Design Issues in DNA Microarray Based Studies

Richard Simon, D.Sc.
Chief, Biometric Research Branch
National Cancer Institute
rsimon@nih.gov
http://linus.nci.nih.gov/~brb

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?

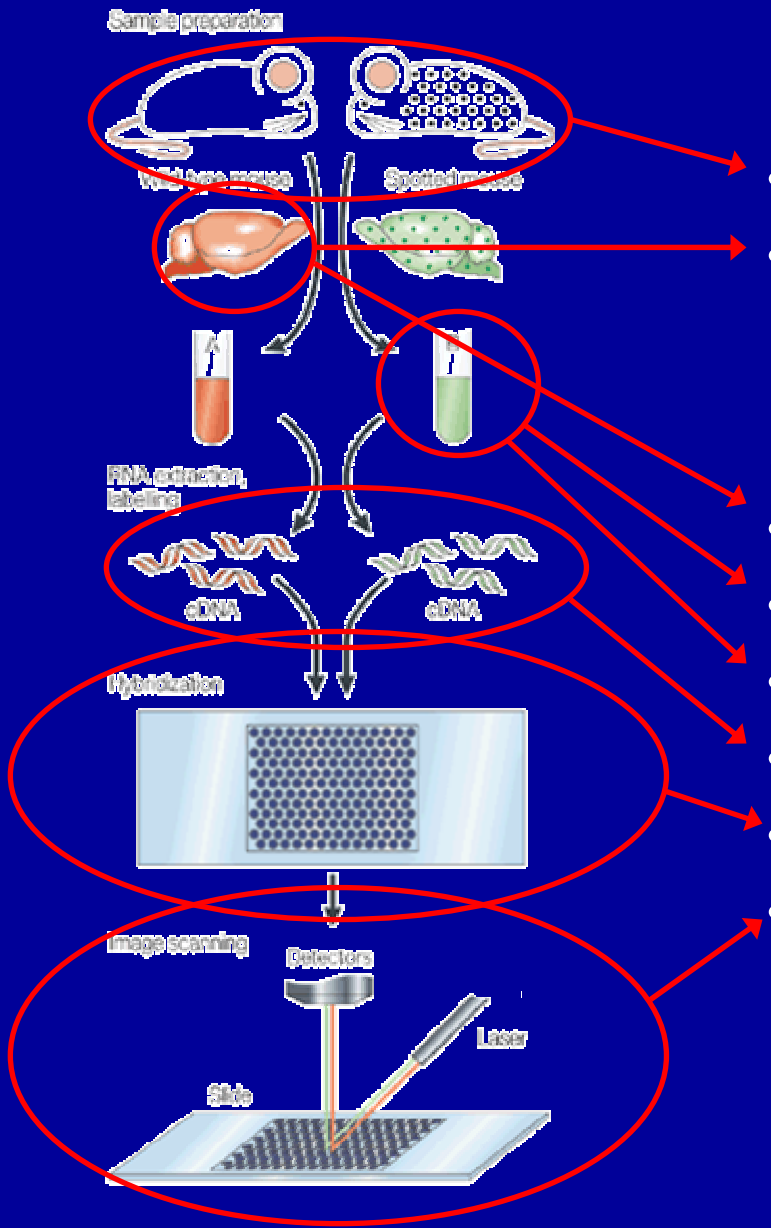
- Not a clustering problem
  - Global similarity measures generally used for clustering arrays may not distinguish classes
- Supervised methods
- Requires multiple biological samples from each class

Class comparison requires multiple biological samples from each class

- Comparing two RNA samples is not the same as comparing two tissues or two biological conditions
- Some statisticians and software producers forget this
  - Formulas for comparing red intensity to green intensity on one cDNA array
  - Methods for comparing intensities on two GeneChips™

How many replicates should I do?

Sources of Variability
(cDNA Array Example)



- Biological Heterogeneity in Population
- Specimen Collection/ Handling Effects
  - Tumor: surgical bx, FNA
  - Cell Line: culture condition, confluence level
- Biological Heterogeneity in Specimen
- RNA extraction
- RNA amplification
- Fluor labeling
- Hybridization
- Scanning
  - PMT voltage
  - laser power

(Geschwind, *Nature Reviews Neuroscience*, 2001)

Levels of Replication

- RNA sample divided into multiple aliquots
- Multiple RNA samples from a specimen
- Multiple subjects from population(s)

Levels of Replication

- For comparing classes, replication of samples should generally be at the “subject” level because we want to make inference to the population of “subjects”, not to the population of sub-samples of a single biological specimen.

