Molecular Biometry in the Genomic Era
DNA Microarrays

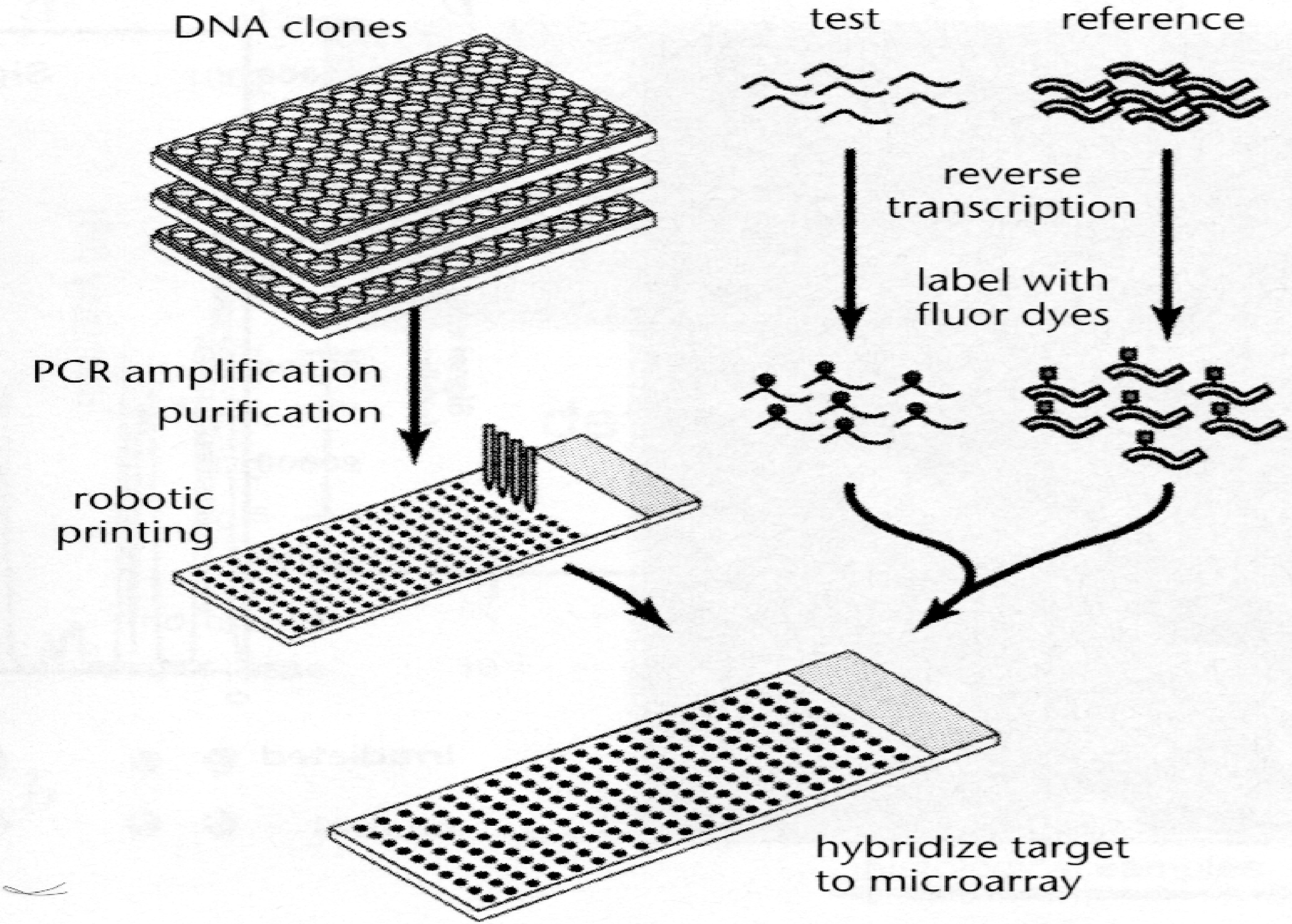
Richard Simon, D.Sc.

