Machine learning in high-throughput genomics and proteomics

(Part II)

Xuegong Zhang, Ph.D.
Professor of Pattern Recognition and Bioinformatics
Tsinghua University
zhangxg@tsinghua.edu.cn

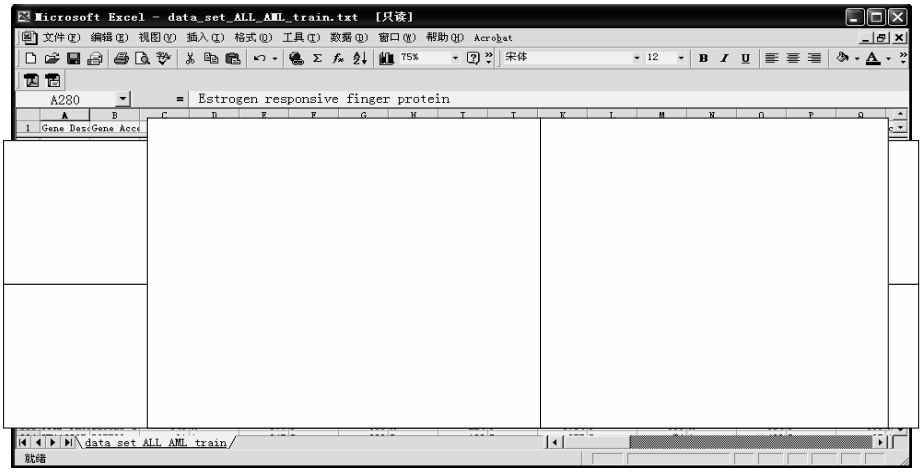
Outline

- Background
  - Microarray and mass-spectrometry technologies
  - Typical bioinformatics problems
  - Basic concepts of machine learning
- Microarray data mining with learning machines
  - Classification, clustering and gene selection
  - Problems, solutions, and possible pitfalls
- Further topics
  - Important considerations and open problems
  - Systems biology study after microarray data mining
  - New types of arrays

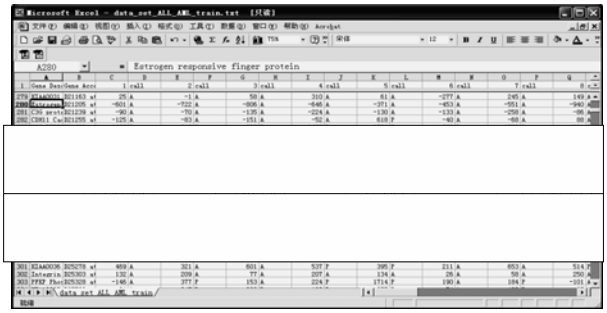
Part II
Microarray data mining with learning machines

- Clustering
- Classification
- Gene selection
- Dimension reduction

Classification study with multiple arrays



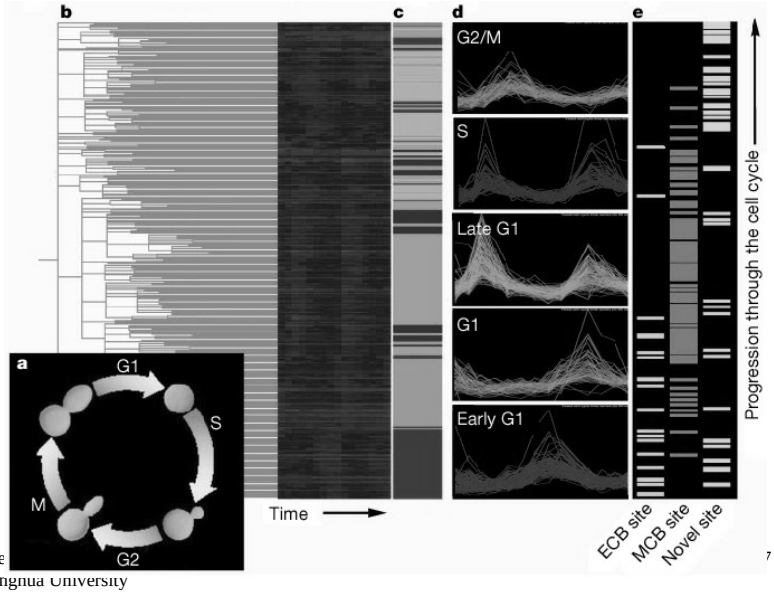
Studying the genes



Finding correlated genes

- Correlated → co-regulated
- Correlated in expression →
  Correlated in function
- Supervised learning
  - Annotating genes according to similar genes
- Unsupervised learning
  - Clustering genes with similar expression patterns
  - Building gene networks

Example: annotating gene functions according to expression patterns

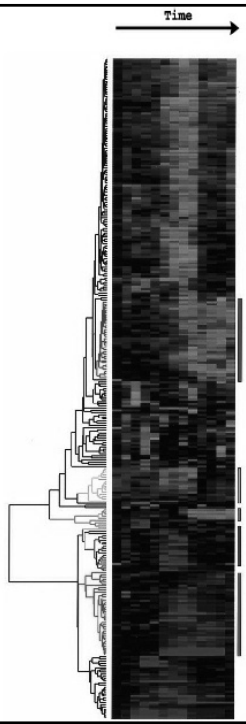


D.J. Lockhart & E.A. Winzeler, Genomics, gene expression and DNA arrays, *Nature*, 405:827-836, 2000

Xue
Tsinghua University

Example:

Inferring gene functions by clustering the genes by their expression profiles.



Clustered display of data from time course of serum stimulation of primary human fibroblasts. Foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15min, 30min, 1hr, 2hr, 3hr, 4hr, 8hr, 12hr, 16 hr, 20hr, 24hr. The final datapoint was from a separate unsynchronized sample. Data were measured by using a cDNA microarray with elements representing approximately 8600 distinct human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points.

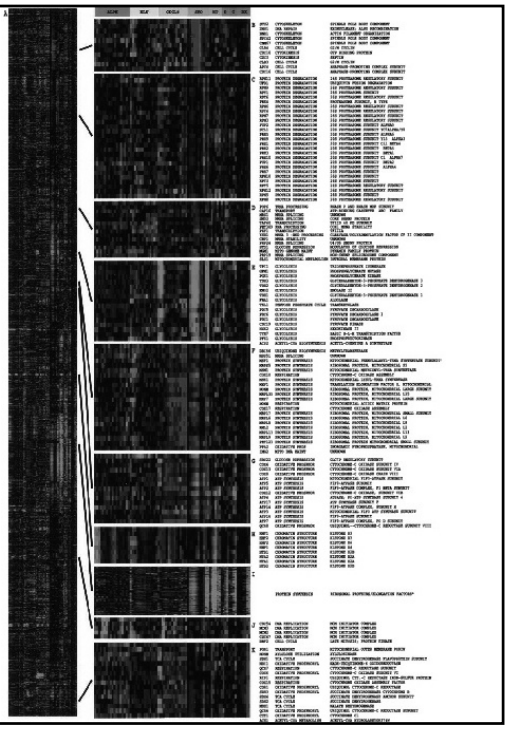
The color scale ranges from saturated green for log ratios –3.0 and below to saturated red for log ratios 3.0 and above. Five separate clusters are indicated by colored bars and by identical coloring of the corresponding region of the dendrogram. The sequence-verified named genes in these clusters contain multiple genes involved in (A) cholesterol biosynthesis, (B) the cell cycle, (C) the immediate-early response, (D) signaling and angiogenesis, and (E) wound healing and tissue remodeling. These clusters also contain named genes not involved in these processes and numerous uncharacterized genes.