Can I do just one cDNA array if I pool specimens?

- Pooling does not average systematic differences in experimental procedures that come after pooling; e.g. labeling and hybridization
- Inference is limited to the specific RNA pools, not to the populations since there is no estimate of variation among pools
- Statistical inference even to compare the specific pools requires replicate labeling and hybridization of the RNA pools

Analysis Strategies for Class Comparisons

- Compare classes on a gene by gene basis using statistical tests
  - Control for the large number of tests performed

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) ≤ 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) ≤ γ
  - 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
  - $P_{(i)} N / i \leq FDR$
- Consider genes with the i'th smallest P values as statistically significant

Step-down Permutation Procedures (Korn *et al.*, 2001)

Want procedures to allow statements like:

FD Procedure: "We are 95% confident that the (actual) number of false discoveries is no greater than 2."

FDP Procedure: "We are 95% confident that the (actual) proportion of false discoveries does not exceed .10."

Step-down Permutation Procedures

- "Step-down"
  - Sequential testing (smallest to largest p-value), adjusting critical values as you go
  - Less conservative than uniform critical value methods
- Permutation-based
  - Independent of distribution
  - Preserve/exploit correlation among tests by permuting each profile *as a unit*

FD Algorithm

To be (1-α)100% confident that the (actual) number of false discoveries is ≤ *u*:

- Automatically reject $H_{(1)}, H_{(2)}, \ldots, H_{(u)}$.
- For $r > u$, having rejected $H_{(r-1)}$, reject $H_{(r)}$ if $P_{(r)} < y(\alpha)_u$
- $y(\alpha)_u$ = α quantile of the permutation distribution of the (u+1)st smallest p value
- Once a hypothesis is not rejected, all further hypotheses are not rejected.

Notes

- FD procedure with $u = 0$ reduces to step-down FWE procedure (Westfall and Young, 1993)
- Ties and missing data can be handled
- Takes advantage of correlation structure of data
- Particularly useful with small sample size
- Computationally intensive
  - Included in BRB-ArrayTools
- Allowing a few errors may buy a lot in power to detect “true discoveries”

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

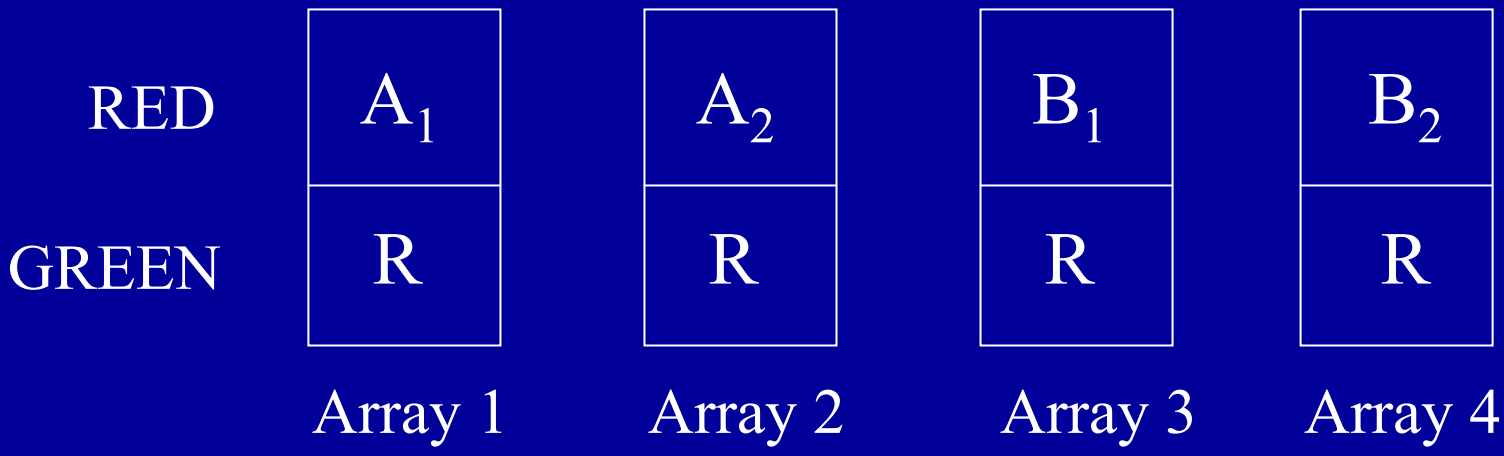
Notes

- Proof of FDP procedure requires asymptotic arguments, so control is only approximate for small samples
- Takes advantage of correlation structure of data
- Particularly useful with small sample size
- Ties and missing data can be handled
- Computationally intensive
  - Included in BRB-ArrayTools
- Allowing a small proportion of errors may buy a lot in power to detect “true discoveries”

