

Cancer Classification Using Informative Gene Profiles

Xue-wen Chen

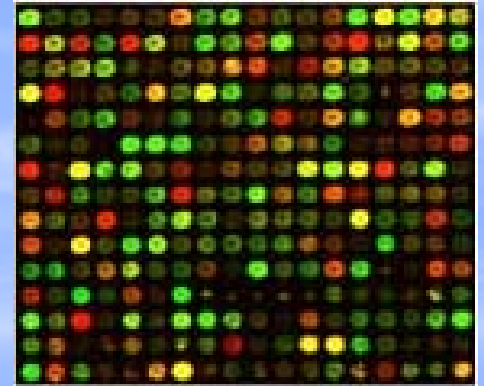
**Bioinformatics and Computational Life-Sciences Laboratory
The university of Kansas**

Interface 2004, Baltimore

OUTLINE

- Introduction
- Microarray Data Analyses
- Bootstrapped GA/Margin Methods
- Experiment Results
- Discussions

INTRODUCTION



- **Traditional biology:** one (or few) gene in one experiment, hard to capture the “whole picture” of gene function
- **Microarray:** monitor thousands of genes on a single chip simultaneously; provides a better understanding of the interactions among genes; helps explore the underlying genetic causes of many human diseases.

MICROARRAY: CANCER CLASSIFICATION

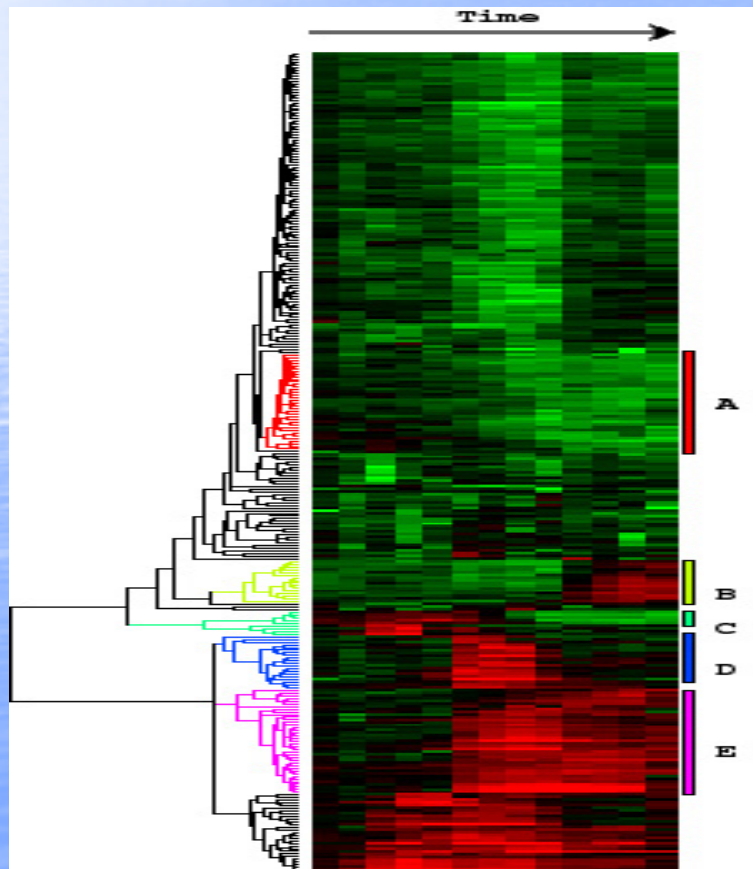
- **Microarray** has been successfully applied to cancer classification problems
- **According to Dudoit, Fridlyand, and Speed**, there are three main problems related to microarray based cancer classification:
 - Cancer discovery (clustering)
 - Cancer classification into known classes (supervised learning)
 - Identification of gene “markers” (gene selection)

OUTLINE

- Introduction
- *Microarray Data Analyses*
- Bootstrapped GA/Margin Methods
- Experiment Results
- Discussions

UNSUPERVISED METHODS: CLUSTERING

Partition genes (or samples) into homogeneous groups in order to explore the similarity among genes