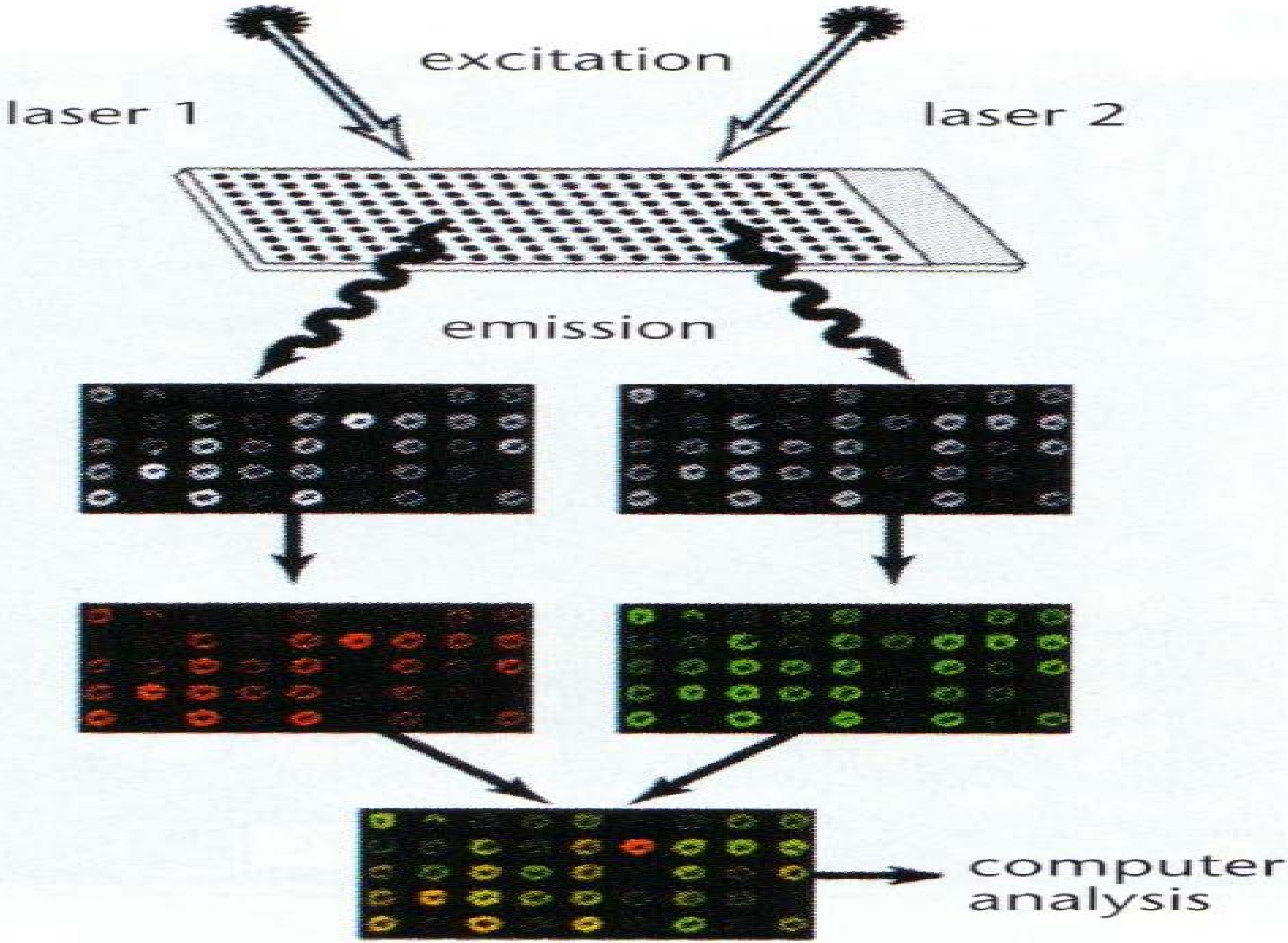
Chief, Biometric Research

National Cancer Institute

rsimon@nih.gov

http://linus.nci.nih.gov/~brb


DNA clones
test
reference
reverse transcription
label with fluor dyes
PCR amplification purification
robotic printing
hybridize target to microarray


excitation
laser 1
laser 2
emission
computer analysis

[Affymetrix] Hybridization Array

Microarray Myths

- That the greatest challenge is managing the mass of micro-array data
- That pattern-recognition or data mining are the most appropriate paradigm for the analysis of micro-array data
- That cluster analysis is the generally appropriate method of data analysis
- That comparing tissues or experimental conditions is based on looking for red or green spots on a single array

