Machine learning in high-throughput genomics and proteomics

(Part III)

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

Outline

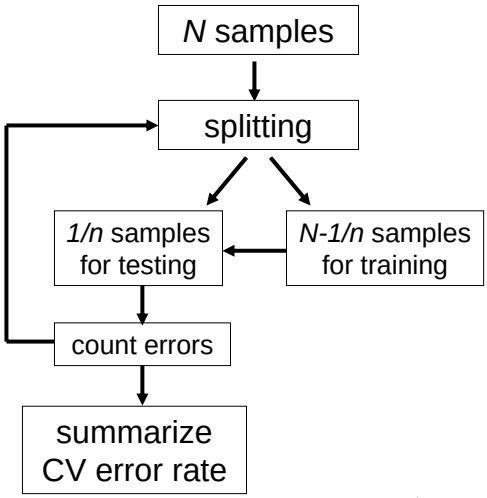
- Background
  - Microarray and mass-spectrometry technologies
  - Typical bioinformatics problems
  - Basic concepts of machine learning
- Microarray data mining with learning machines
  - Classification, clustering and gene selection
  - Problems, solutions, and possible pitfalls
- Further topics

Part III
Further Topics

- Assessing of the results
  - Open questions
- Follow-up study after microarray mining
  - New types of arrays

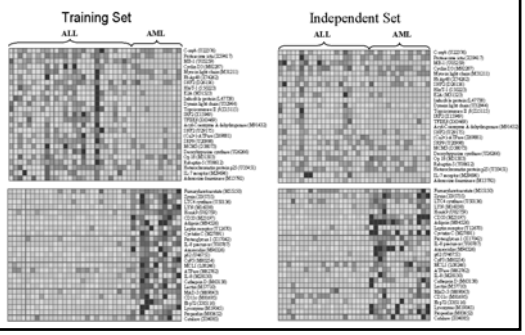
Classification Error Rates

- Training error
  - Apparent Error (AE)
- Error on independent test sets
  - Test Error
- Cross validation (CV) error
  - Leave-one-out (LOO)
  - *n*-fold CV



Example: ALL/AML classification (T. Golub et al, 1999)

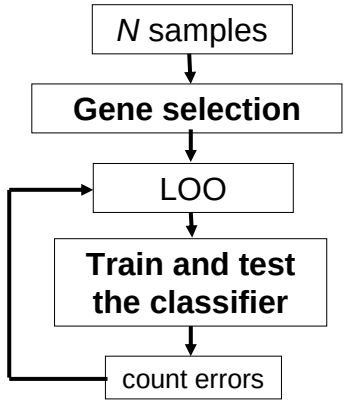
- Method:
  - Gene selection (with the SNR they defined)
  - Classification (weighted voting)
  - CV and independent test
- Result:
  - training error (CV): 0 error, 2 no-calls (would be error)(5.26%)
  - test: 5 no-calls (would be error) (14.7%)
- Observation:
  - Molecular classification of cancer types is promising
  - **The CV error is so different with the test error (?)**



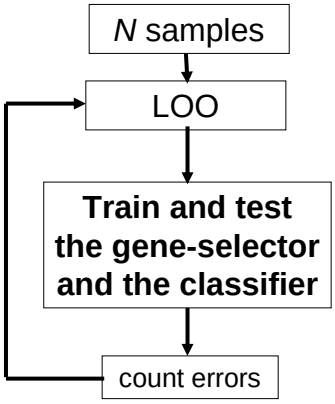
Xuegong Zhang
Tsinghua University

Two schemes of Cross Validation

- Scheme 1 (CV1)



- Scheme 2 (CV2)

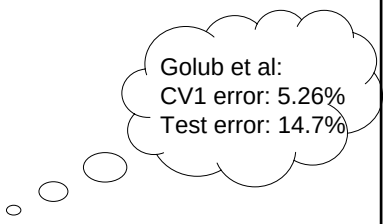


Xuegong Zhang
Tsinghua University

Zhang et al, 2001, unpublished

Example: R-SVM on ALL/AML classification

- Method:
  - R-SVM
  - CV2 and independent test
- Result:
  - training (CV2): 1 error (2.63%)
  - test: 1 error (2.94%)
- Observation:
  - R-SVM reaches a better accuracy
  - CV2 error is a better estimate of the test error



Zhang et al, 2001, unpublished

Artificial “Fake-class” Data

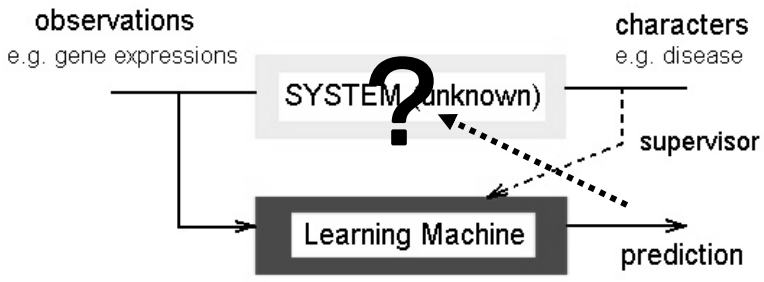
- Two fake "classes" of 20 cases each, generated from the same Gaussian model of 1000 fake genes.