- **Hierarchical Clustering** (Eisen et al. Proc. Natl. Acad. Sci. 1998)
- **SOMs** (Tamayo et al. Proc. Natl. Acad. Sci., 1999)
- **K-means** (Tavazoie et al. Nature Genetics, 1999)
- **More**

SUPERVISED LEARNING

- **Learning (Training) Task**
 - Given: Expressed gene profiles of cells and their class labels
 - Learn: Models distinguishing cells of one class from cells in other classes (genes are features)
- **Classification (Test) Task**
 - Given: Expression profile of a cell whose class is unknown
 - Test: Predict the class to which this cell belongs

SUPERVISED LEARNING METHODS

- **Neural Networks** (Mateos et al. 2002)
- **K-nearest Neighbors** (Theilhaber et al. 2002)
- **Support Vector Machines** (Brown et al. 2000)
- **Fisher Discriminant Analysis** (Dudoit et al. 2002)
- **Decision Trees** (Dubitzky et al. 2000)
- **And more**

CHALLENGES IN LEARNING MICROARRAY DATA

- **High dimensionality:** in microarray data analysis, the number of features (genes) is normally much larger than the # of training samples.
- **Often noisy and not normally distributed** (Hunter et al. 2001, bioinformatics)
- **Too many features are not desirable in learning:** poor generalization is expected (or overfitting).
- **Essential to reduce the # of genes to use**

GENE SELECTION (MARKER IDENTIFICATION)

- Feature selection is essential to reduce the test errors in microarray data classification.
- Given such huge amount of data, we need to remove genes irrelevant to the learning problems
- For diagnostics or identification of therapeutic targets, a small subset of discriminant genes is needed

GENE SELECTION

Golub et al. (1999): $[\text{mean}(+) - \text{mean}(-)] / [\text{std}(+) + \text{std}(-)]$.

Xing et al. (2001): information gain to rank genes.

Long et al. (2001): t-test with a Gaussian model

Furey et al. (2000): the Fisher score

Newton et al. (2001): a Gamma-Gamma-Bernoulli model

Kerr et al., (2000): ANOVA A F-statistics

Dudoit et al. (2002): a nonparametric t-test

Bo and Jonassen (2002), Inza et al. (2002): Forward selection

Khan et al. (2001): PCA

Li et al. (2001): GA/knn

more ...

Univariate vs. Multivariate

Filter vs. wrapper

IN THIS PAPER

- **A method for:**
 - Cancer classification and gene identification
 - Simultaneously
- **Wrapper methods**

OUTLINE

- Introduction
- Microarray Data Analyses
- **Bootstrapped GA/Margin Methods**
- Experiment Results
- Discussions

Gene Selection: General Idea



Criterion function: should generalize (predict) well (wrapper); particularly important in microarray data classifications, since very limited training samples are available.

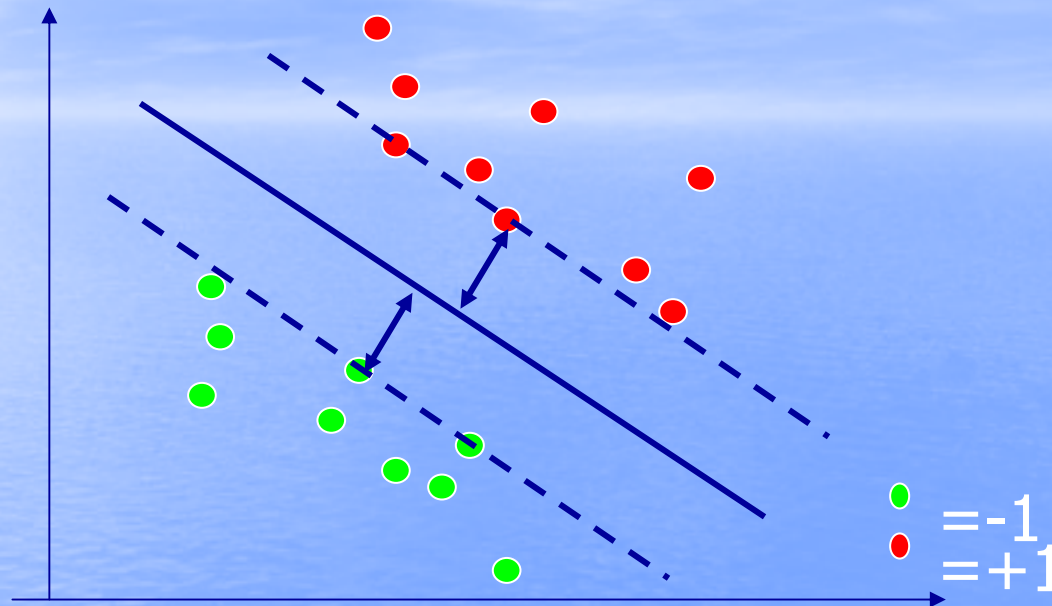
Search algorithms: efficient for very high-d data (e.g., # genes ~ 2000) in terms of both computation time and solutions