Microarray Myths

- That reference rna for two-channel arrays must be biologically relevant
- That multiple testing issues can be ignored
- That multiple testing issues are insurmountable
- That complex classification algorithms such as neural networks perform better than simpler methods for class prediction
- That pre-packaged analysis tools are a good substitute for collaboration with statistical scientists in complex problems

Design and Analysis Methods Should Be Tailored to Study Objectives

- Class Comparison (supervised)
  - For predetermined classes, establish whether gene expression profiles differ, and identify genes responsible for differences
- Class Prediction (supervised)
  - Prediction of phenotype using information from gene expression profile
- Class Discovery (unsupervised)
  - Discover clusters among specimens or among genes

Class Comparison Examples

- Establish that expression profiles differ between two histologic types of cancer
- Identify genes whose expression level is altered by exposure of cells to an experimental drug

Class Prediction Examples

- Predict from expression profiles which patients are likely to experience severe toxicity from a new drug versus which will tolerate the drug well
- Predict which breast cancer patients will relapse within two years of diagnosis versus which will remain disease free

Class Discovery Examples

- Discover previously unrecognized subtypes of lymphoma
- Identify co-regulated genes

Do Expression Profiles Differ for Two Defined Classes of Arrays?

- Not a clustering problem
  - Global similarity measures generally used for clustering arrays may not distinguish classes
- Supervised methods
- Requires multiple biological samples from each class

Analysis Strategies for Class Comparisons

- Compare classes on a gene by gene basis using statistical tests
  - Control for the large number of tests performed
  - Types of statistical significance tests
    - t-tests or F-tests
    - permutation tests
    - pooled variance or shared variance t and F tests
    - Analysis of variance of log intensities
- Global tests

Controlling for Multiple Testing

- Bonferroni control of familywise error (FWE) rate at level α
  - 95% confident that FD=0
- Expected Number of False Discoveries – E(FD)
- Expected Proportion of False Discoveries – E(FDP)

*False discovery = declare gene as differentially expressed (reject test) when in truth it is not differentially expressed

Simple Procedures

- Control E(FD) $\leq u$
  - Conduct each of k tests at level u/k
  - e.g. To limit of 10 false discoveries in 10,000 comparisons, conduct each test at p<0.001 level
- Control E(FDP) $\leq \gamma$
  - FDR procedure

Controlling the Expected False Discovery Proportion

- Compare classes separately by gene and compute significance levels
- Rank genes in order of significance
  - $P_{(1)} < P_{(2)} < \ldots < P_{(N)}$
- Find largest index i for which
  - $P_{(i)}N / i \leq \mathrm{FDR}$
- Consider genes with the i'th smallest P values as statistically significant

Control of the Probability FDR < γ

- $G_i(k)$ = permutation estimate of the probability of $\leq$ k genes with $p \leq P_{(i)}$
- $\Pr(FDR \leq k/i) \cong G_i(k)$
- Select smallest i for which $G_i(\gamma i) < \alpha$
- Include in gene list those with (i-1)st smallest p values

Global Test
Do Expression Profiles Of Melanoma Lesions Differ By Response to Vaccine Rx

- 13 metastatic lesions with cr to rx
- 21 other pre-rx lesions
- 6108 gene microarrays
- 18 genes with $p<0.001$ by t test
- < 5% of 10,000 permutations give 18 or more genes with $p<0.001$ by t test

Sample Size Planning

GOAL: Identify genes differentially expressed in a comparison of pre-defined classes of specimens on two-color arrays using reference design

- Total sample size when comparing two equal sized, independent groups:

$$n = 4\sigma^2(z_{\alpha/2} + z_\beta)^2/\delta^2$$

where δ = mean log-ratio difference between classes

σ = standard deviation

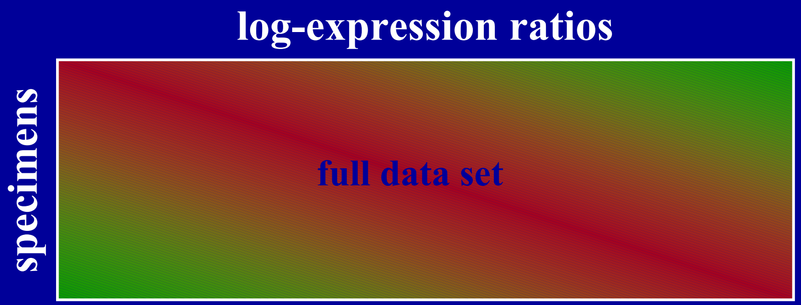
$z_{\alpha/2}, z_\beta$ = standard normal percentiles