M.B. Eisen, P.T. Spellman, P.O. Brown & D. Botstein, Cluster analysis and display of genome-wide expression patterns, *PNAS*, **95**: 14863-14868, 1998

Xuegong Zhang
Tsinghua University

Hierarchical clustering and the heat map:

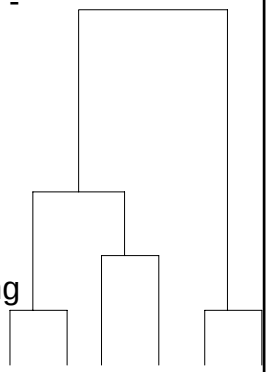
the importance of a good illustration



M.B. Eisen, P.T. Spellman, P.O. Brown & D. Botstein, Cluster analysis and display of genome-wide expression patterns, *PNAS*, **95**: 14863-14868, 1998

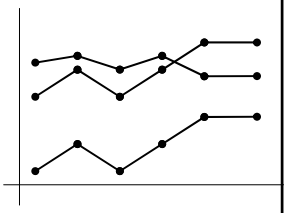
Hierarchical Clustering

- Bottom-up: start with every single sample - singleton clusters
- At each step, merge the two closest clusters into a new cluster
- At the last step, all samples are merged into one cluster
- Show the clustering procedure as a tree (the *dendrogram*), with nodes representing clusters and length of branches reflecting the similarity between merged clusters

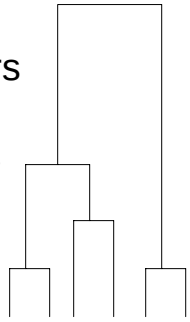


Sokal, RR and Michener, CD, A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409 - 38, 1958

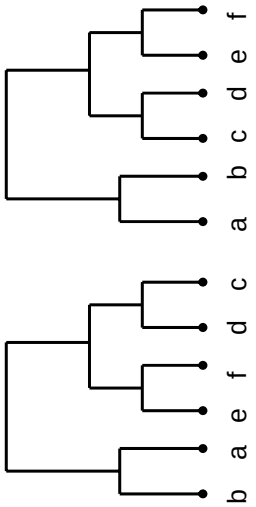
Hierarchical Clustering (cont)



- The distance (similarity) measure
  - Euclidean distance
  - Pearson's correlation coefficient
  - Editing distance (for sequences)
  - …
- Distance/similarity between two clusters
  - Single linkage
    - The distance between the two nearest samples
  - Complete linkage
    - The distance between the two farthest samples
  - Average linkage
    - The average distance between all samples



The orders in dendrograms

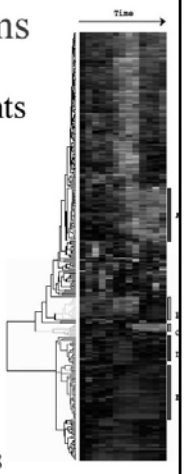


Ordered dendrograms

2^{n-1} linear orderings of n elements
(n= # genes or conditions)

Maximizing adjacent similarity is impractical. So order by:
- Average expression level,
- Time of max induction, or
- Chromosome positioning
- Closest uncle

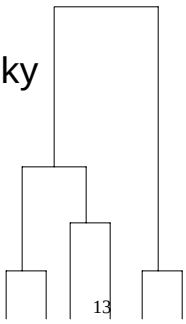
Wing H. Wong's Stat215 lecture, 2002



Eisen98

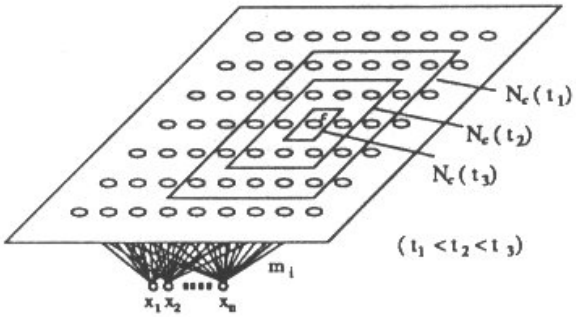
Discussion on Hierarchical Clustering

- Very good illustration
- Flexible, no need to pre-define cluster numbers
- Hierarchical cluster structure makes results more explainable
  - **But be careful not to over-explain it.**
- The distance metric/scaling could be tricky
  - A problem for all clustering methods!
- May not always be stable
  - Can be sensitive to certain single samples

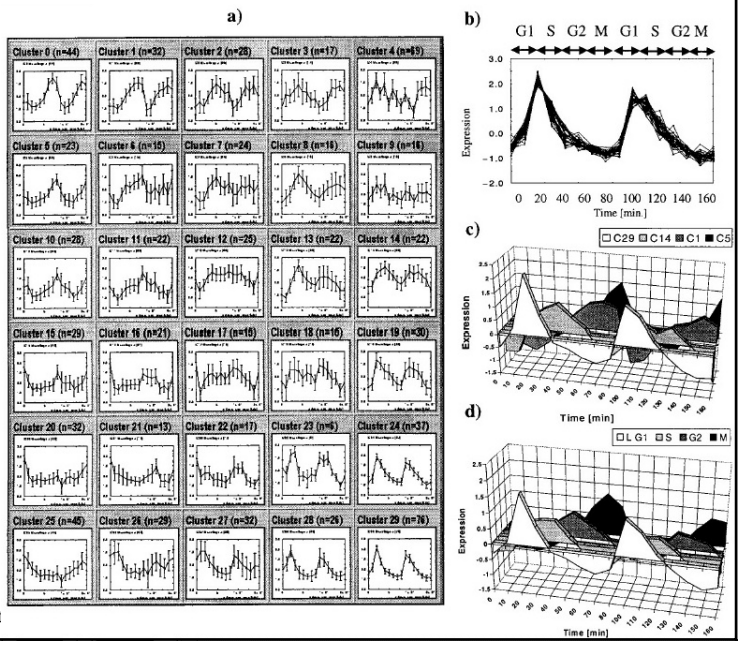


The self-organizing map (SOM)

(Kohonen, 1984)

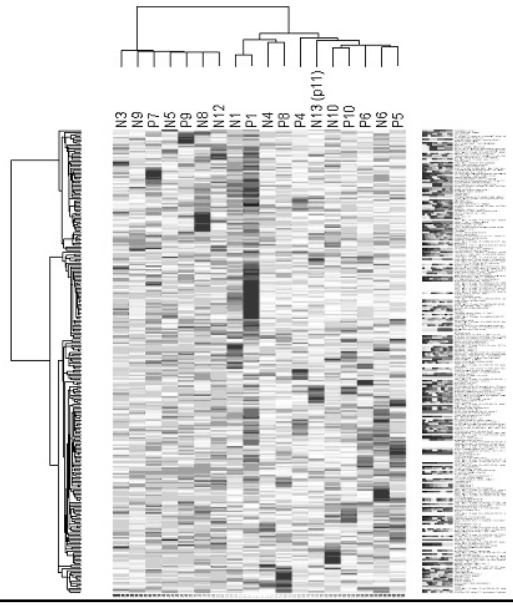


Example: Interpreting patterns of gene expression

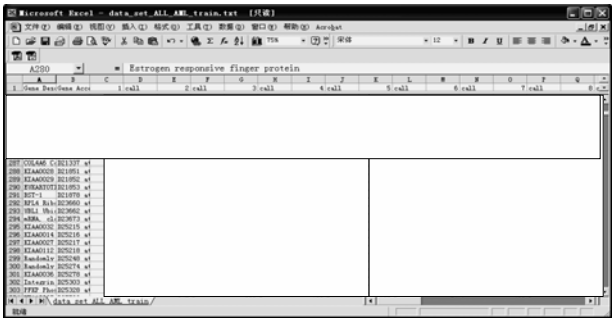


P. Tamayo, …, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *PNAS*, 96:2907-2912, 1999

Analyzing the genes AND the samples



Finding differentially expressed genes



Comparing Two Independent Samples
T-test (normal distribution assumption)

I	II
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

sample 1: $X_1,\cdots,X_n \quad X \sim N(\mu_X,\sigma^2)$

sample 2: $Y_1,\cdots,Y_m \quad Y \sim N(\mu_Y,\sigma^2)$

pooled sample variance $s_P^2 = \dfrac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$

the null hypothesis

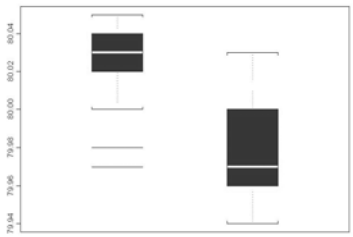
$H_0: \quad \mu_X = \mu_Y$

alternative hypotheses

two-sided $H_1: \quad \mu_X \neq \mu_Y$

one-sided $H_2: \quad \mu_X > \mu_Y$

$H_3: \quad \mu_X < \mu_Y$



t-statistic $t = \dfrac{\overline{X}-\overline{Y}}{s_{\overline{X}-\overline{Y}}} = \dfrac{\overline{X}-\overline{Y}}{s_P\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$

The *p-value*

A finding is described as statistically significant, when it can be demonstrated that the probability of obtaining such a finding by chance only, is relatively low.