Margin: ability to generalize; used as the criterion function

GAs: better performance than SFS, much faster than exhaustive search; used as the search algorithm

Bootstrapping: because of limited training samples

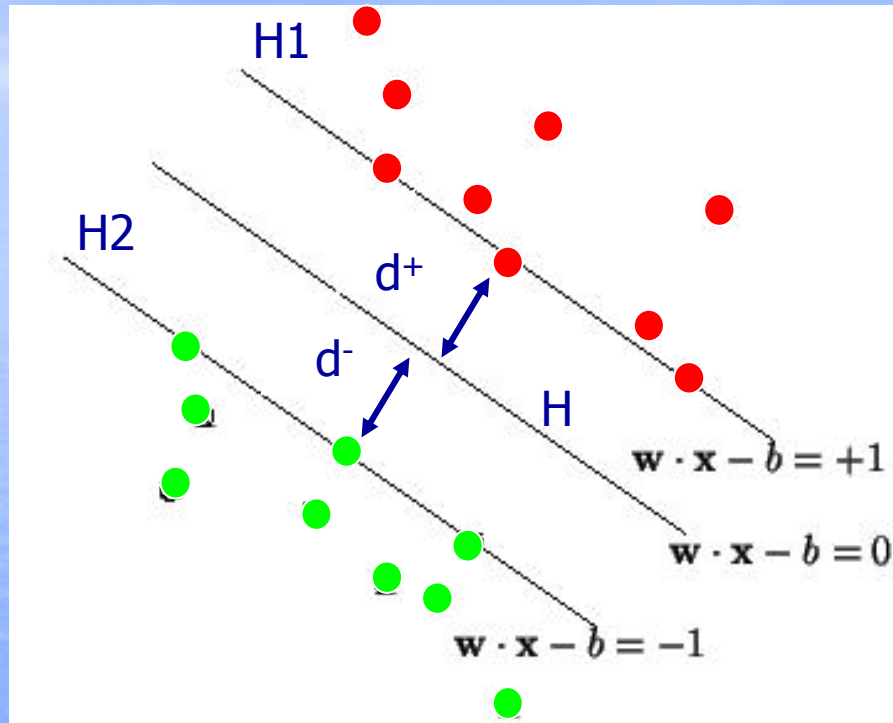
MAXIMUM MARGIN



maximizing the **margin** (the minimum distance between a hyperplane that separates two classes and the closest training samples to the decision surface).

Motivation: Obtain tightest possible bounds for generalization ; is capable of avoiding overfitting

MARGIN



- Define the hyperplane H such that:
 $x_i \cdot w + b \geq +1$ when $y_i = +1$
 $x_i \cdot w + b \leq -1$ when $y_i = -1$

MARGIN

- In order to maximize the margin, we need to minimize $\|w\|$.
with the constraints: no datapoints between H1 and H2:

$$y_i(x_i \cdot w + b) - 1 \geq 0$$

- Equivalently (a dual problem), maximize:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