# selected genes	CV1 error	CV2 error
1000	0.5	0.5
500	0.275	0.5
200	0.1	0.575
100	0.025	0.475
50	0.025	0.5
30	0.025	0.475
20	0	0.475

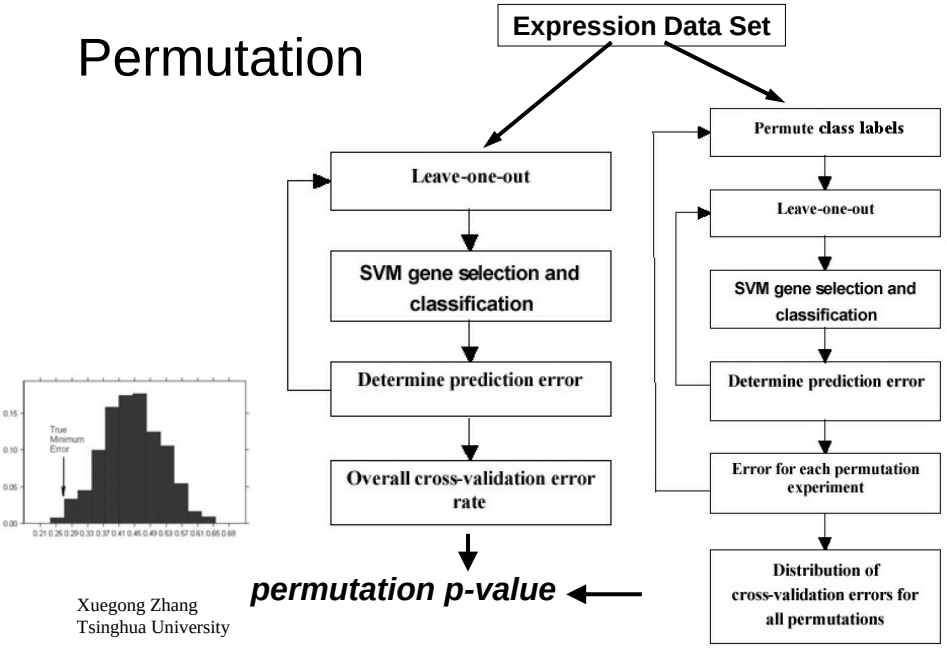
Cross validate the gene selection step!

- The timing of leaving the validation example is critical when sample size is small
  - x **Scheme1 CV (CV1)**:
    gene selection – LOO CV classification
  - ✓ **Scheme2 CV (CV2)**:
    LOO – gene selection and classification
  - ! ***Scheme1 CV may result in very biased misleading conclusions***

Significance of classification



Permutation



Xuegong Zhang
Tsinghua University

Example: ALL/AML classification (Zhang, 2001, unpublished)

- Method: R-SVM (CV and independent test)
- Result:
  training (CV2): 1 error (2.63%)
  test: 1 error (2.94%)
- Difference of CV schemes:
  R-SVM CV1 error: 0 Permutation p-value: 0.168
  R-SVM CV2 error: 0.0263 Permutation p-value: 0.000

Xuegong Zhang
Tsinghua University

Zhang et al, 2001, unpublished

More experiments with R-SVM

Data set	Scheme1 CV		Scheme2 CV		Test error on independent set
	Min Error	**P-value**	**Min Error**	**P-value**	
Leukemia: ALL/AML	0	0.168	0.026	0.000	0.029
Lung/Breast cancer (20 cases)	0	0.08	0.050	0.00	0.039
Lung/Breast cancer (198 cases)	0	0.00	0	0.00	-
Breast cancer: LN+/-	0.112	0.219	0.382	0.166	-
Breast cancer: ER+/-	0	0.000	0	0.000	-

