# The State of the Art of Microarray Informatics

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

# Design and Analysis of Microarray Experiments

- State of Practice
- State of Knowledge Base

# 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 without filling the literature with spurious results
- 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

# Finding Genes Differentially Expressed in Red vs Green Channels

- Many methods address significance of ratio on one array
- Many methods address significance of difference in intensities between two arrays (e.g. Affymetrix software)
- Many methods address significance of difference in expression levels between two rna samples
- **These are generally not the biologically meaningful questions**

# It is Important to Distinguish Among 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.

# Supervised Methods Are Better Than Unsupervised Methods (Cluster Analysis) For Class Comparison and Prediction

# 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

# 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

# Multiple testing procedures:
## Identifying differentially expressed genes while controlling for false discoveries*

- *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) ≤ $\gamma$
  - FDR procedure
- Bonferroni control of familywise error (FWE) rate at level $\alpha$
  - Conduct each of k tests at level $\alpha$/k
  - At least $(1-\alpha)100\%$ confident that FD = 0

# Controlling the False Discovery Rate

- 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

# Additional Procedures

- "SAM" - Significance Analysis of Microarrays
  - Tusher *et al.*, *PNAS,* 2001
  - Estimate FDR
  - Statistical properties unclear
- Empirical Bayes
  - Efron *et al.*, *JASA,* 2001
  - Related to FDR
- Step-down permutation procedures
  - Korn *et al.,* 2001 (http://linus.nci.nih.gov/~brb)
  - Control number or proportion of false discoveries

# 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 $\delta$ = mean log-ratio difference between classes

$\sigma$ = standard deviation

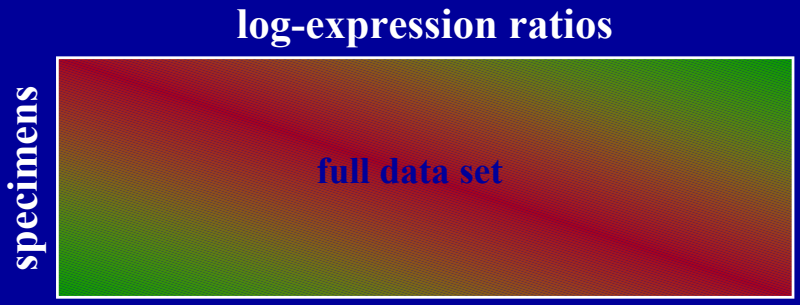
$z_{\alpha/2}$, $z_{\beta}$ = standard normal percentiles

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

# Class Prediction

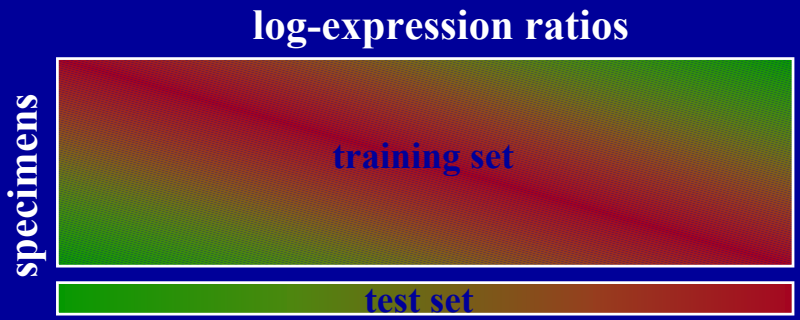
- Predict membership of a specimen into pre-defined classes
  - mutation status
  - poor/good responders
  - long-term/short-term survival

# Non-cross-validated Prediction




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.

Percentage of simulated data sets with *m* or fewer misclassifications

| *m* | Non-cross-validated class prediction | Cross-validated class prediction |
|---|---|---|
| 0 | 99.85 | 0.60 |
| 1 | 100.00 | 2.70 |
| 2 | 100.00 | 6.20 |
| 3 | 100.00 | 11.20 |
| 4 | 100.00 | 16.90 |
| 5 | 100.00 | 24.25 |
| 6 | 100.00 | 34.00 |
| 7 | 100.00 | 42.55 |
| 8 | 100.00 | 53.85 |
| 9 | 100.00 | 63.60 |
| 10 | 100.00 | 74.55 |
| 11 | 100.00 | 83.50 |
| 12 | 100.00 | 91.15 |
| 13 | 100.00 | 96.85 |
| 14 | 100.00 | 100.00 |

# 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 $\alpha$).
- 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.

