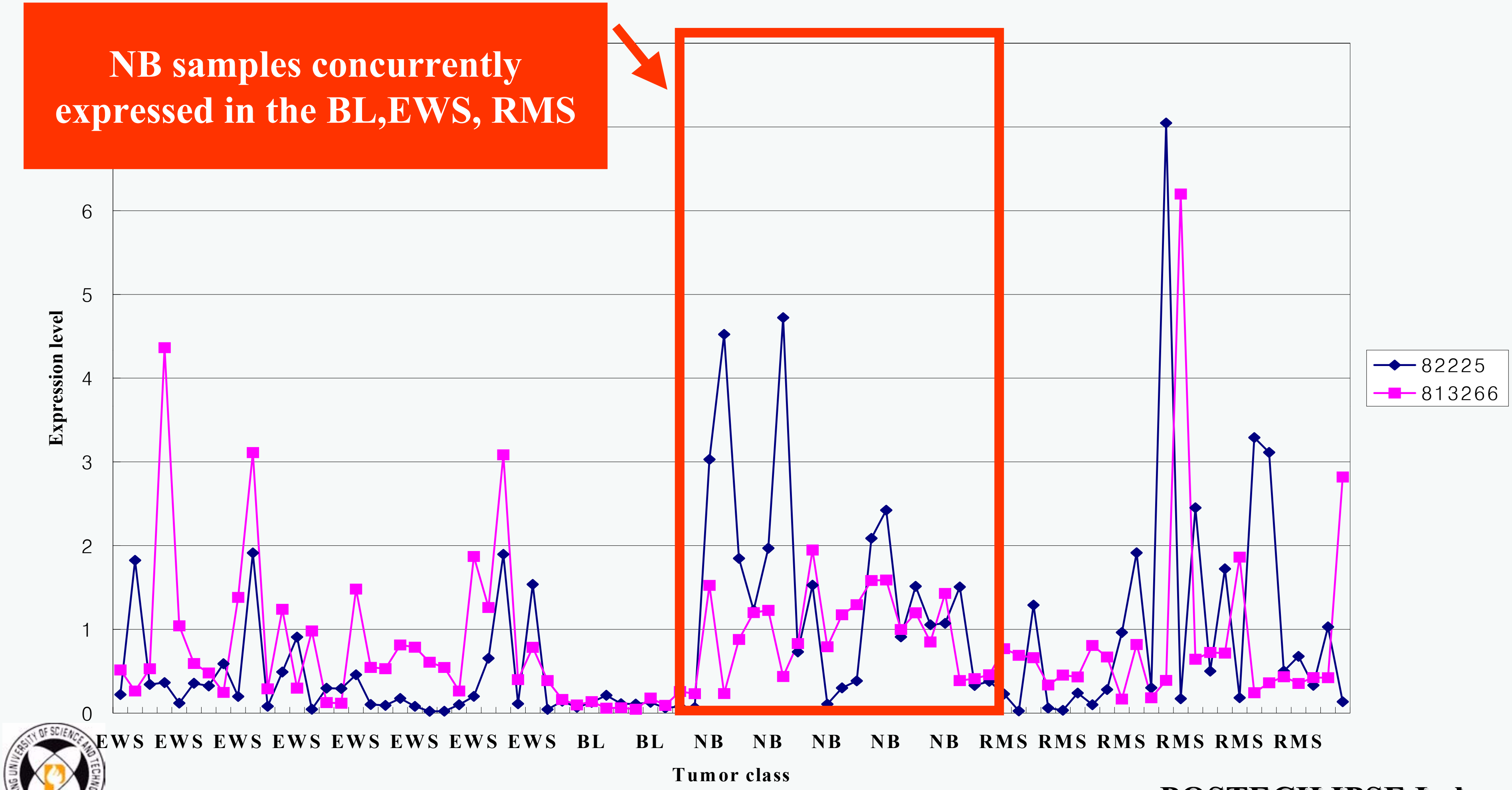


POSTECH IPSE Lab.

The expression profile of potential marker gene (2)