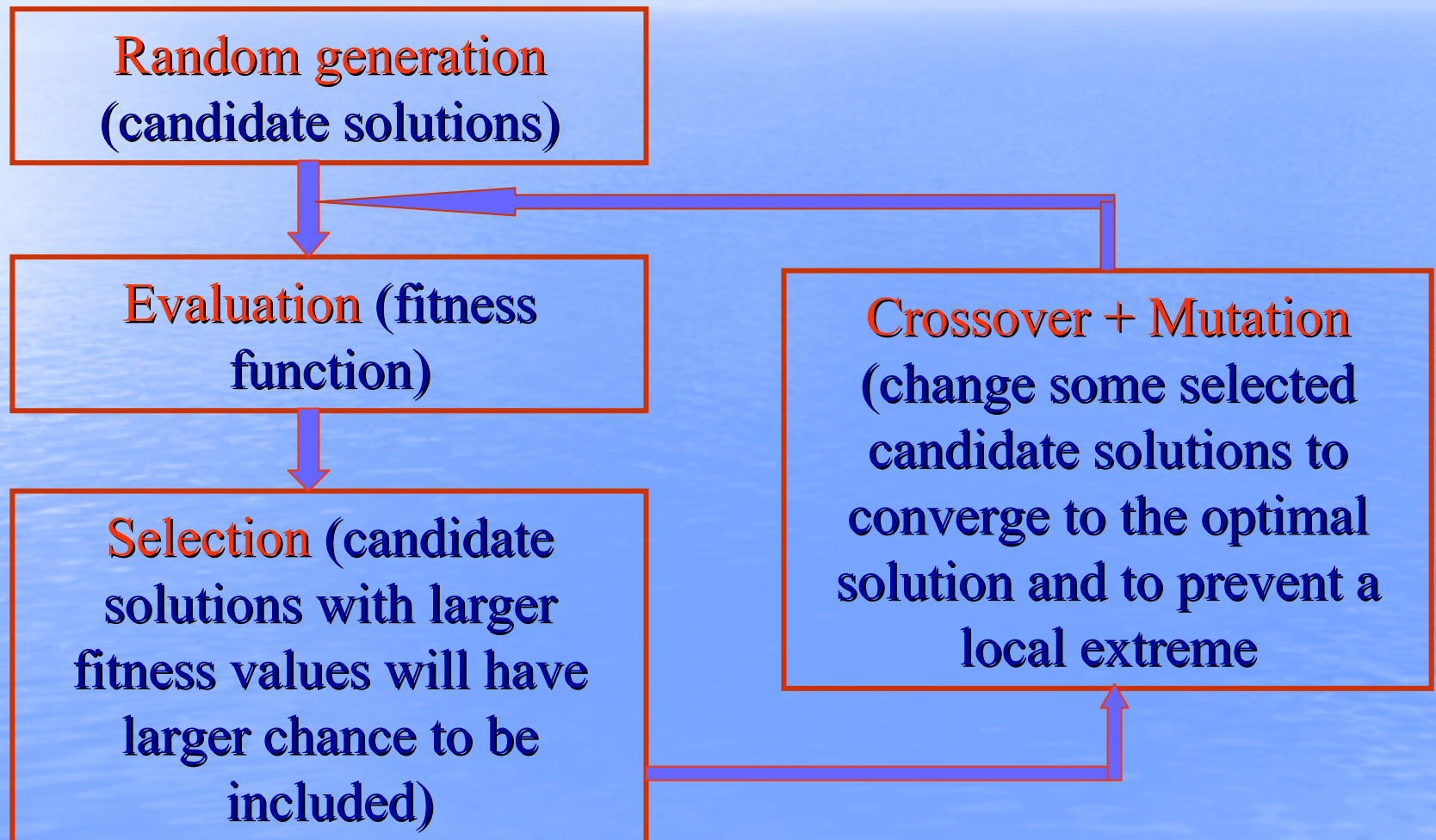
with respect to the α_i 's, subject to $\alpha_i \geq 0$ and $\sum_{i=1}^m \alpha_i y_i = 0$

Margin:

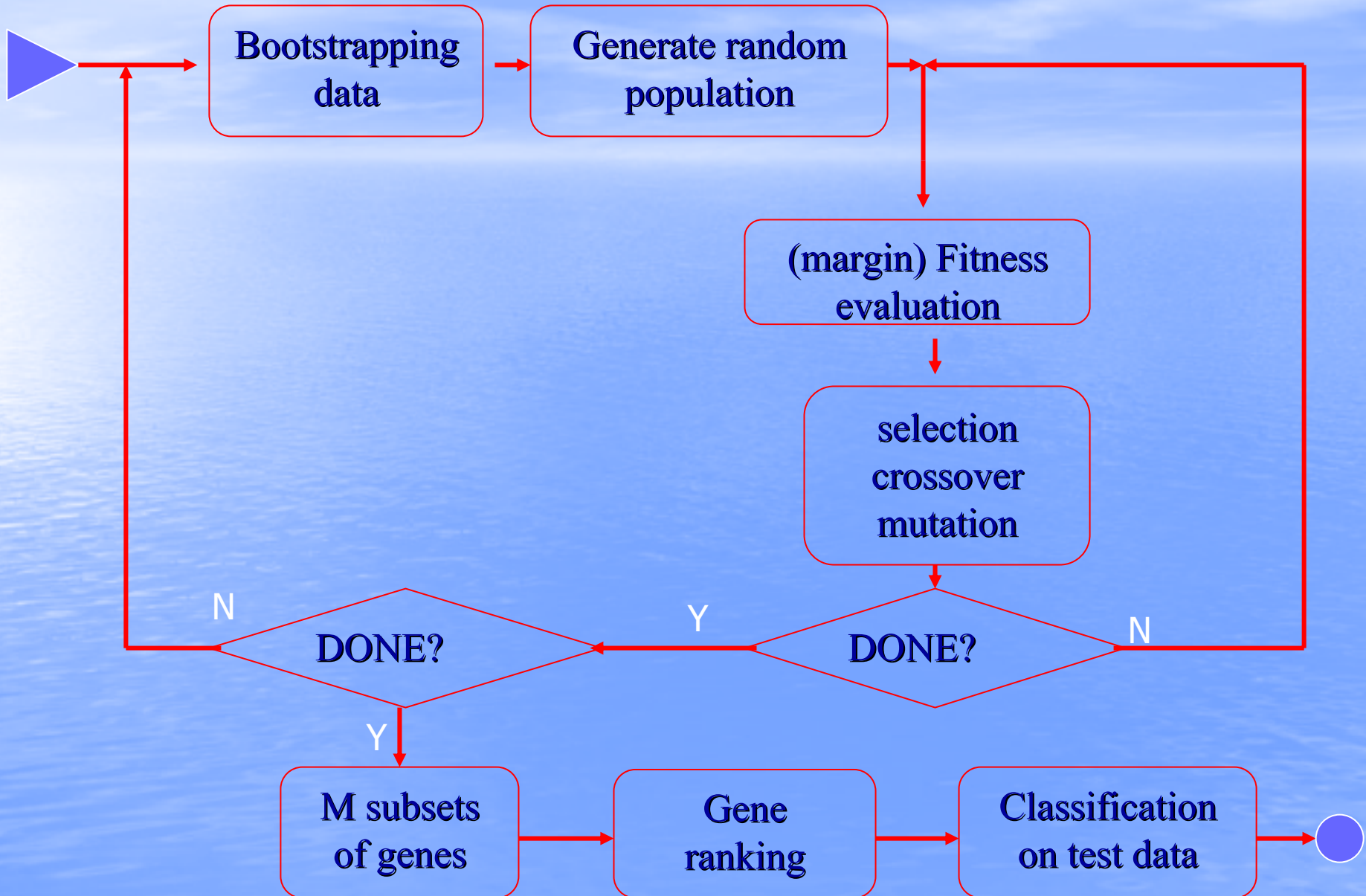
$$d = \frac{2}{\sqrt{\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)}}$$

GENETIC SEARCH ALGORITHMS

Goal: identify the best subsets of genes evaluated by margin



BOOTSTRAPPED GA/MARGIN:



OUTLINE

- Introduction
- Microarray Data Analyses
- Bootstrapped GA/Margin Methods
- Experiment Results
- Discussions

Dataset 1: Colon Cancer

- Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999) Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, **96**, 6745-6750.