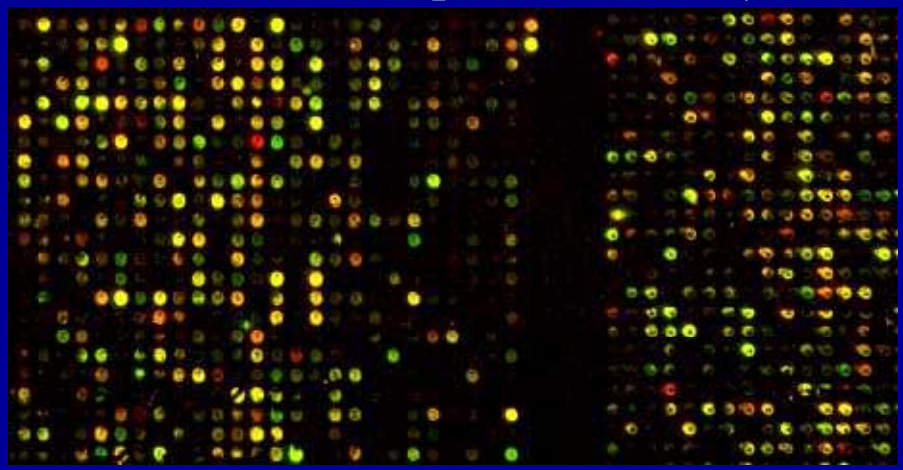
# Advantages of Composite Variable Classifier

- 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 |

# Composite Variable Classifier
# CLL Mutational Status
# 18 Samples

| Nominal Alpha | Number of DEGs | X-validation Errors | Permutation p value | Misclassific's in 10 new samples |
|---|---|---|---|---|
| 0.001 | 56 | 1 | 0.001 | 1 |
| 0.0001 | 7 | 5 | 0.107 | 1 |

# Quadratic Discriminant Analysis

- Assumes that log-ratios (log intensities) have a multi-variate Gaussian distribution.
- The two classes have different mean vectors and potentially different covariance matrices.
- Using the training data, estimate the mean vector and covariance matrix for each class.

# Diagonal Linear Discriminant Analysis

- Full QDA performs poorly when G >N. One can help somewhat by selecting the G genes to include based on univariate discrimination power.
- The number of parameters can be dramatically reduced by assuming that the variances are the same in the two classes and that covariance among genes can be ignored. This reduces the number of parameters to 3G. This is DLDA. It has performed as well as much more complex methods in comparisons conducted by Dudoit et al.

# Diagonal Linear Discriminant Analysis

- Golub's Weighted Voting Method and Radmacher et al's Compound Variable Predictor are similar to DLDA.
- These methods, as well as other, are generally implemented with feature (gene) selection based on univariate classification power. In performing cross-validation to estimate mis-classification rate, the gene selection step **must** be repeated starting with the full set of genes for each leave-one-out training set.

# Neural Network Classification

Kahn et al. Nature Med. 2001

- A perceptron with no hidden nodes and a linear transfer function at each node.
- Inputs are first 10 principal components
  - The linear combinations of the genes that have greatest variation among samples and are orthogonal
- The method is essentially equivalent to DLDA based on the 10 PC's as predictors
- Authors didn't cross-validate the computation of the 10 PC's.

# 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!

# Cluster Analysis of Samples

- For discovering unanticipated structure and subsets of tissues

# Cluster Analysis is Subjective

- Cluster algorithms always produce clusters
- Different distance metrics and clustering algorithms may find different structure using the same data.

# Establishing That a "Disease" Can Be Molecularly Dissected into Sub-Diseases Requires More Than The Existence of Clusters

- Reproducibility with different clustering methods
- Statistical significance of clusters
- Reproducibility of clusters under data perturbations
- Reproducibility of clusters with separate rna samples of the same tissues

# Evaluating Clusters of Samples

- Assessing statistical significance of clusters
  - McShane *et al*., Bioinformatics (In Press) http://linus.nci.nih.gov/~brb
- Reproducibility of clusters under perturbations
  - McShane *et al*.
  - Kerr and Churchill (*PNAS*, 2001)
- Estimating number of clusters
  - Tibshirani *et al*., *JRSS B*, 2002

# State of the Art of Microarray Bioinformatics

- Software that incorporates valid, published, statistical methods and that encourages good experimental design

# 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

# Collaborators

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