Class Comparison: Allocation of Specimens to cDNA Array Experiments

- Reference Design
- Balanced Block Design
- Loop Design
  - Kerr and Churchill

Reference Design

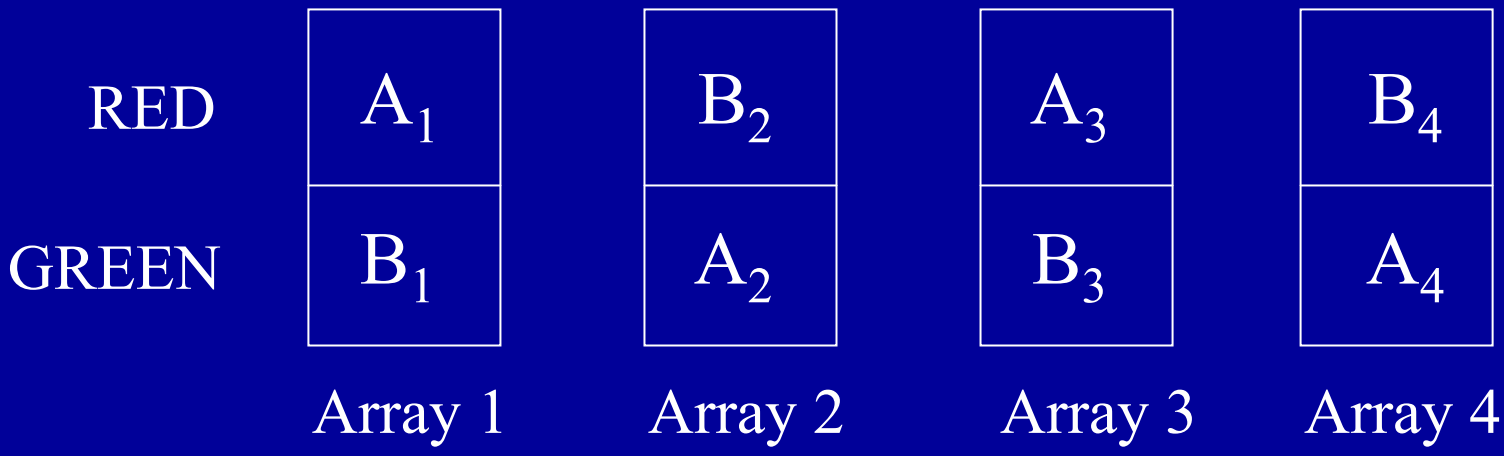


A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

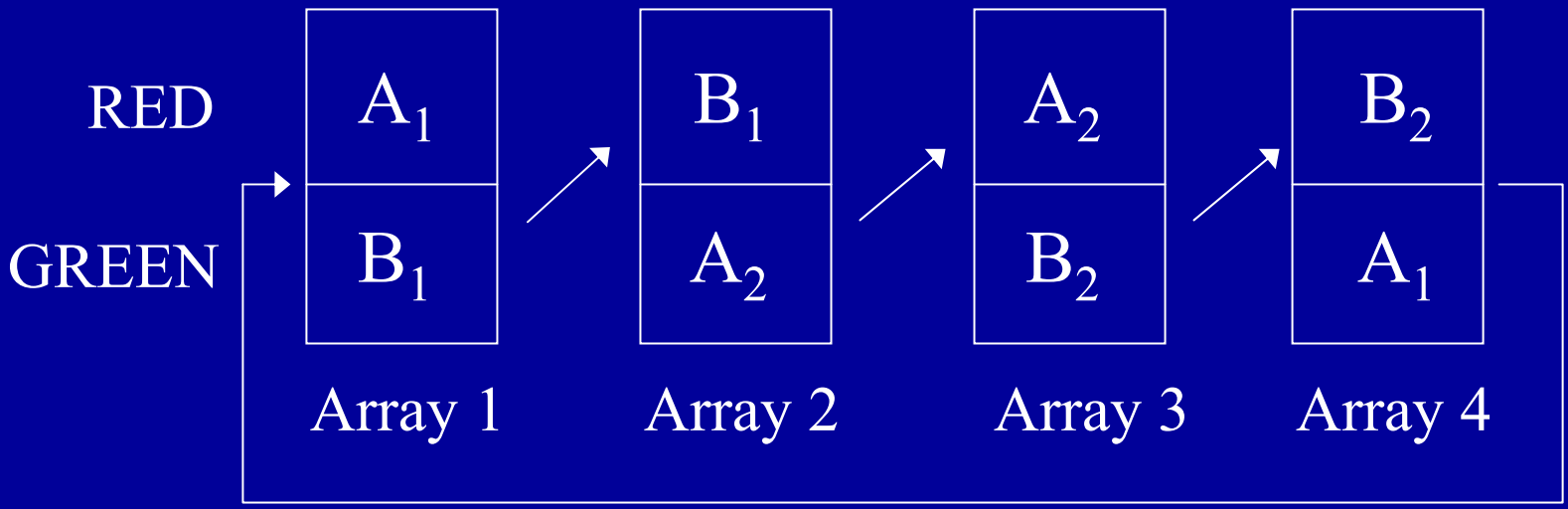
Balanced Block Design



A_i = ith specimen from class A

B_i = ith specimen from class B

Loop Design



A_i = aliquot from *i*th specimen from class A

B_i = aliquot from *i*th specimen from class B

(Requires two aliquots per specimen)

Model-based Methods for Analysis of cDNA Array Data:
ANOVA for Logarithm of Background Adjusted Intensities

- First Stage Normalization Model
  - Array
  - Dye
  - Array * Dye
  - Variety (Class)
  - Sample within variety

ANOVA for Logarithm of Background Adjusted Intensities

- Gene-Variety Second Stage Models Fitted to Residuals from Normalization Model
  - Gene
  - Array by Gene
  - Variety by Gene
  - Sample within Variety by Gene
- Gene-Variety Models Fitted Separately by Gene

Gene-Variety Model

- $r = G_g + AG_{ag} + VG_{vg} + SG_{sg} + \varepsilon$
- $\varepsilon \sim N(0,\sigma_g^2)$
- Efficiency of design based on variance of estimators of VG_{ig}-VG_{jg}
- To study efficiency, assume $SG_{sg} \sim N(\mu_g,\tau_g^2)$

Comparison of Designs

- For class discovery, a **Reference** design is preferable because of large gains in cluster performance.
- For class comparisons . . .
  - With a fixed number of arrays, **Block** design is more efficient than **Loop** or **Reference** design
    - Block design precludes clustering
    - Block designs focus on single type of comparison