P-Value: the probability that a variate would assume a value greater than or equal to the observed value strictly by chance.

---- false positive rate

Alpha Value: the acceptable level of the p-value

Multiple Comparison

- Family-wise Error Rate (FWER):
  - the probability of yielding one or more false positive in all simultaneous hypotheses.

  i.e. *FWER* = P(*#false-positive* >= 1)
- Bonferroni correction:
  - *m* hypotheses, each test have false positive rate less than a/m → FWER<a

	Cancer patients				Normal patients			p-value	Bonferroni*
1	1.16	0.99	1	0.84	-0.79	-0.8	-0.9	0.0000	expressed
2	0.75	0.91	0.96	0.68	-0.68	-0.73	-0.85	0.0000	expressed
3	-0.01	0.3	0.08	0.25	-0.63	-0.91	-0.98	0.0005	expressed
4	-0.09	-0.22	-0.12	0.14	-0.64	-1.21	-1.01	0.0032	
5	0.43	0.69	0.47	0.56	-0.24	-0.29	0.14	0.0038	
6	-0.41	-0.46	-0.73	-0.52	0.41	0.09	-0.07	0.0055	
7	0.36	0.26	0.12	0.34	-0.16	-0.67	-0.29	0.0064	
8	0.37	0.74	0.69	0.97	-0.37	-0.17	0.25	0.0137	
9	0.2	0.15	0.02	0.08	-0.05	-0.4	-0.47	0.0159	
10	-0.13	-0.31	-0.19	-0.15	-1.03	-0.41	-0.91	0.0166	
11	-0.43	-0.44	-0.25	-0.57	0.49	-0.02	-0.1	0.0262	
12	0.03	0.36	0.03	0.28	-0.06	-0.04	-0.17	0.0535	
:									
:									
:									
48	-0.15	0.5	-0.28	-0.04	0.16	-0.11	0.08	0.8734	
49	0.11	0.09	-0.3	0.77	0.05	0.03	0.54	0.8958	
50	-0.29	0.29	0.44	-0.15	0.21	0.21	-0.15	0.9328	

$$\text{Bonferroni*}: \alpha = 0.05, \text{m} = 50$$

X
Tsinghua University

False Discovery Rate (FDR) (Benjamini 1995)

- Measure the expectation of proportion of false positives out of total number of rejected null hypotheses

$$FDR = \frac{\#\text{ false positive}}{\#\text{ declared positive}}$$

FDR controlling procedure (Benjamini 1995):

1. Consider hypotheses $H_1, H_2, \ldots, H_m$ based on corresponding p - values $P_1, P_2, \ldots, P_m$.
2. Order p - values $P_{(1)} \le P_{(2)} \le \ldots \le P_{(m)}$. Let $H_{(i)}$ be the corresponding hypothesis to $P_{(i)}$.
3. Let k be the largest i for which $\text{P}_{(\text{i})} \le \frac{i}{m}\alpha^*$. Reject all $\text{H}_{(\text{i})}, i = 1, \ldots, k$.
4. The above procedures controls FDR at level α^*.

Comparison of FWER (Bonferroni method) and FDR (Benjamini)

	Cancer patients				Normal patients			p-value	Bonferroni*	FDR method**
1	1.16	0.99	1	0.84	-0.79	-0.8	-0.9	0.0000	expressed	0.0010
2	0.75	0.91	0.96	0.68	-0.68	-0.73	-0.85	0.0000	expressed	0.0020
3	-0.01	0.3	0.08	0.25	-0.63	-0.91	-0.98	0.0005	expressed	0.0030
4	-0.09	-0.22	-0.12	0.14	-0.64	-1.21	-1.01	0.0032		0.0040
5	0.43	0.69	0.47	0.56	-0.24	-0.29	0.14	0.0038		0.0050
6	-0.41	-0.46	-0.73	-0.52	0.41	0.09	-0.07	0.0055		0.0060
7	0.36	0.26	0.12	0.34	-0.16	-0.67	-0.29	0.0064		0.0070
8	0.37	0.74	0.69	0.97	-0.37	-0.17	0.25	0.0137		0.0080
9	0.2	0.15	0.02	0.08	-0.05	-0.4	-0.47	0.0159		0.0090
10	-0.13	-0.31	-0.19	-0.15	-1.03	-0.41	-0.91	0.0166		0.0100
11	-0.43	-0.44	-0.25	-0.57	0.49	-0.02	-0.1	0.0262		0.0110
12	0.03	0.36	0.03	0.28	-0.06	-0.04	-0.17	0.0535		0.0120
:										
:										
:										
48	-0.15	0.5	-0.28	-0.04	0.16	-0.11	0.08	0.8734		0.0480
49	0.11	0.09	-0.3	0.77	0.05	0.03	0.54	0.8958		0.0490
50	-0.29	0.29	0.44	-0.15	0.21	0.21	-0.15	0.9328		0.0500

Bonferroni* : $\alpha = 0.05$

Benjamini** : $\alpha^* = 0.05$

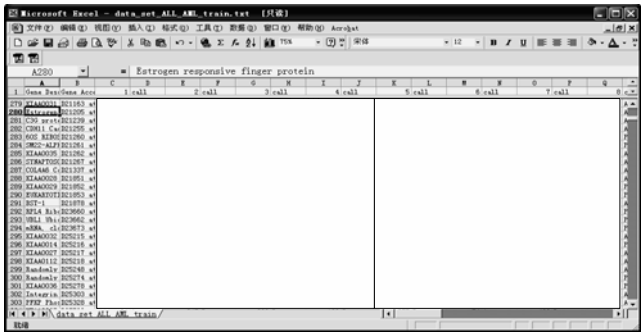
Three levels of significance

- P-value
  - For the hypothesis I tested, what is the probability I claim a wrong positive finding?
- Bonferroni corrected p-value
  - For a set of hypotheses I tested, what is the probability I claim a wrong positive finding?
- FDR
  - For a set of hypotheses I tested, what is the expected proportion of wrong claims among the positive findings I made?

Further reading:

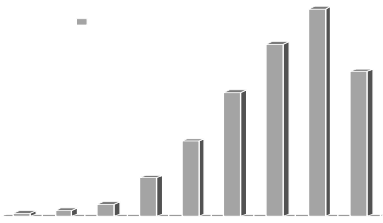
John D. Storey and Robert Tibshirani, Statistical significance for genomewide studies, *PNAS*, vol. 100, no. 16 9440-9450 Aug 5, 2003

Classification



Microarray-based cancer study: a very hot topic

- SCI query result with “microarray” and “cancer”:
  - 5,490 documents up to 09/16/2006



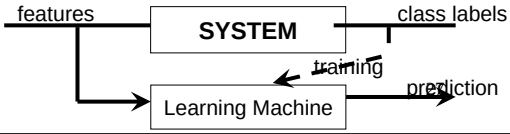
- T.R.Golub, …, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286: 531-537, 1999

  has been cited 2317 times by 09/16/2006 (according to SCI)

Classifying the samples

Gene \ Case	Case 1	Case 2	...	Case 9	...
gene a	123.4	234.5		56.3	
gene b	1234.5	5678.9		78.9	
gene c	765.4	43.2		3456.1	
gene d	211.0	985.0		12.3	
gene e	432.9	543.4		102.7	
...					