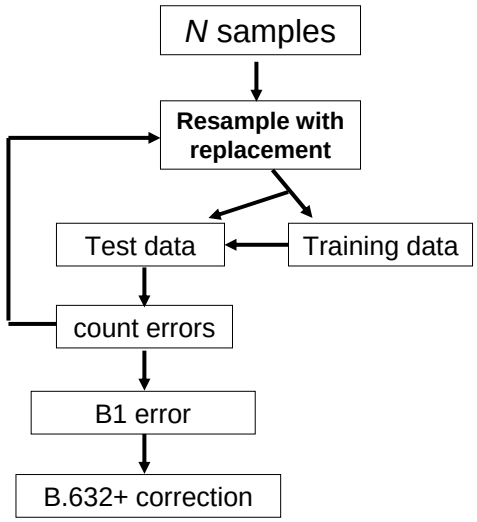
The bootstrap error

$$B.632+ = (1-w)\,\mathrm{AE} + w\mathrm{B1}$$

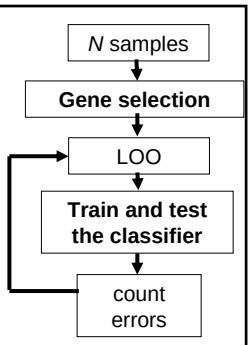
AE: apparent error rate (training error)

B1: Bootstrap error rate

C. Ambroise & G.J. McLachlan, Selection bias in gene extraction on basis of microarray gene-expression data, *PNAS*, **99**(10): 6567-6572, 2002



Why CV1 results can be so biased?



- *"Information leaking"*
- Leaving one sample out can affect the selection greatly and result in very different classification performance.
  - Is the selection of genes based on a small sample set stable or reliable?

→→→ questions about gene selection

Q1. How stable are the gene selection results?

Zhang et al, 2001, unpublished

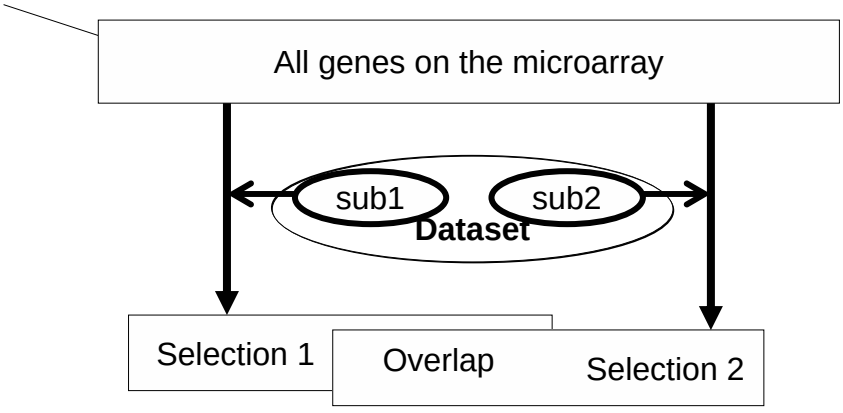
Different methods on the Leukemia Data

Method	selected genes	Classification performance	**Genes overlapping with baseline genes**
T. Golub et al. (baseline)	50 genes	5.26% CV1 error 14.7% test error	**baseline**
I. Guyon et al (SVM-RFE)	16 genes	0 CV1 error 0 test error	**< 25%**
X. Zhang (R-SVM)	50 genes	0 CV1 error 2.63% CV2 error 2.94% test error	**14 genes**
X. Zhang, after removing the 50 baseline genes	50 genes	0 CV1 error 7.89% CV2 error 5.88% test error	**0 genes**

--- Different methods select different genes.

The same method on different data sets:

--- Experiments on exclusive subsets

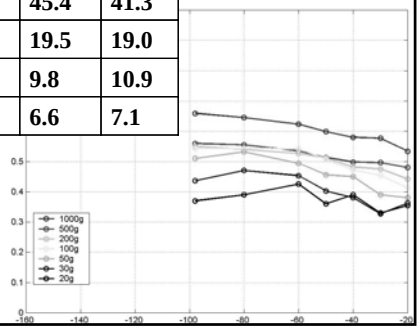


Zhang et al, 2001, unpublished

Same method on different data sets:
--- Observation on the lung-vs-breast data

SampleSize / #genes	98	80	60	50	40	30	20
1000	**659.2**	**644.9**	**623.0**	**598.9**	**579.7**	**576.5**	**534.4**
500	**280.0**	**277.6**	**267.2**	**256.5**	**249.2**	**247.9**	**240.1**
200	**110.0**	**108.0**	**105.2**	**102.4**	**96.4**	**94.9**	**88.4**
100	**54.3**	**54.3**	**54.2**	**50.7**	**47.2**	**45.4**	**41.3**
50	**25.5**	**26.6**	**24.7**	**22.8**	**22.5**	**19.5**	**19.0**
30	**13.1**	**14.1**	**13.6**	**12.1**	**11.4**	**9.8**	**10.9**
20	**7.4**	**7.8**	**8.5**	**7.2**	**7.8**	**6.6**	**7.1**

Average number of genes overlapping in the two selections from exclusive subsets with R-SVM