- Choose α small, e.g. $\alpha = .001$

Class Prediction

- Predict membership of a specimen into pre-defined classes
  - mutation status
  - poor/good responders
  - long-term/short-term survival

Non-cross-validated Prediction



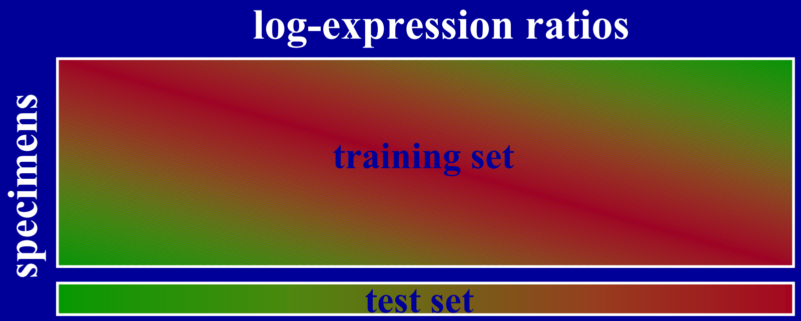
log-expression ratios

specimens

full data set

1. Prediction rule is built using full data set.
2. Rule is applied to each specimen for class prediction.

Cross-validated Prediction (Leave-one-out method)



log-expression ratios

specimens

training set

test set

1. Full data set is divided into training and test sets (test set contains 1 specimen).
2. Prediction rule is built **from scratch** using the training set.
3. Rule is applied to the specimen in the test set for class prediction.
4. Process is repeated until each specimen has appeared once in the test set.