Not matched results

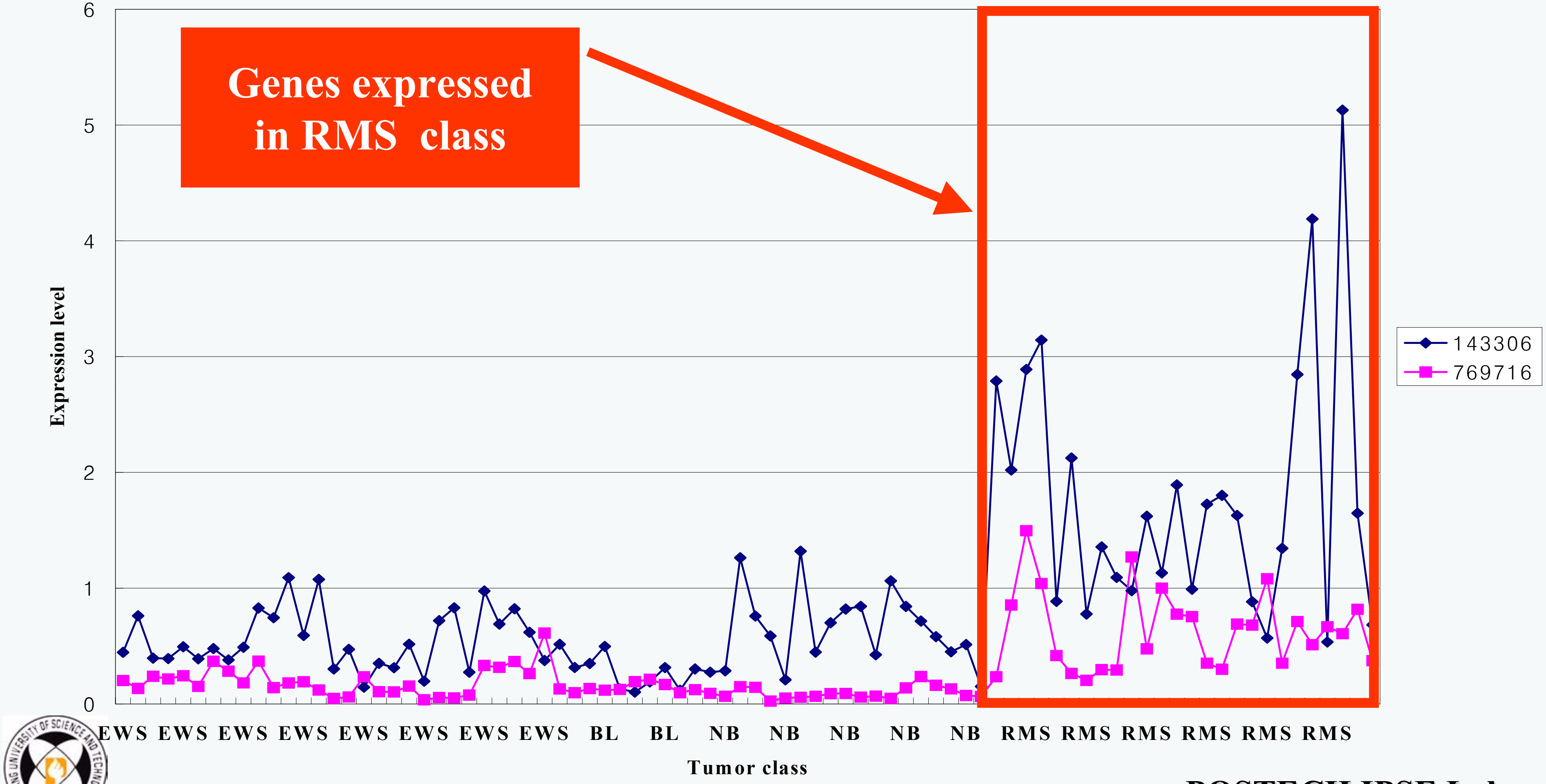
NB samples concurrently expressed in the BL,EWS, RMS



POSTECH IPSE Lab.

The expression profile of potential marker gene (3)