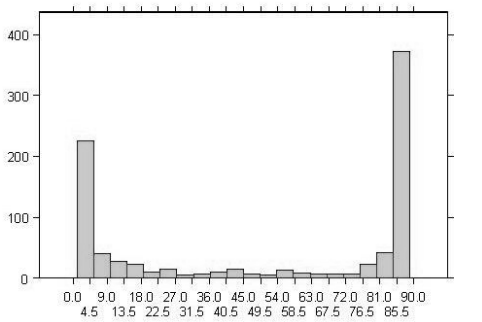


Xuegong Zhang
Tsinghua University

Zhang et al, 2001, unpublished

The case of less separable samples

- LN+/- example:
  - 89 LOO experiments
  - Selecting 500 genes in each run
    - Totally 864 genes are selected
    - 161 (18.6%) genes appear in all 89 selections
    - 388 (44.9%) genes appear in 84 or more selections
    - 225 (26.0%) genes appears in only 5 or less selections



Number of times for a gene to be selected in the 89 LOO experiments

Xuegong Zhang
Tsinghua University

20

Zhang et al, 2001, unpublished

Overfitting in selection due to SVM? --- simpler selection methods also suffer

SampleSize / #genes	98	80	60	40	20
1000	**568.2**	**576.6**	**471.8**	**436.2**	**324.4**
500	**271.2**	**269.8**	**206.2**	**181.2**	**122.2**
200	**100.4**	**96.6**	**69.6**	**59.2**	**36.2**
100	**47.0**	**44.2**	**31.4**	**25.2**	**15.6**
50	**22.0**	**20.6**	**15.8**	**11.2**	**4.6**
30	**14.0**	**12.0**	**10.0**	**7.4**	**2.2**
20	**9.4**	**9.0**	**7.4**	**4.6**	**1.4**

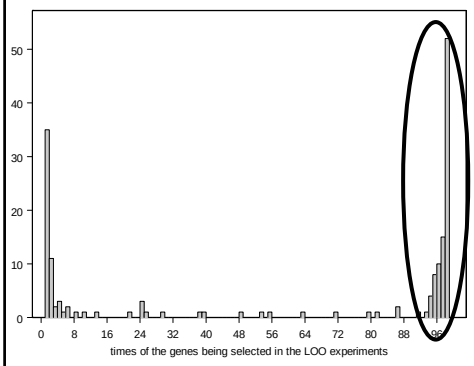
Average number of genes stable between the two selections with two exclusive subsets by t-test, the breast-vs-lung dataset

Observation:

- Different methods select different groups of genes.
- Each specific datasets also highlight different groups of genes.

Zhang and Lu et al, 2002, unpublished

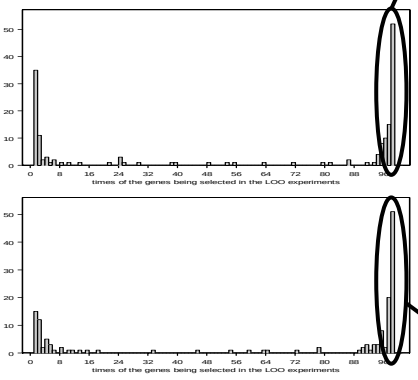
Improvement:
Voting with LOO or bootstrap selection



- Vote for the majority of genes selected with the LOO or bootstrap resampling experiments.

Zhang et al, 2002, unpublished

The performance of LOO selection on the lung-vs-breast dataset



Voting threshold	#genes from dataset 1	#genes from dataset 2	# overlapping genes
98	**52**	**51**	**28 (54.4%)**
>=97	67	71	37 (53.6%)
>=90	91	93	48 (52.2%)
>=1	165	148	84 (26.8%)

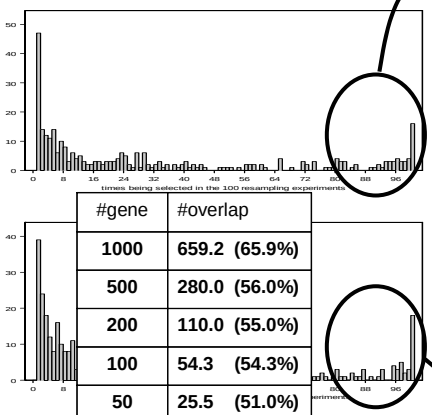
comparison:

stableness of two one-time selections.