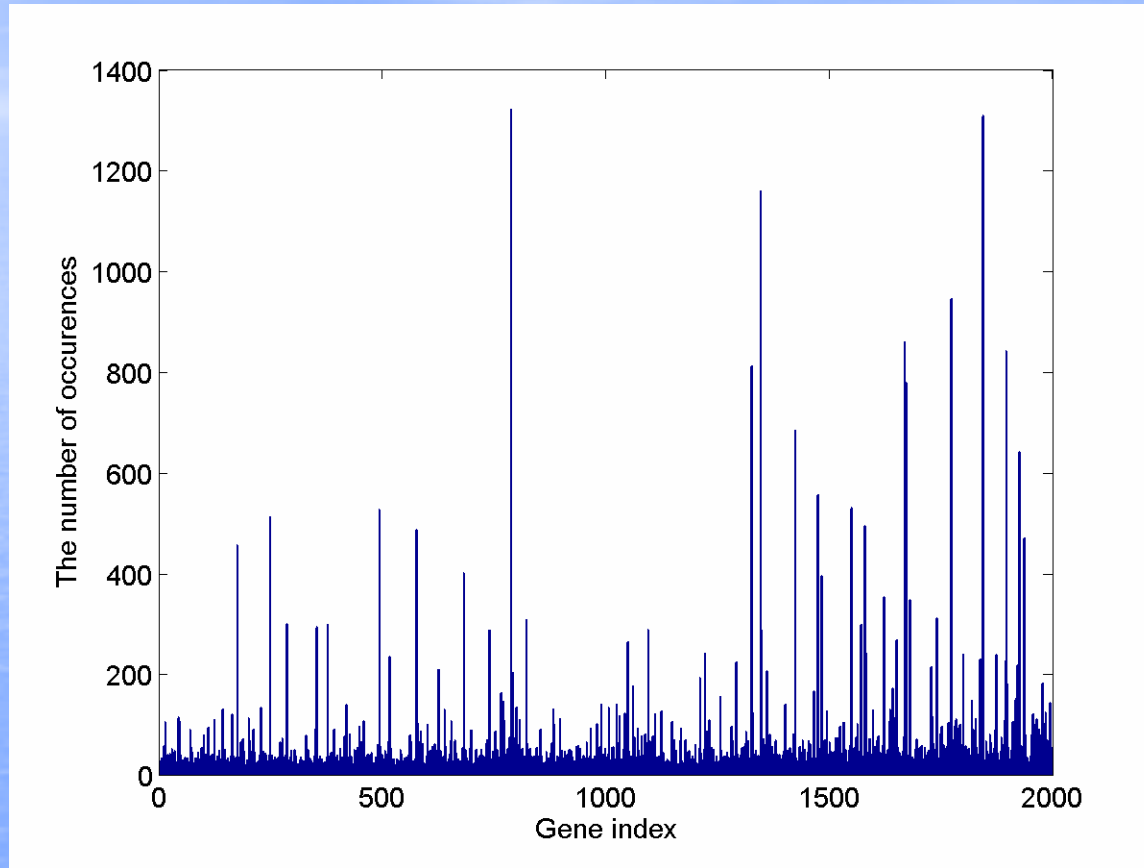
<u>Cancer</u>	<u># samples</u>	<u># genes</u>	<u>task</u>
Colon	62 (22 normal + 40 cancer)	2000	cancer/normal

GENE SELECTION

- 3000 bootstrapping datasets
- Each data set contains 18 normal + 36 cancer
- Genes are ordered based on the number of occurrences

GENE SELECTION

The Number of Occurrence



Gene Index

Interferon-induced > 1321 times, while sparc precursor = 0.

CANCER CLASSIFICATION

The top 50 genes are used for cancer classification

Classifier: linear SVMs

300 bootstrapping tests (12 normal + 25 cancer)

Compared to GA/3-NN (Li et al. 2001) with top 50 genes

	GA/Margin	GA/knn
Training data	0	0
Test data	950	1622

LEUKEMIA DATASET

Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Lo, Downing, Caligiuri, Bloomfield, Lander "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" in *Science* Vol. 286, 1999

<u>Cancer</u>	<u># samples</u>	<u># genes</u>	<u>task</u>
Leukemia	72 (47 ALL + 25 AML)	7129	AML/ALL
		1800	