  - With a fixed number of specimens, **Reference** design is more efficient than **Loop** or **Block** design when intra-class variance is “large”.

When Should You Do Reverse-Label Pairs of 2-color arrays?

- Always?
- Never?
- Sometimes?

Dye Bias

- Average differences among dyes in label concentration, labeling efficiency, photon emission efficiency and photon detection are corrected by normalization procedures
- Gene specific dye bias may not be corrected by normalization

Gene Specific Dye Bias

- Gene specific dye bias is common in my experience
- Gene specific dye bias is generally of very small magnitude

Comparison of Classes of Samples Using A Reference Design

- Consistently label reference with same label
- Reverse label pairs are not needed
- Gene specific dye bias does not bias class comparisons since the label bias applies equally to all classes
- Power for detecting class differences could be effected if dye bias is severe

Comparison of Two Classes of Samples Using A Block Design

- Paired comparisons
  - Tumor vs normal tissue from same patient
  - BRCA1 mutated vs BRCA1 non mutated paired by stage of disease
- Unpaired comparisons randomly blocked onto arrays
- The most efficient design is a balanced block design with no reverse arrays of the same two specimens

Comparison of Experimental Specimens to Internal Reference Using a Reference Design

- Comparison to pooled normal tissue reference
  - Inference limited to that pool
- Some reverse label array pairs are necessary to estimate dye bias; e.g. 5-10 pairs. Rest of arrays should be consistently labeled
- ANOVA model based comparison of specimen averages to internal reference pool adjusted for dye bias