… and finding the “guilty” genes



Xuegong Zhang
Tsinghua University

The famous ALL/AML Dataset

- ALL - acute lymphoblastic leukemia
- AML - acute myeloid leukemia

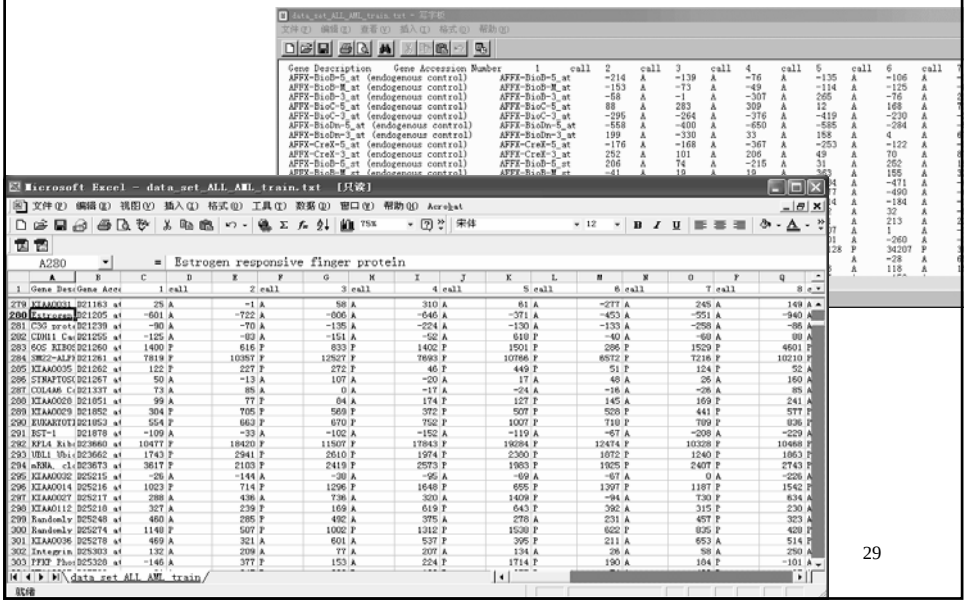
- The data: the expression levels of 6,817 genes (by Affymetrix Hu6800)
- 38 training samples: 27 ALL, 11 AML
- 34 test samples: 20 ALL, 14 AML

Xuegong Zhang
Tsinghua University

28

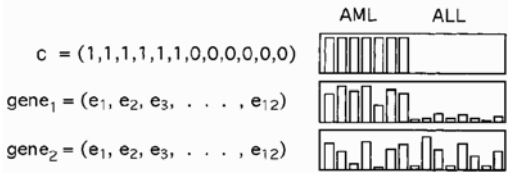
Feel of the Dataset

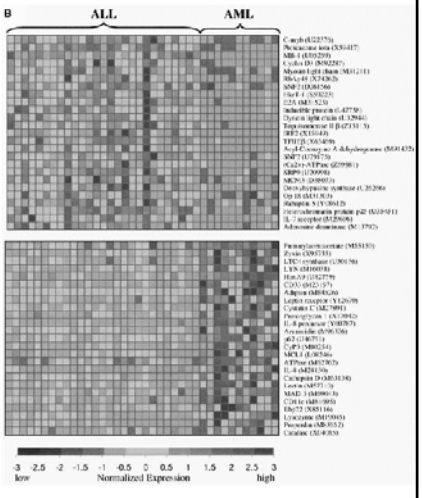
http://www.genome.wi.mit.edu/MPR



29

The "weighted voting" approach





$$P(g,c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$$

weighted vote V
= weight(g)•distance(x,b)

$$PS = \frac{V_{winner} - V_{loser}}{V_{winner} + V_{loser}}$$

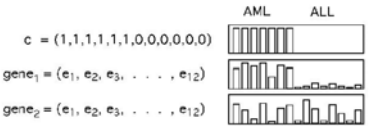
Gene Selection: Filtering vs. Wrapper

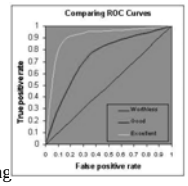
Unsupervised Gene Filtering

- e.g.
  - Fold-change: exclude those with little variation
  - Expression value: exclude those whose expression values are too low (under noise level) or too high (possibly outlier)
  - Present/Absent (AFFY chips): exclude those with too many absent calls according to Affy
  - Unreliable genes: exclude those that was detected as outlier genes in the pre-processing
  - …

Gene Selection with ***Filtering Methods***

- According to criterion of the single genes to the discrimination, e.g.
  - Correlation with classes
    e.g. the WI's SNR:
  - T-test $z_i = \frac{\bar{x}_i^+ - \bar{x}_i^-}{s_i + s_0}$
  - Fisher criterion $F_i = \frac{\left(\bar{x}_i^+ - \bar{x}_i^-\right)^2}{\sum_{class+}(x_i - \bar{x}_i^+)^2 + \sum_{class-}(x_i - \bar{x}_i^-)^2}$
  - ROC





Note:
1. Multi-gene filtering criteria are also possible, but not many reported applications.
2. The filtering criteria are not related with the follow-up classification.

Gene Selection with ***Wrapper Methods***

- Start from all genes
- Use all candidate genes to build certain classifier, e.g. SVM, NN, and use the classification performance to evaluate the genes
- Evaluate the contribution of every gene in the classifier
  - Linear classifiers: according to weights or according to the separation of certain samples (e.g. class centers) – weighted mean difference
  - Nonlinear classifiers: According to sensitivity of each gene to the output (e.g. partial derivative of the NN/SVM output with respect to the gene expression)
- Recursive selection: select a new subset of candidate genes

General schemes of gene selection

- Two-Step Procedure:
  - Gene selection (with certain stand-alone methods)
  - Classification (using the selected genes)

- Recursive Procedure
  - Classification (with all genes)
  - Gene selection (according to classification)
  - Classification (using the selected genes)

- Hybrid Procedure:
  - e.g. certain initial gene selection before the recursive procedure

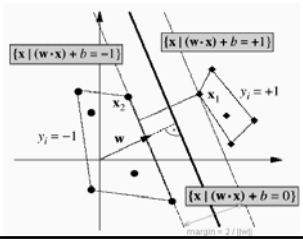
R-SVM and SVM-RFE

Further reading:

Xuegong Zhang, Xin Lu, Qian Shi, Xiu-qin Xu, Hon-chiu E Leung, Lyndsay N Harris, James D Iglehart, Alexander Miron, Jun S Liu and Wing H Wong, Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data, *BMC Bioinformatics*, **7**:197, 2006 (10Apr2006)

Choosing Linear SVM for the task

- Good generalization esp. for very sparse samples in high-dimensional spaces
- Linear SVM
  - Least complexity, proper for the very small sample size



R-SVM: Recursive SVM for classification and gene selection [Zhang & Wong, 2001]

- SVM training
- Select a subset of genes that gives the *best performance*:
- Loop (select a subset → redo SVM training and gene ranking)

R-SVM: Recursive SVM for classification and gene selection [Zhang & Wong, 2001]