#gene	#overlap
1000	**659.2 (65.9%)**
500	**280.0 (56.0%)**
200	**110.0 (55.0%)**
100	**54.3 (54.3%)**
50	**25.5 (51.0%)**
30	**13.1 (43.3%)**
20	**7.4 (37.0%)**

Zhang et al, 2002, unpublished

The performance of resampling selection on the lung-vs-breast dataset



Voting threshold	# genes from dataset 1	# genes from dataset 2	# overlapping genes in the two selections
=100	16	18	5 (29.4%)
>=90	43	39	23 (56.1%)
>=80	**57**	**52**	**34 (62.4%)**
>=70	68	64	41 (62.1%)
>=60	76	76	48 (63.2%)
>=50	86	89	50 (57.1%)
>=1	308	325	217 (68.6%)

#gene	#overlap
1000	**659.2 (65.9%)**
500	**280.0 (56.0%)**
200	**110.0 (55.0%)**
100	**54.3 (54.3%)**
50	**25.5 (51.0%)**
30	**13.1 (43.3%)**
20	**7.4 (37.0%)**

Observations

- Bootstrapping or LOO selection seems to improve the stableness of gene selection
- Still there is a big variation between selections from different datasets
  - Possible explanations:
    - Insufficient sampling
    - Biological heterogeneity

Q2. Have we caught all the guilty genes, or just some unlucky ones?

Another view on the different results on the Leukemia Data

Method	selected genes	Classification performance	**Genes overlapping with baseline genes**
T. Golub et al. (baseline)	50 genes	5.26% CV1 error 14.7% test error	**baseline**
I. Guyon et al (SVM-RFE)	16 genes	0 CV1 error 0 test error	**< 25%**
X. Zhang (R-SVM)	50 genes	0 CV1 error 2.63% CV2 error 2.94% test error	**14 genes**
X. Zhang, after removing the 50 baseline genes	50 genes	0 CV1 error 7.89% CV2 error 5.88% test error	**0 genes**

--- There may be many guilty genes still at large!

Zhang et al, 2002, unpublished

Example:

to remove the top informative genes and do it again

Leukemia dataset
(27+11, 7071)

#genes removed	Minimum CV error
0	0.026316
92	0.078947
171	0.184211
240	0.184211
311	0.236842
430	0.289474
476	0.342105

ERsub2 dataset
(19+19, 12558)

#genes removed	Minimum CV error
0	0.000000
20	0.026316
104	0.184211
168	0.236842
229	0.263158
284	0.289474
338	0.315789

Q3. How can we evaluate the significance of ML-selected genes?

False positive and false discovery rates

Statistical decision	Truth	
	Negative (not to be selected)	**Positive (to be selected)**
Declared positive (being selected)	**Type I error False Positive (FP)**	**Correct True Positive (TP)**
Declared negative (not being selected)	**Correct True Negative (TN)**	**Type II error False Negative (FN)**

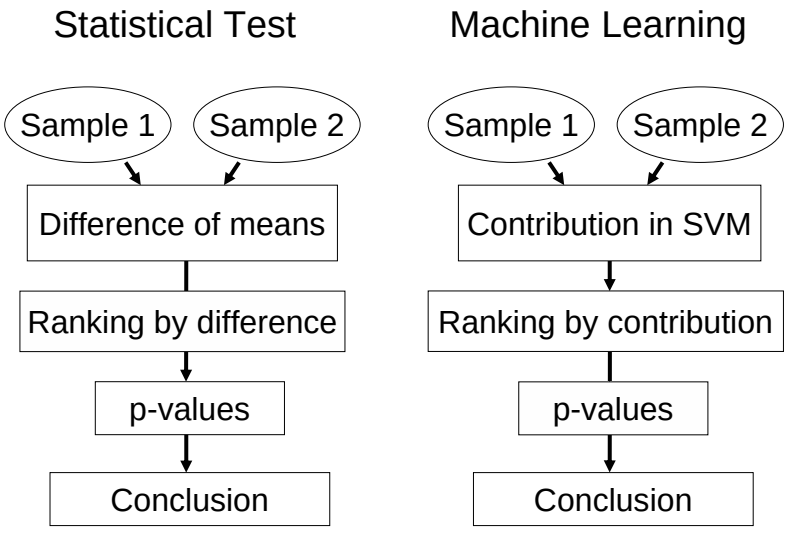
$$\text{false positive rate} = \text{FP}/(\text{FP}+\text{TN})$$