Dobbin, Shih, Simon ANOVA

$$r_{gadvf} = G_g + GA_{ga} + GD_{gd} + GV_{gv} + GF_{gf} + \varepsilon$$

r is background adjusted, normalized intensity

gene *g*

array *a*

dye *d*

variety *v* (0=ref, 1=experimental)

individual *f*

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_{β} = standard normal percentiles

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

- π = proportion of genes on array that are differentially expressed between classes
- N = number of genes on the array
- FD = expected number of false discoveries
- TD = expected number of true discoveries
- FDR = FD/(FD+TD)

- $FD = \alpha(1-\pi)N$
- $TD = (1-\beta)\ \pi N$
- $FDR = \alpha(1-\pi)N/\{\alpha(1-\pi)N + (1-\beta)\ \pi N\}$
- $= 1/\{1 + (1-\beta)\pi/\alpha(1-\pi)\}$

Controlling Expected False Discovery Rate

π	α	β	FDR
0.01	0.001	0.10	9.9%
	0.005		35.5%
0.05	0.001		2.1%
	0.005		9.5%

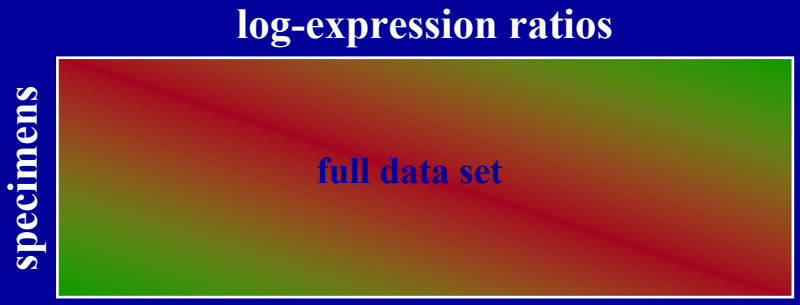
Total Number of Samples for Two Class Comparison

α	β	δ	σ	Total Samples
0.001	0.05	1 (2-fold)	0.5 human tissue	25
			0.25 transgenic mice	12 (t approximation)

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



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)

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.