- SVM training
- Select a subset of genes that gives the *best performance*:
  - Minimal error ---- turns to be always zero on training set
  - Maximal separation:

$$S = \frac{1}{n_1} \sum_{\mathbf{x}^+ \in class1} f(\mathbf{x}^+) - \frac{1}{n_2} \sum_{\mathbf{x}^- \in class2} f(\mathbf{x}^-)$$

$$S = \sum_{i=1}^{d} w_i m_i^+ - \sum_{i=1}^{d} w_i m_i^- = \sum_{i=1}^{d} w_i (m_i^+ - m_i^-)$$

$$s_i = w_i (m_i^+ - m_i^-)$$

  - Ranking the genes according to their *s*
- Loop (select a subset → redo SVM training and gene ranking)

SVM-RFE: SVM - Recursive Feature Elimination

I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning*, **46**: 389-422, 2002

- SVM training
- Select a subset of genes that gives the *best performance*:
  - Genes' contribution:

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$$

$$s_i^{SVM-RFE} = w_i^2$$

$$S_{SVM-RFE} = \mathbf{w} \cdot \left(\sum_{\mathbf{x}_j^+ : SVs\ in\ class1} \alpha_j \mathbf{x}_j^+ - \sum_{\mathbf{x}_j^- : SVs\ in\ class2} \alpha_j \mathbf{x}_j^- \right) = \mathbf{w} \cdot \mathbf{w}$$

  - Ranking the genes according to their *s*
- Loop (select a subset → redo SVM training and gene ranking)

R-SVM vs. SVM-RFE

- Same recursive selection procedure
- Different philosophy in ranking criteria
  - SVM-RFE:

$$S_{SVM-RFE} = \mathbf{w} \cdot \left(\sum_{\mathbf{x}_j^+ : SVs\ in\ class1} \alpha_j \mathbf{x}_j^+ - \sum_{\mathbf{x}_j^- : SVs\ in\ class2} \alpha_j \mathbf{x}_j^- \right) = \mathbf{w} \cdot \mathbf{w}$$

$$s_i^{SVM-RFE} = w_i (r_i^+ - r_i^-) = w_i^2$$

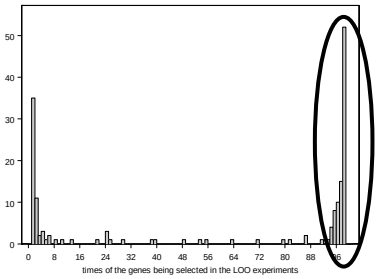
  - R-SVM:

$$S_{R-SVM} = \mathbf{w} \cdot \left(\frac{1}{n_i} \sum_{\mathbf{x}^+ \in class1} \mathbf{x}^+ - \frac{1}{n_2} \sum_{\mathbf{x}^- \in class2} \mathbf{x}^- \right) = \mathbf{w} \cdot (\mathbf{m}^+ - \mathbf{m}^-)$$

$$s_i^{R-SVM} = w_i (m_i^+ - m_i^-)$$

- Experiments show that R-SVM is more robust to noise.

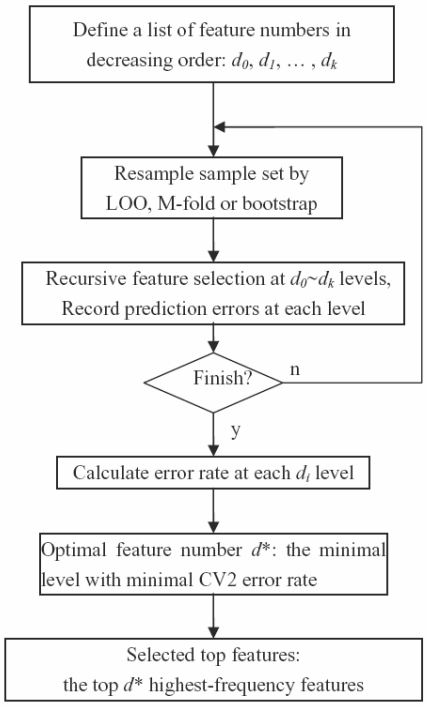
Gene selection with R-SVM



- Vote for the most frequently selected genes in the LOOCV experiments

- Previous simpler strategy:
  - LOOCV to find the gene selection level where minimal CV error is obtained
  - Redo R-SVM on the whole data to select that number of genes

The work flow of R-SVM



Comparison on Simulated Data-G (with gene outliers)

Table 1. Comparison of R-SVM and SVM-RFE on Data-G (with gene outliers)

Level[a]	ReduceSV[b]	P(sv-diff)[c]	ReduceTest[d]	P(test-diff)[e]	ImproveRec[f]	P(rec-diff)[g]
800	4.01%	1.81E-42	-7.70%	4.72E-03	-3.90%	1.71E-39
600	5.77%	1.74E-49	-2.50%	4.64E-01	-1.70%	5.21E-15
500	6.83%	2.75E-51	-4.00%	1.62E-01	-0.30%	0.079189
400	8.35%	3.26E-60	2.80%	3.48E-01	1.10%	4.48E-06
300	9.33%	3.83E-58	7.40%	3.65E-02	3.70%	1.77E-31
200	8.22%	1.28E-48	19.20%	6.36E-09	6.30%	5.79E-44
150	8.55%	1.51E-53	19.50%	1.16E-08	7.10%	9.76E-46
100	4.97%	6.20E-22	11.90%	1.83E-04	6.00%	6.43E-40
90	5.84%	1.66E-27	13.70%	4.20E-06	4.60%	1.07E-30
80	5.17%	8.20E-29	12.40%	4.14E-06	4.50%	7.12E-29
70	4.14%	1.46E-27	8.50%	4.77E-04	3.80%	1.05E-24
60	3.10%	1.23E-20	10.20%	3.14E-05	3.40%	4.99E-24
50	2.27%	2.01E-15	10.20%	4.11E-06	2.90%	2.37E-21

Comparison on Simulated Data-S (with sample outliers)

Table 2. Comparison of R-SVM and SVM-RFE on Data-S (with sample outliers)

Level[a]	ReduceSV[b]	P(sv-diff)[c]	ReduceTest[d]	P(test-diff)[e]	ImproveRec[f]	P(rec-diff)[g]	ReduceOSV[h]	P(osv-diff)[i]
800	3.25%	4.49E-41	-65.19%	5.65E-36	-10.14%	3.36E-75	50.37%	5.97E-35
600	5.80%	1.90E-57	-70.27%	3.04E-35	-7.14%	5.18E-56	72.28%	1.10E-49
500	7.02%	8.20E-63	-59.63%	1.81E-37	-5.13%	3.37E-39	80.54%	1.17E-56
400	8.26%	1.68E-67	-41.43%	8.31E-25	-2.57%	4.53E-12	89.04%	2.51E-64
300	7.72%	1.20E-58	-19.14%	2.18E-13	0.75%	4.92E-02	93.44%	7.46E-65
200	7.21%	4.54E-51	-6.53%	2.56E-04	4.00%	7.15E-16	93.91%	1.47E-61
150	9.13%	1.29E-71	2.63%	1.20E-01	6.47%	8.41E-23	93.59%	6.27E-61
100	8.30%	1.42E-64	5.56%	8.04E-04	7.69%	3.50E-22	92.44%	1.33E-61
90	8.36%	2.01E-72	4.31%	1.15E-02	6.99%	8.74E-19	91.37%	2.60E-61
80	8.01%	6.63E-71	4.45%	1.99E-02	6.99%	9.33E-18	90.26%	2.65E-60
70	7.17%	1.29E-67	6.59%	3.78E-04	7.52%	2.80E-16	88.56%	7.55E-62
60	6.67%	2.65E-65	6.16%	2.32E-03	7.27%	5.72E-13	86.38%	2.60E-62
50	5.82%	1.08E-58	7.70%	1.34E-04	7.42%	3.71E-12	83.82%	1.23E-61

Comparison on Simulated Data-R (generated from real data)