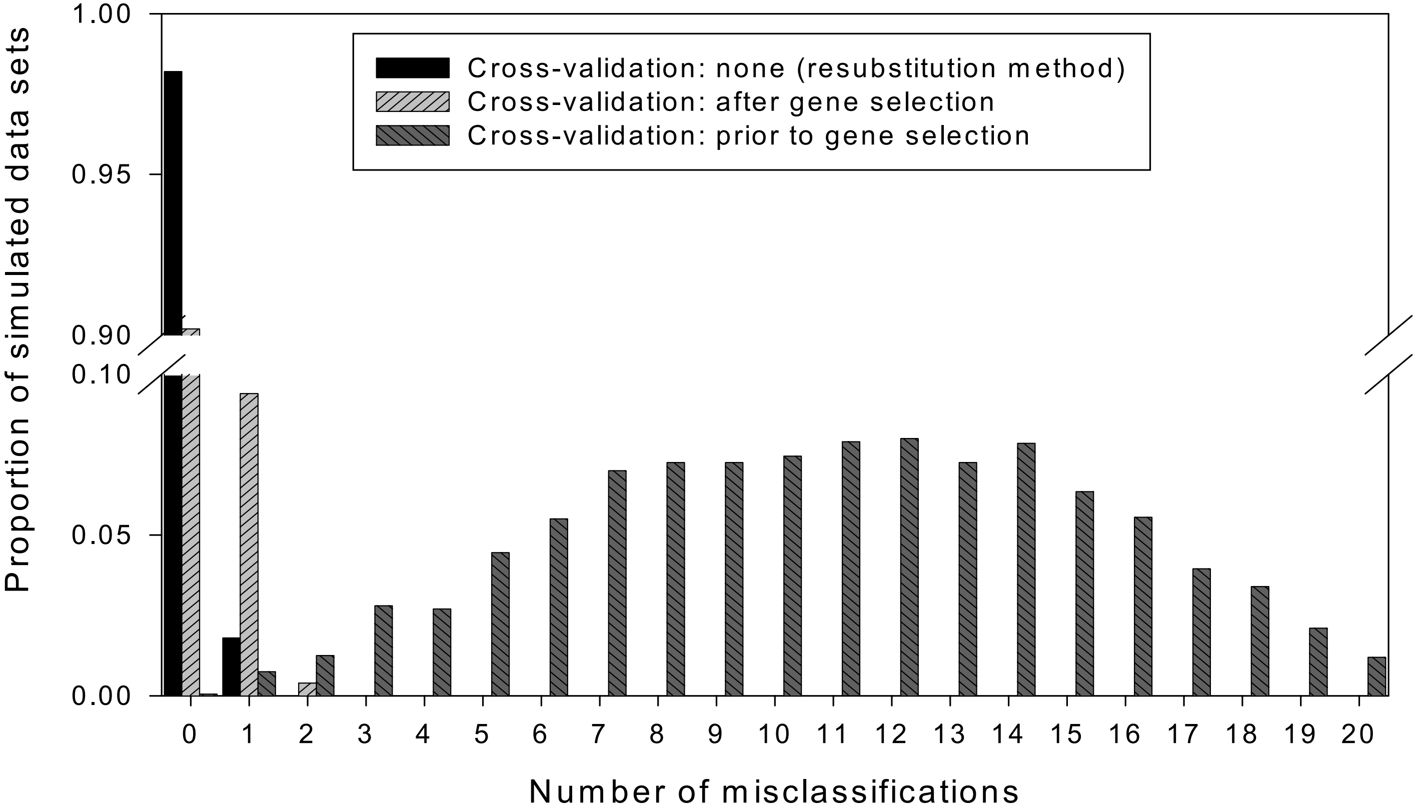
Prediction on Simulated Null Data

Generation of Gene Expression Profiles

- 14 specimens
- Log-ratio measurements on 6000 genes
- $\sim N(\mathbf{0}, \mathbf{I}_{6000})$ for all genes and all samples
- Can we distinguish between the first 7 specimens (Class 1) and the last 7 (Class 2)?

Prediction Method

- Compound covariate prediction
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.


Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Proportion of simulated data sets
1.00
0.95
0.90
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Number of misclassifications

Selection of a Class Prediction Method

"Note that when classifying samples, we are confronted with a problem that there are many more attributes (genes) than objects (samples) that we are trying to classify. This makes it always possible to find a perfect discriminator if we are not careful in restricting the complexity of the permitted classifiers. To avoid this problem we must look for very simple classifiers, compromising between simplicity and classification accuracy." (Brazma & Vilo, *FEBS Letters*, 2000)

Weighted voting method: distinguished between subtypes of human acute leukemia (Golub *et al.*, *Science*, 1999)

Support vector machines: classified ovarian tissue as normal or cancerous (Furey *et al.*, *Bioinformatics*, 2000)

Clustering-based classification: applied to above data sets and others (Ben-Dor *et al.*, *J Comput Biol*, 2000)

Compound covariate prediction: distinguished between mutation positive and negative breast cancers (Hedenfalk *et al.*, *NEJM*, 2001)

The Compound Covariate Predictor (CCP)

- We consider only genes that are differentially expressed between the two groups (using a two-sample *t*-test with small α).
- The CCP
  - Motivated by J. Tukey, *Controlled Clinical Trials*, 1993
  - Simple approach that may serve better than complex multivariate analysis
  - A compound covariate is built from the basic covariates (log-ratios)

$$\mathrm{CCP}_i = \sum_j t_j \, x_{ij}$$

t_j is the two-sample *t*-statistic for gene *j*.

x_{ij} is the log-ratio measure of sample *i* for gene *j*.

Sum is over all differentially expressed genes.

- Threshold of classification: midpoint of the CCP means for the two classes.

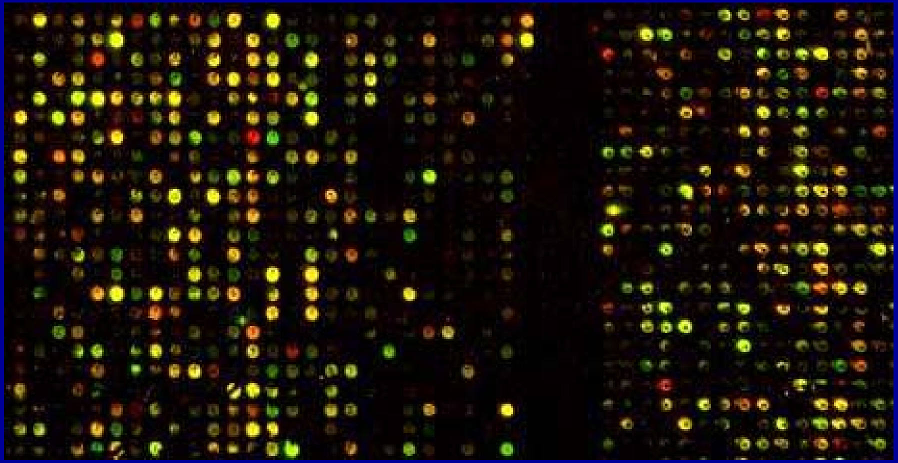
Advantages of Composite Variable Classifier