$$\text{FWER} = \text{P}(\text{FP} > 1) \le m \cdot \alpha / m$$

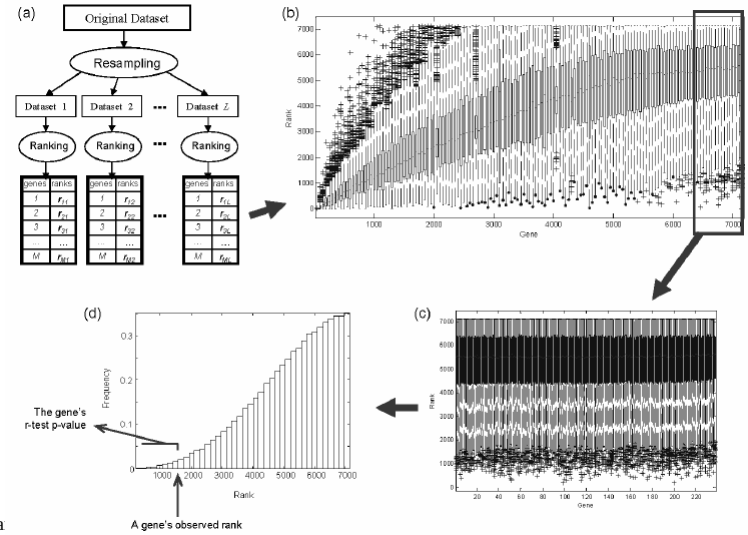
$$\text{false discovery rate} = \text{FP}/(\text{FP}+\text{TP})$$

--- Hard to apply for genes selected with machine learning approaches.

Q: Significance of ML ranking



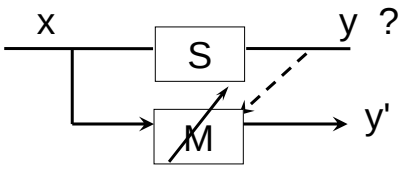
C. Zhang, X. Lu and X. Zhang, Significance of gene ranking for classification of microarray samples, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(3): 312-320, July-Sept, 2006

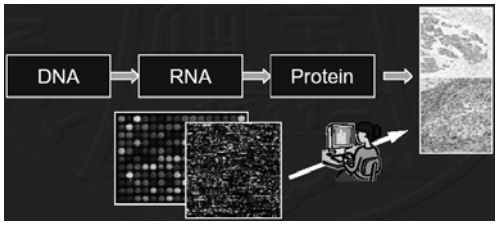


Xuegong Zha
Tsinghua University

Discussion

The scenarios when most ML methods were invented are different with current biology applications

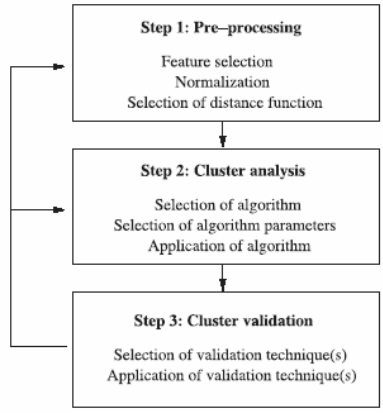




- Machine Learning:
  - Assumption:
    - Existence of the (y, **x**) dependence
    - Sufficient, i.i.d. data
  - Black-box ok
- Biology Applications:
  - The existence of the (y, **x**) dependence is to be tested
  - Very limited, noisy data
  - Results explainable

Validating Unsupervised Learning Results

- External measures
  - According to some external knowledge
  - Caution for bias and subjectivity
- Internal measures
  - Quality of the clusters according to the data
  - Compactness and separation
  - Stability
  - …



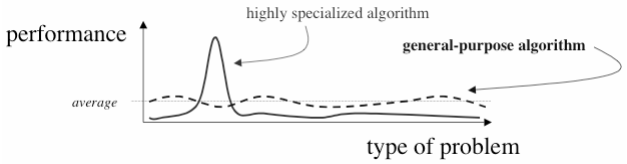
J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, *Bioinformatics*, **21**(15): 3201-3212, 2005

Comparing Methods?

- My personal view:
  - Focusing on our specific biological question
  - Current biological data might not be good as a standard for generic comparison of computational methods
  - Understanding the underlying assumption and characteristics of different methods are more important