Table 3. Comparison of R-SVM and SVM-RFE on Data-R

Level[a]	ReduceSV[b]	P(sv-diff)[c]	ReduceTest[d]	P(test-diff)[e]	ImproveRec[f]	P(rec-diff)[g]
800	15.35%	1.24E-53	-3.59%	1.26E-05	-3.60%	1.50E-23
600	18.65%	3.14E-56	-7.06%	4.09E-04	2.69%	2.20E-09
500	19.58%	7.71E-58	-6.46%	1.79E-03	9.18%	1.24E-37
400	21.07%	1.80E-63	-2.74%	3.22E-05	17.32%	4.25E-59
300	22.51%	5.12E-67	-4.64%	1.26E-05	24.14%	5.43E-65
200	22.16%	9.38E-68	-0.93%	1.83E-04	30.64%	2.25E-71
150	21.78%	4.57E-64	-3.44%	8.74E-04	29.14%	5.86E-71
100	21.01%	3.21E-57	0.31%	3.22E-05	29.95%	7.74E-69
90	22.57%	1.88E-60	-2.52%	3.52E-03	27.51%	9.74E-66
80	22.88%	1.67E-65	1.84%	7.85E-05	27.92%	4.03E-62
70	21.42%	2.96E-59	0.59%	4.09E-04	27.16%	1.15E-58
60	20.20%	1.64E-55	6.16%	1.83E-04	26.83%	2.55E-60
50	18.67%	4.40E-52	4.23%	8.74E-04	25.89%	9.63E-53
40	15.37%	5.66E-46	8.99%	4.69E-06	25.39%	1.09E-55
30	11.85%	6.90E-33	9.61%	1.67E-06	24.19%	2.07E-45
20	7.87%	2.19E-18	11.43%	3.22E-05	20.86%	1.09E-34

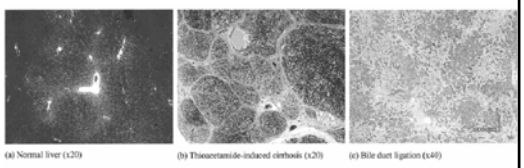
Example: R-SVM to find proteomics markers for liver cirrhosis

Proteomics 2004, *4*, 3235–3245 DOI 10.1002/pmic.200400839

Molecular classification of liver cirrhosis in a rat model by proteomics and bioinformatics

Xiu-Qin Xu[1], Chon K. Leow[1, 2], Xin Lu[3], Xuegong Zhang[4], Jun S. Liu[3], Wing-Hung Wong[3], Arndt Asperger[5], Sören Deininger[5] and Hon-chiu Eastwood Leung[1]

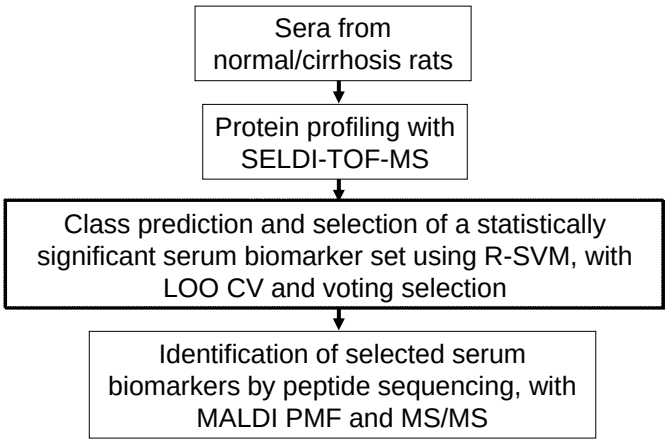
- Diagnosis of liver cirrhosis
  - Biopsy: invasive, potential risk of internal bleeding
  - CT scanning: not able to detect early cirrhosis accurately
  - At present, there are no sensitive and specific serum or plasma markers available
  - cDNA microarray: need liver tissue by an invasive procedure
  - 2DE: not good for hydrophobic proteins, low abundant proteins and low molecular weight proteins
  - SELDI-TOF-MS: good resolution, surfaces for different proteins
- Material:
  - Normal rat (n=8)
  - Liver cirrhosis rat (n=22)
  - Liver fibrosis rat (n=5)



(a) Normal liver (x20) (b) Thioacetamide-induced cirrhosis (x20) (c) Bile duct ligation (x40)

Figure 1. Liver sections stained with Masson Trichrome. (A) normal liver, (B) cirrhotic liver, and (C) liver after bile duct ligation. The fold magnification in 1a) and 1b) was × 20; while 1c) was × 40.

The Method



3 ways of doing R-SVM on the SELDI data

- Data range: 1-10kDa
- Biomarker Wizard: R-SVM on *biomarkers* detected with Ciphergen's software
  - 78 biomarkers → 6 important markers
    - 1743.12, 3515.68, 3537.26, 4186.07, 4902.63, 8201.04 Da
- Point-to-Point: R-SVM on all 4607 points resampled from 1-10kDa
  - 7 important regions selected, covering all the previous 6 markers, but centering at different points
    - 1744.56, 3513.31, 3515.07, 3518.60, 3520.36, 4187.13, 8209.99 Da
- Sliding Window:
  - Scanning with a sliding window to pre-select some candidate regions (21 with MeanDist (Intra/Inter)<0.75), then select 6 markers by R-SVM:
    - 1743.12, 1787.89, 3515.68, 3537.26, 6207.55, 8201.04 Da

Table 1. Overall error rates of different statistical biomarkers selection approaches

Statisical Method	CV2 (external CV) errors		CV1 (internal CV) errors	
	False positive (Type 1) count	False negative (Type 2) count	False positive (Type 1) count	False negative (Type 2) count
Point-to-point RSVM	2	2	1	1
Sliding window selection	2	0	2	0
Biomarker Wizard RSVM	2	2	1	1
Type Specific Error Rate	7.7%	0 to 2.9%	3.8 to 7.7%	0 to 1.4%
Overall Error Rate	2.1 to 4.2%		2.1%	
Overall sensitivity	97.1 to 100%			
Overall specificity	92.3%			

Xuegong Zhang
Tsinghua University

The 3495 Da protein

- The 1743, 3515 and 3537 Da peaks were selected as important markers by 3 tests, and they are mostly dicharged peptide or sodium adducts of the 3495 Da peak.
- Taken together, … the 3495 Da peak was a fragment of some unknown histidine-rich glycoprotein.

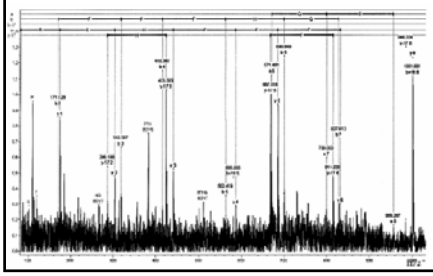


Figure 7. Peptide sequencing of the 1001 Da peak using Ultraflex MALDI TOF/TOF mass spectrometer. X-axis is the *m/z* value; the y-axis is the signal intensity.

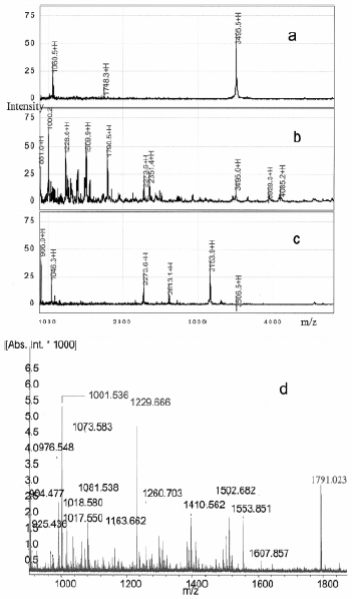


Figure 6. On-chip digestion of the purified 3495 Da protein. (A) Mass spectrum before trypsin digestion on WCX chip. (B) Mass spectrum after 3 h of on-chip trypsin digestion at 37°C. (C) Mass spectrum of trypsin alone after 3 h of on-chip digestion on WCX chip. PMF of tryptic digest after transferring to an AnchorChip and analyzed using an Ultraflex MALDI TOF/TOF mass spectrometer.