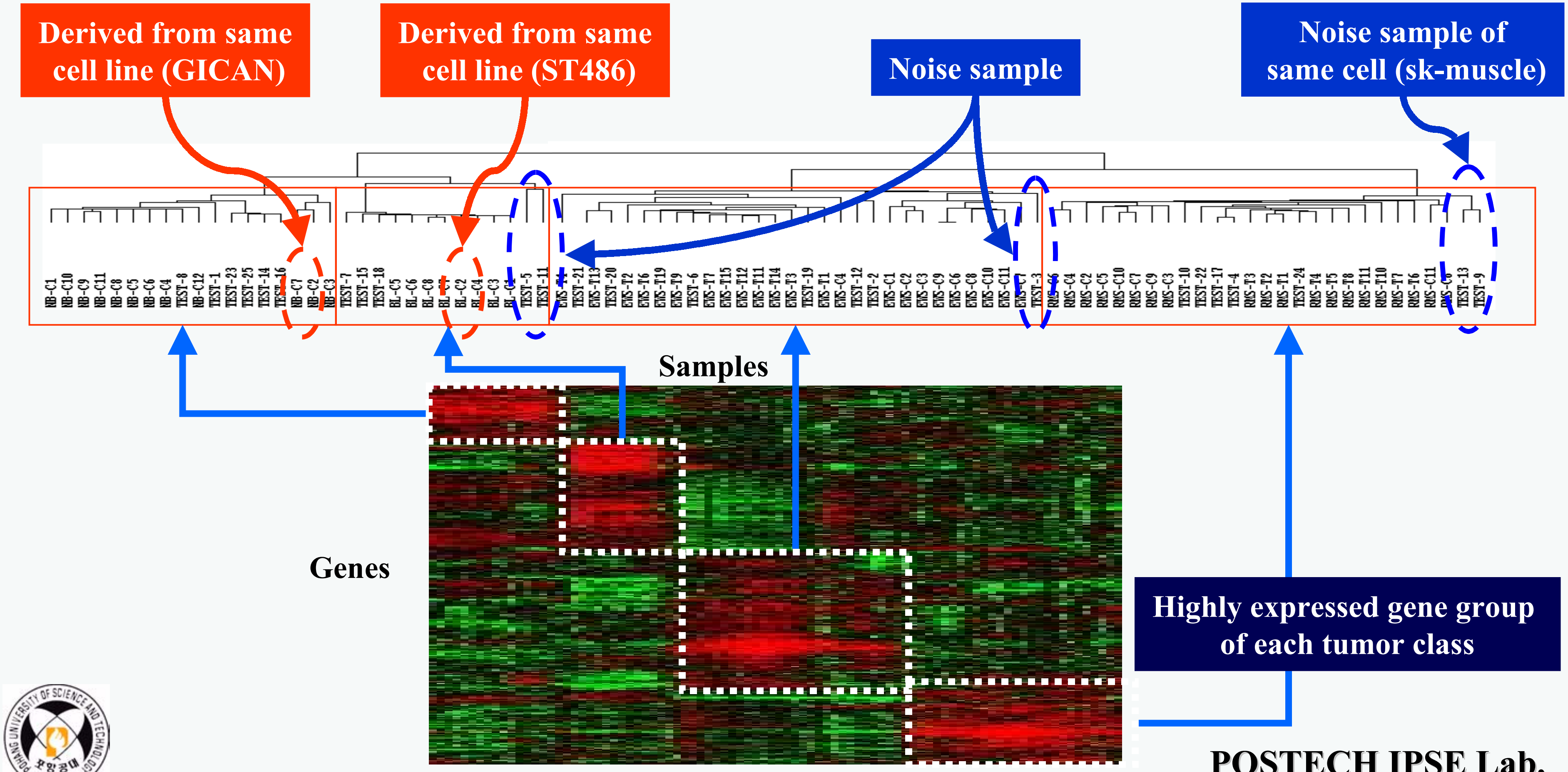
New discovered potential marker gene



POSTECH IPSE Lab.

Interpretation of the analysis results

- Hierarchal clustering based on 311 selected potential marker genes
- Correct classification for each class



Derived from same cell line (GICAN)

Derived from same cell line (ST486)

Noise sample

Noise sample of same cell (sk-muscle)

Highly expressed gene group of each tumor class

Interpretation of the analysis results: Biological validation

Marker genes for cancer classes

Gene Image ID	Cancer class	Biological gene function
1435862	EWS	antigen identified by monoclonal antibodies 12E7, F21 and O13
291756	EWS	tubulin, beta, 5
43733	EWS	glycogenin 2
52076	EWS	olfactomedinrelated ER localized protein
377731	EWS	glutathione S-transferase M5
784224	RMS	fibroblast growth factor receptor 4
470128	RMS	Myosin IC
296448	RMS	insulin-like growth factor 2 (somatomedin A)
207274	RMS	Human DNA for insulin-like growth factor II (IGF-2); exon 7 and additional ORF
461425	RMS	Myogenesis
377671	RMS	integrin, alpha 7
823886	RMS	Smooth muscle myosin heavy chain isoform SMemb [human, umbilical cord, fetal

Marker genes not matched for cancer classes

Gene Image ID	Chip data Cancer class	Normal Cancer class	Gene Image ID	Chip data Cancer class	Normal Cancer class
823886	Not BL	RMS	782488	All class	Not NB
897667	EWS	RMS	814773	EWS	NB
162208	BL	RMS	29054	NB	RMS
626502	BL	RMS	308231	NB	RMS
785793	BL	RMS	823886	NB	RMS
868304	BL	RMS	377048	NB	RMS
781018	BL	RMS			

Contributions

- Accurate multivariate classification method based on Bayesian method
- Potential marker gene selection method
- Simple and easy procedure for data analysis
- 250 new candidate marker genes discovered
- new hypothesis testing based on the candidate marker genes for drug discovery or cancer research

Biotechnology meets data mining

- Time to dance!!!
- **Contacts** between the established ‘data mining community’ and ‘bio/medical scientists’ seem to be **rare**
- There will be more dances, and new biotechnology will be forthcoming as we learn the steps

Coming dance !!!!

Questions ?

Contacts and full paper request: sw74@postech.ac.kr