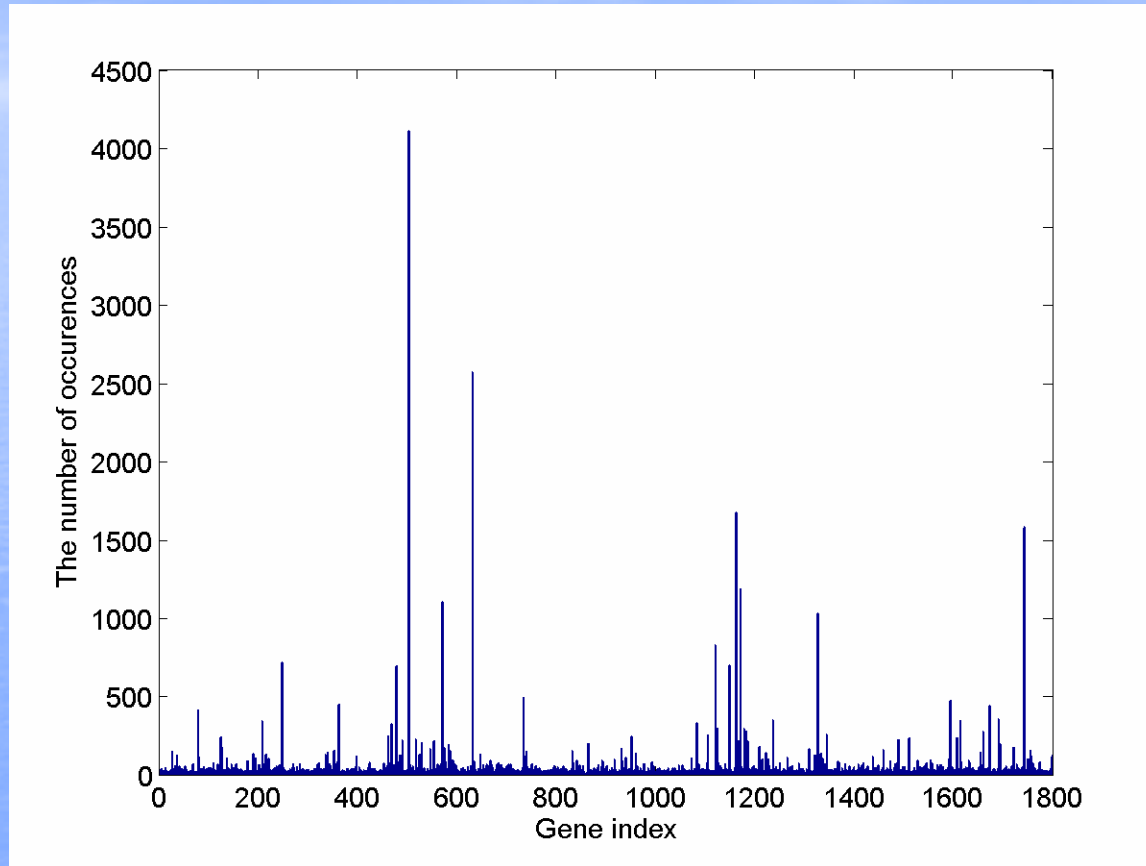
Training and test sets were prepared under different expression cond.

GENE SELECTION

- 4500 bootstrapping datasets
- Each data set contain 17 AML + 35 ALL
- Genes are ordered based on the number of occurrences

GENE SELECTION

The Occurrence



Gene Index

CANCER CLASSIFICATION

The top 50 genes are used for cancer classification

Classifier: linear SVMs

500 bootstrapping tests (35 ALL + 17 AML)

Compared to GA/3-NN (Li et al. 2001) with top 50 genes

	GA/Margin	GA/knn
Training data	0	0
Test data	259	722

COMPUTATIONAL CONSIDERATIONS

- **Individual ranking:** about 1 second
- **Forward selection:** about 10 seconds
- **GA/SVM selection:** about 5 hours
- **Exhaustive search:** about 5 months? (the selection of five features out of 86 took ~ 2 wks; the total combination # = 35M; out of 2000 (10^{14}))
- **The data collection** and preparation may take several months or years. It is reasonable that the data analysis takes a few hours

CONCLUSIONS

- A multivariate wrapper method is proposed for both gene identification and cancer classification
- Generalize well
- Need to test on more datasets