The R-SVM LOOCV feature selection and classification results on the rat cirrhosis data

Table 4. The CV results on the rat cirrhosis data

Level [a]	R-SVM		SVM-RFE	
	CV2 [b]	AveSV [c]	CV2 [b]	AveSV [c]
93	4.2%	14.75	4.2%	14.75
80	4.2%	11.91	4.2%	14.74
70	4.2%	9.95	4.2%	14.73
60	3.2%	9.22	4.2%	13.91
50	3.2%	9.03	4.2%	13.82
40	3.2%	9.02	4.2%	14.65
30	3.2%	8.95	4.2%	13.65
20	3.2%	8.93	4.2%	9.98
18	4.2%	8.14	4.2%	9.97
16	4.2%	8.08	3.2%	7.26
15	4.2%	7.60	3.2%	7.15
14	4.2%	7.54	3.2%	7.94
13	6.3%	7.58	4.2%	7.98
12	6.3%	7.41	4.2%	8.05
11	6.3%	7.65	4.2%	8.02
10	6.3%	7.64	3.2%	9.83
9	5.3%	6.50	3.2%	8.83
8	4.2%	5.97	4.2%	7.01
7	4.2%	6.73	4.2%	6.05
6	4.2%	5.98	3.2%	5.97
5	5.3%	5.94	4.2%	5.05

51

Example: R-SVM to find proteomics markers for breast cancer

[Q. Shi et al, 2005]

Table 6. The CV results on the human breast cancer dataset

Level [a]	R-SVM		SVM-RFE	
	CV2 [b]	MeanSV [c]	CV2 [b]	MeanSV [c]
98	28.7%	54.65	28.70%	54.65
88	27.9%	50.10	29.40%	55.25
79	29.4%	49.28	30.10%	52.21
71	29.4%	47.48	30.90%	50.88
63	27.9%	44.65	27.90%	48.42
56	27.2%	42.50	27.90%	46.02
50	27.9%	40.04	26.50%	40.13
45	25.7%	38.65	26.50%	40.25
40	24.3%	37.04	27.90%	34.88
36	23.5%	35.16	27.90%	34.51
32	22.1%	33.26	27.90%	30.75
28	22.8%	32.04	27.20%	27.77
25	22.1%	31.24	30.90%	24.61
22	22.1%	31.15	34.60%	23.93
19	22.8%	32.10	30.10%	26.79
17	25.7%	33.26	29.40%	31.28
15	23.5%	35.68	25.70%	35.10
13	19.9%	37.40	26.50%	42.15
11	22.1%	37.83	25.00%	46.03
9	21.3%	42.01	24.30%	50.18
8	17.6%	44.07	22.10%	49.93
7	23.5%	50.29	20.60%	51.43
6	22.1%	54.73	20.60%	52.39
5	22.1%	57.98	20.60%	52.18
4	22.8%	59.75	25.00%	58.92
3	27.2%	78.90	32.40%	77.46

Xuegong Zhang
Tsinghua University

Observations

- R-SVM vs. SVM-RFE:
  - R-SVM and SVM-RFE are close in terms of CV performances
  - R-SVM is more robust to noise and outliers
  - R-SVM can recover more informative genes
  - R-SVM has better generalization ability
- Comparing with univariate methods (WV):
  - R-SVM performs better in terms of CV classification accuracy
  - WV reveal more of the differentially expressed genes

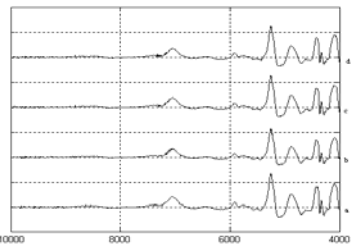
Example:

Discrimination and feature selection of geographic origins of traditional Chinese medicine herbs with NIR spectroscopy

S. Liu, X. Zhang, S. Sun, *Chinese Science Bulletin*, **50**(2): 179-184, 2005

Background:

- The efficiency of some traditional herbal medicines depends on the geographic origin and the growth condition of the herbs.
- Herbs are mixtures of many unknown compounds
- Infrared spectrometry is a key technique in identifying medical compounds, but success have not been widely reported on TCM.



The NIR derivative spectrums of Baizhi from different origins (a. Henan, b. Hebei, c. Sichuan, d. Zhejiang）

Data and Method

Table 1 The geographic origins of Baizhi samples

Origin	Henan	Hebei	Sichuan	Zhejiang	Total
Number of samples in Set-A	64	60	75	70	269
Number of samples in Set-B	20	18	24	22	84
Total	84	78	99	92	353

Table 2 The geographic origins and growth conditions of the Danshen samples

	Growth Condition	Geographic origin						Total
		Shandong	Shanxi	Henan	Sichuan	Zhejiang	Hebei	
Number of samples in Set-A	Wild	50	30	30	0	0	0	110
	Cultivated	70	40	25	30	35	40	240
	Total	120	70	55	30	35	40	350
Number of samples in Set-B	Wild	14	9	10	0	0	0	33
	Cultivated	22	13	9	9	10	13	76
	Total	36	22	19	9	10	13	109
Total		156	92	74	39	45	53	459

- Pre-processing
- Re-sampling the spectrums
- Scanning for the effective frequency range according to CV performance on training set
- R-SVM with Gaussian kernel

$$f(\mathbf{x}) = \mathrm{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i \cdot \mathbf{x}) + b^*\}$$

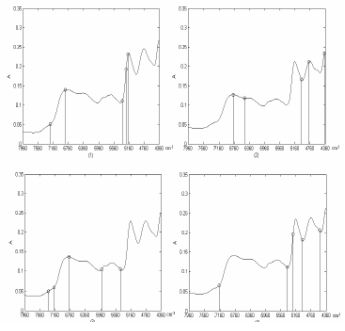
$$DQ(k) = (1/2)\sum_{i,j=1}^{n} \alpha_i^* \alpha_j^* y_i y_j [K(\mathbf{x}_i, \mathbf{x}_j) - K(\mathbf{x}_{i(-k)}, \mathbf{x}_{j(-k)})]$$

- One-vs-all scheme for multi-category classification
- Test on independent data set
- GA to select a unified group of features for all classes based on training set, then test on independent set

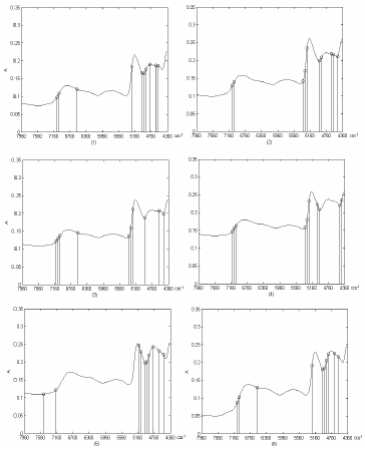
Results

Table 4 The classification accuracy with recursive SVM with different number of features, for Danshen

Number of features	Accuracy for Classification of Origins							Accuracy for wild/cultivated discrimination
	with classifier1	with classifier2	with classifier3	with classifier4	with classifier5	with classifier6	**Final accuracy**	
72	98.3%	88.6%	89.1%	100%	100%	97.5%	95.1%	95.1%
50	98.3%	90%	85.5%	100%	100%	100%	95.1%	95.7%
30	98.3%	88.6%	87.3%	100%	100%	100%	95.1%	95.1%
20	98.3%	82.9%	81.8%	100%	100%	97.5%	92.9%	93.1%
10	94.2%	74.3%	70.9%	100%	100%	97.5%	91.4%	91.1%
5	95%	42.9%	63.6%	100%	94.3%	100%	80.6%	91.7%



The 5 features for discriminating the 4 origins of Baizhi (*angelicae dahuricae radix*)