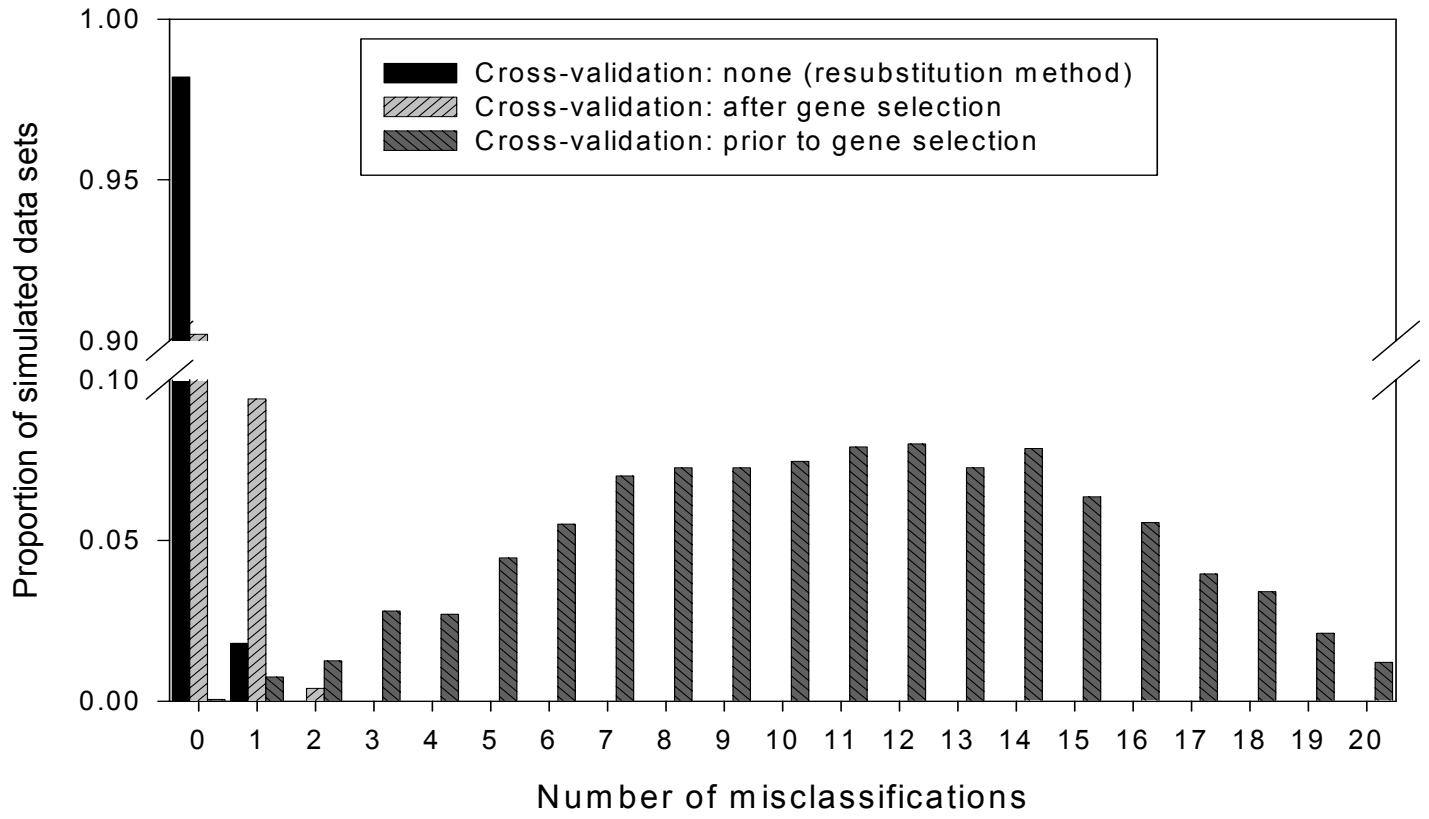
Prediction on Simulated Null Data

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.

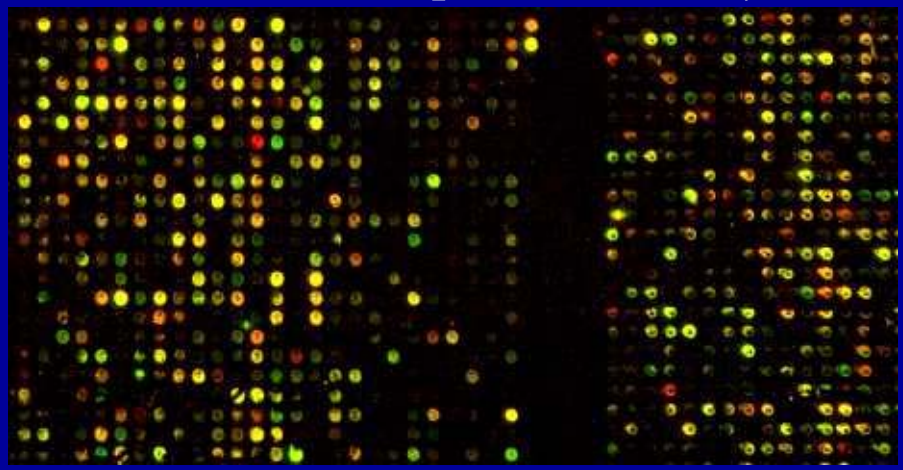
Advantages of Composite Variable Classifier

- 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 kernel 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

References

- Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics (In Press).
- Dobbin K, Shih J, Simon R. Statistical design of reverse dye microarrays. Submitted for publication.
- Simon R, Radmacher MD, Dobbin K. Design of studies with DNA microarrays. Genetic Epidemiology 23:21-36, 2002.
- Radmacher MD, McShane LM and Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511, 2002.

References

- McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods of assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics (In Press).
- Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: Class prediction methods. Journal of the National Cancer Institute. (In Press)
- Korn EL, McShane LM, Troendle JF, Rosenwald A and Simon R. Identifying pre-post chemotherapy differences in gene expression in breast tumors: a statistical method appropriate for this aim. British Journal of Cancer 86:1093-1096, 2002.

References

- Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: Application to high dimensional genomic data. Submitted for publication.

- Wright G, Simon R. A hierarchical random variance model for the analysis of DNA microarray data. Submitted for publication.

References (Applications)

Hedenfalk I, et al. Gene expression profiles of hereditary breast cancer. New England Journal of Medicine 344:549, February 22, 2001.

Bittner M, et al. Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. Nature 406:536-540, 2000.

Lymphoma/Leukemia Molecular Profiling Project. The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large B-cell Lymphoma. New Engl J Med 346:1937-47.

Collaborators

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