- Does not over-fit data
  - Incorporates influence of multiple variables without attempting to select the best small subset of variables
  - Does not attempt to model the multivariate interactions among the predictors and outcome
  - A one-dimensional classifier with contributions from variables correlated with outcome

Gene-Expression Profiles in Hereditary Breast Cancer

cDNA Microarrays

Parallel Gene Expression Analysis



- Breast tumors studied:
  - 7 *BRCA1*+ tumors
  - 8 *BRCA2*+ tumors
  - 7 sporadic tumors

- Log-ratios measurements of 3226 genes for each tumor after initial data filtering

RESEARCH QUESTION

Can we distinguish *BRCA1*+ from *BRCA1*– cancers and *BRCA2*+ from *BRCA2*– cancers based solely on their gene expression profiles?

Classification of hereditary breast cancers with the compound covariate predictor

Class labels	Number of differentially expressed genes	m = number of misclassifications	Proportion of random permutations with m or fewer misclassifications
BRCA1^{+} vs. *BRCA1*$^{-}$	9	1 (0 *BRCA1*$^{+}$, 1 *BRCA1*$^{-}$)	0.004
BRCA2^{+} vs. *BRCA2*$^{-}$	11	4 (3 *BRCA2*$^{+}$, 1 *BRCA2*$^{-}$)	0.043

Linear Methods of Class Prediction

- Compound covariate predictor
- Gollub’s weighted voting method
- Diagonal linear discriminant analysis
- Linear support vector machines
- Perceptrons with linear transfer functions and principal component inputs

Comparison of discrimination methods

Speed et al

In this field many people are inventing new methods of classification or using quite complex ones (e.g. SVMs). Is this necessary?

We did a study comparing several methods on three publicly available tumor data sets: the Leukemia data set, the Lymphoma data set, and the NIH 60 tumor cell line data, as well as some unpublished data sets.

We compared NN, FLDA, DLDA, DQDA and CART, the last with or without aggregation (bagging or boosting).

The results were unequivocal: simplest is best!

BRB ArrayTools: An integrated Package for the Analysis of DNA Microarray Data Created by Statisticians for Biologists

http://linus.nci.nih.gov/BRB-ArrayTools.html

BRB ArrayTools

- Based on the experience of Biometric Research Branch staff in analyzing microarray studies and developing methodology for the design and analysis of such studies
- Packaged to be easy to use by biologists

DNA Microarrays and Clinical Trials

- Requires tissue from target organ
  - Gene expression is organ specific
    - blood sample insufficient
  - Impractical for adverse reaction screening
  - Inconvenient and expensive for response screening even for life threatening diseases
  - RNA based so careful specimen handling is necessary

DNA Microarrays and Clinical Trials

- Powerful technology for elucidating mechanisms, identifying targets and treatment effects and for developing therapeutically relevant diagnostic classifications
- Translation of classifications to clinic may require translation to protein based platform

DNA Microarrays and Clinical Trials

- Because of problems of multiplicity and problems of poor analysis, findings require independent confirmation

Two Phase Strategy for Using Microarrays in Clinical Trials

- Develop gene expression based model for predicting response to treatment T using phase II trial data
- Screen patients for eligibility to randomized phase III trial comparing treatment T to placebo or no treatment based on response prediction model

Sample Size Planning

- Total sample size for phase III trial with continuous response endpoint

$$n = 4(z_{\alpha/2} + z_{\beta})^2/(\delta/\ \sigma\ )^2$$

where δ = target difference in mean outcome between treatment and control group

σ = within group standard deviation

$z_{\alpha/2}$, z_{β} = standard normal percentiles e.g. 1.96, 1.28

Effect of Pharmacogenetic Screening on Sample Size

- σ decreases
- δ increases
- 2 fold increase in $\delta/\sigma \Rightarrow 4$ fold reduction in n
- 1.5 fold increase in $\delta/\sigma \Rightarrow 2.25$ fold

  reduction in n
- number of patients screened may not be reduced, but therapeutic benefit ratio for treated patients should be enhanced

Collaborators

- Molecular Statistics & Bioinformatics, NCI
  - Kevin Dobbin
  - Lisa McShane
  - Amy Peng
  - Michael Radmacher
  - Joanna Shih
  - George Wright
  - Yingdong Zhao