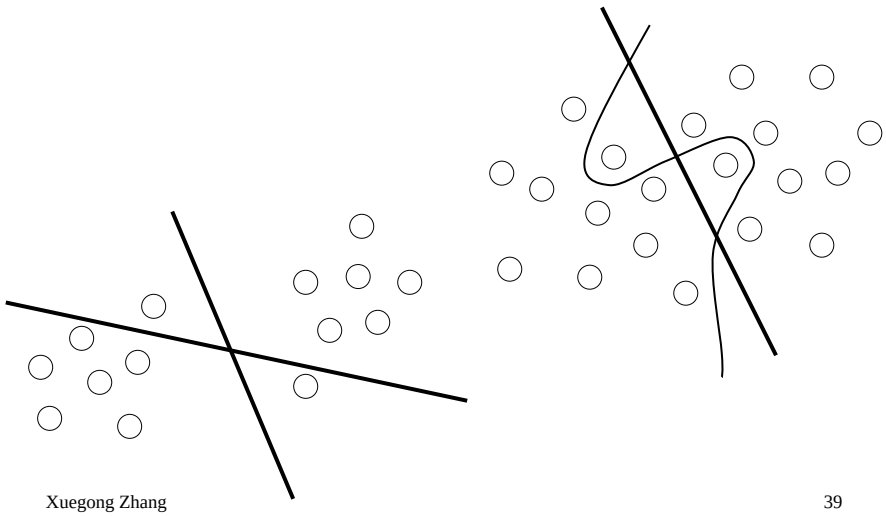
No-free-lunch theorem

- "… all algorithms that search for an extremum of a cost function perform exactly the same, when averaged over all possible cost functions." [Wolpert and Macready, 1995]
- "A general-purpose universal optimization strategy is theoretically impossible, and the only way one strategy can outperform another is if it is specialized to the specific problem under consideration." [Ho and Pepyne, 2002]

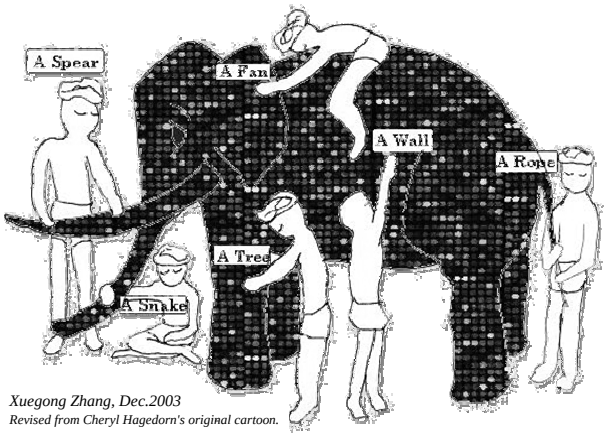


- In Machine Learning, every method has its own conditions and assumptions, and therefore advantages and limitations.

To learn: a challenge forever



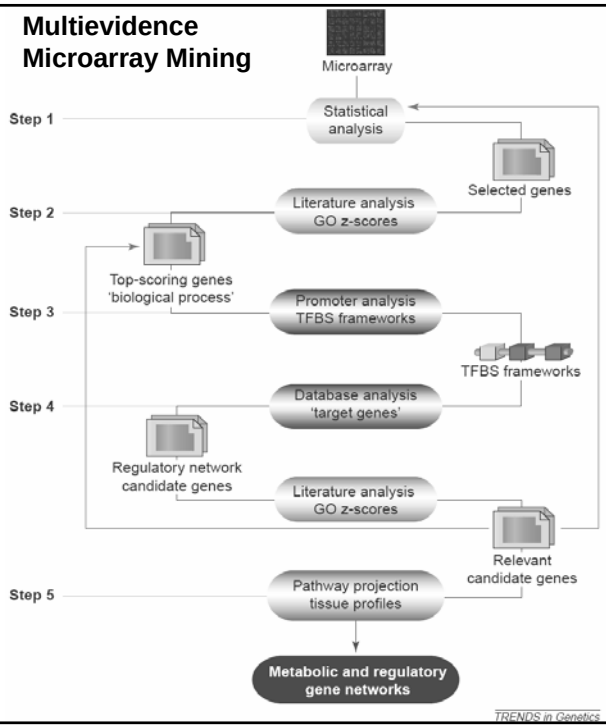
A picture that we should keep in mind



Xuegong Zhang, Dec.2003
Revised from Cheryl Hagedorn's original cartoon.

What do we do after mining microarray data?