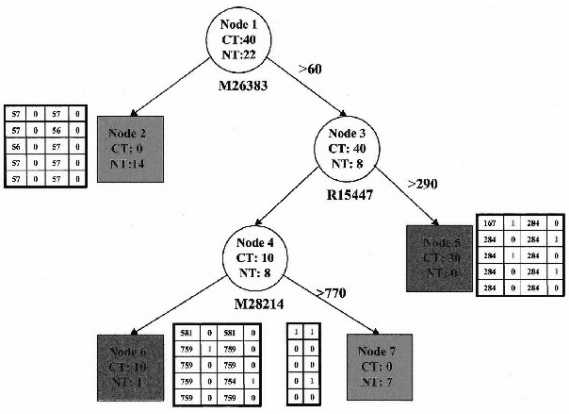


The 10 features for discriminating the 6 origins of Danshen (*salviae miltiorrhizae radix*)

Classification Trees and Random Forests

Classification Trees

- Select genes and separate subsets of samples with cutoffs in a tree manner



Random Forests

(Leo Breiman, *Machine Learning, 45: 5-32, 2001*)
(http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm)

- Many decision trees ➔ Random Forest

- Generate multiple trees by bootstrapping the samples (choosing *N* times with replacement from all *N* available training cases).
- For each node of the tree, randomly choose *m* variables on which to base the decision at that node. Calculate the best split based on these *m* variables in the training set.
- The RF classifier outputs the class that is the mode of the classes output by individual trees.

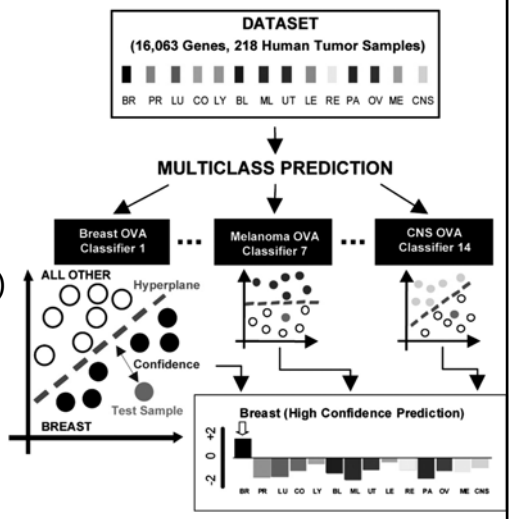
Multi-category Classification

Example:

SVM in multi-class classification of cancers with microarray data

- 14 tumor classes
- Methods: SVM, Recursive Feature Elimination, etc.
  - they concluded SVM performs the best
- Multi-class problem:

  multiple one-over-all (OVA) binary classifiers



S. Ramaswamy et. al. Multiclass cancer diagnosis using tumor gene expression signatures, *PNAS*, **98**(26): 15149-15154, 2001

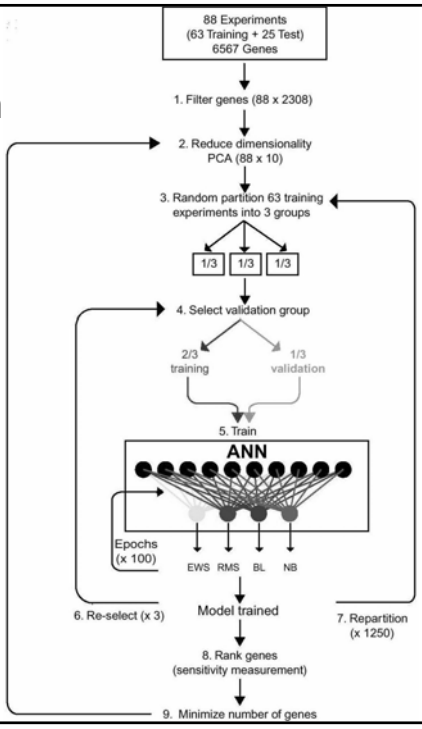
Dimension Reduction and Data Illustration

Dimension Reduction (Feature Transformation)

- To reduce the feature dimension by eliminating redundancy
  - PCA
  - SVD
  - KL Transform
- To eliminate the minor factors that might be due to noise
- To illustrate high-dimensional data on a plane or in a cube

Example: MLP for cancer classification with microarray data

- Multi-class:
  4 classes, one output node per class
- A set of NN models, then voting
- refs.
  - Khan, J. … and Meltzer, P.S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. **7**(6): 673-679, 2001
  - Gruvberger, S. … and Meltzer, P.S. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**: 5979-5984, 2001



Xuegong Zhang
Tsinghua University

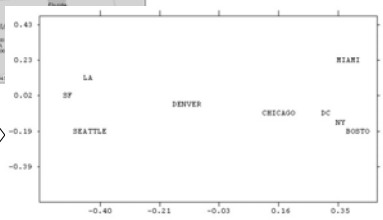
MDS: Multi-Dimensional Scaling

- Showing distance/similarity relations on 2D or 3D

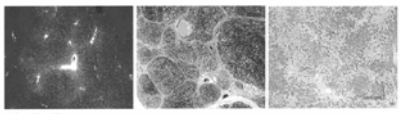
Similarity Matrix → Non-linear Transformation → 2D map

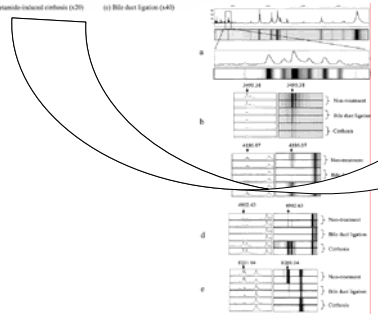
	1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

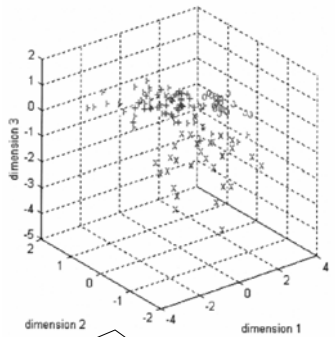
MDS



Example

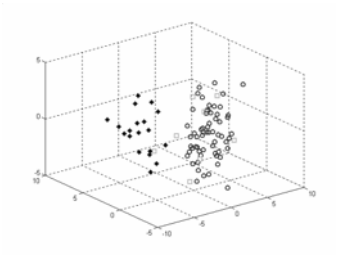


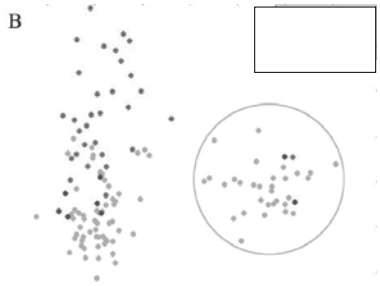


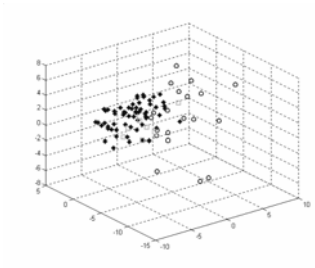


X-Q. Xu, C.K. Leow, X. Lu, X. Zhang, J.S. Liu, W.H. Wong, A. Asperger, S. Deininger, H.E. Leung, Molecular classification of liver cirrhosis in a rat model by proteomics and bioinformatics, *Proteomics*, 4: 3235-3245, 2004

Examples

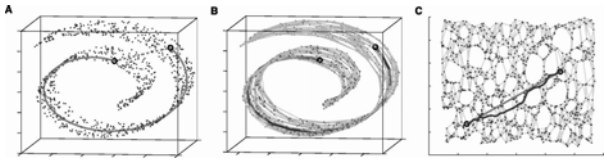






IsoMap

(Tenenbaum, 2000)



On this mapping of the yeast cell-cycle data (Spellman, 1998), gMap visualized groups of genes over-expressed in different cell-cycle stages

Zhang et al, *RECOMB*, 2004