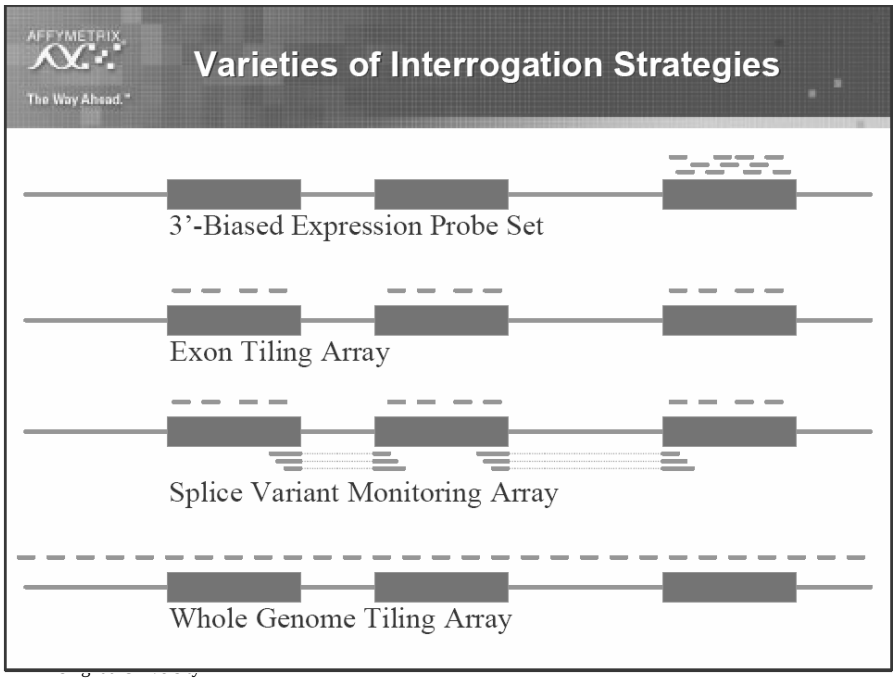
Multievidence Microarray Mining



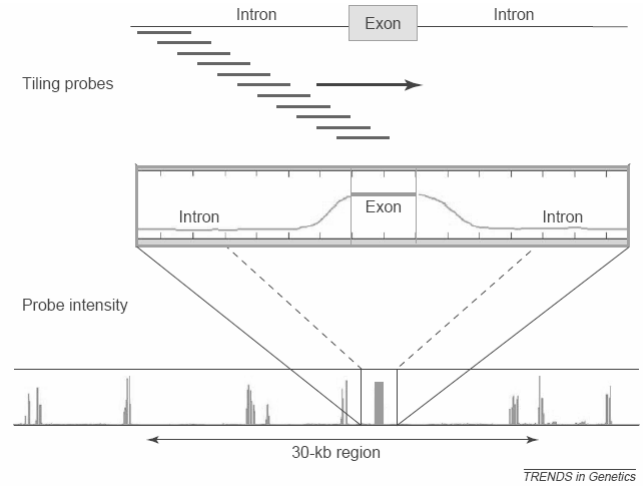
M. Seifert et al, Multievidence microarray mining, *TiG*, 21: 553-558 2005

Strategy for multievidence microarray mining. The colored boxes in the center represent basic tasks of the workflow. The list symbols to the left and to the right of this central column represent intermediate results. The blue pathway from the microarray to the 'metabolic and regulatory networks' symbolizes the flow of analysis. The purple pathways indicate iterations where the result list is used as input for a refined analysis with the respective next task. 42

New types of microarrays



Tiling arrays



J.M. Johnson et al, Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments, *TiG*, 21(2): 93-102, 2005

Beyond the Protein-coding Genes

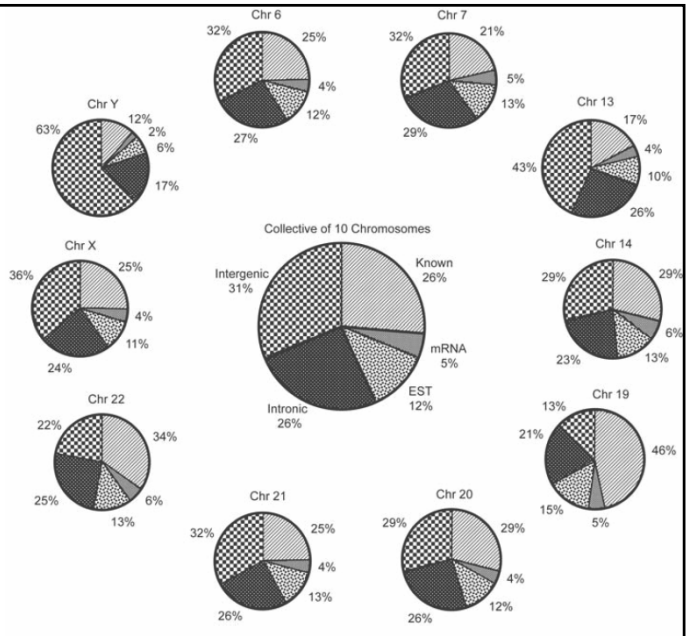


Fig. 1. The correlation of detected transcription in one of eight cell lines to annotations along each of the 10 chromosomes is shown for each chromosome individually and as a collective of all chromosomes. The detected transcription was determined using poly A+ cytosolic RNA from each of the eight cell lines. The annotations used in this correlation are defined in (*15*). The pattern code used in the central pie chart is used in all other pie charts.

J. Cheng et al, Transcriptional map of 10 human chromosomes at 5-nucleotide resolution, *Science*, 308: 1149-1154

Acknowledgements

- Besides the references I cited in the lectures
  - Wing H. Wong, HSPH→Stanford
  - Michael Q. Zhang, CSHL/Tsinghua
  - Cheng Li, HSPH/DFCI
  - Xin Lu, HSPH→UCSD
  - Chaolin Zhang, Tsinghua→CSHL
  - Xuesong Lu, Tsinghua