Development and Validation of Gene Expression Based Diagnostic Classification

Richard Simon

National Cancer Institute

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

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

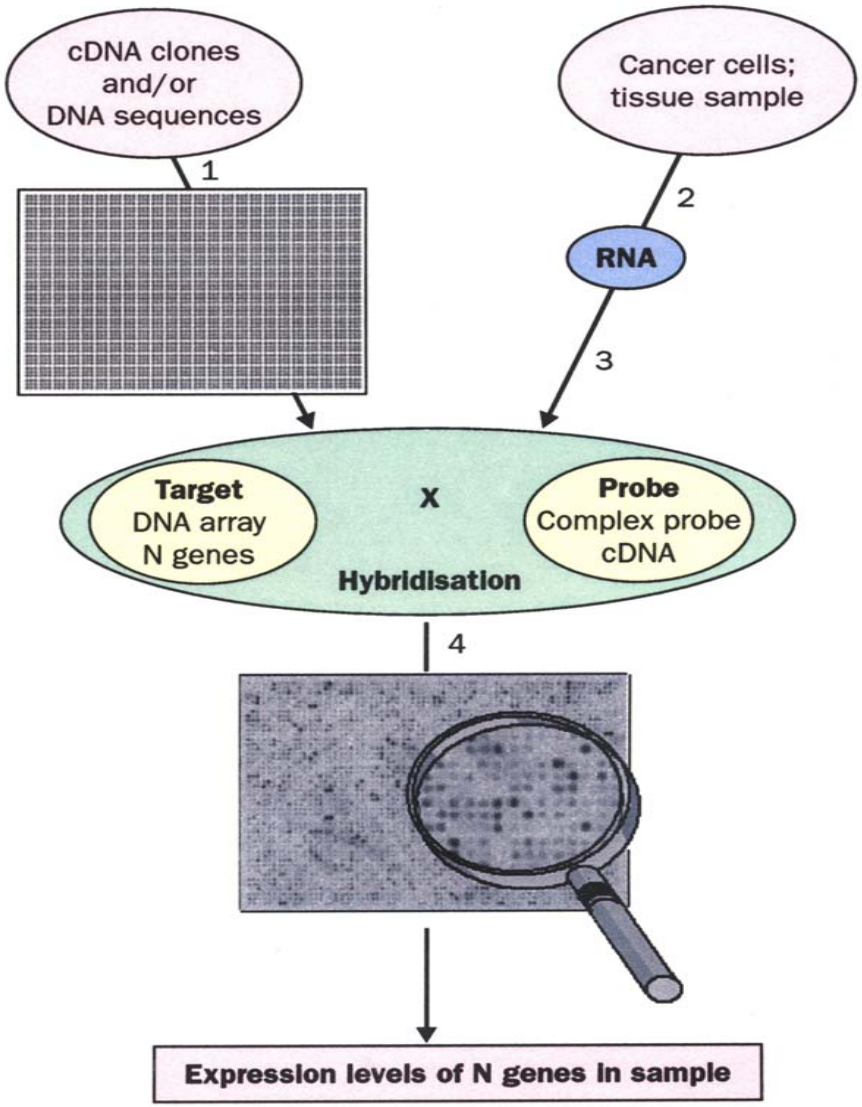
- Powerpoint presentation
- Reprints & Technical Reports
  - Myths & Truths
    About Microarray Expression Profiling
- BRB-ArrayTools software
  - Performs all analyses described

BRB ArrayTools Design Objectives

- Encapsulates BRB experience in analysis of data and development of methods
- Educating biologists in microarray data analysis
- Easy user interface
  - Excel front-end
- Ease of data loading
  - integrated
- Drill-down linkage to genomic databases
- State-of-the-art analytic tools
  - Based on BRB critically evaluating literature
- Easily extensible
  - R add-ins
- Portable
  - Non-proprietary
  - Free for non-commercial use

- *Design and Analysis of DNA Microarray Investigations*
  - R Simon, EL Korn, MD Radmacher, L McShane, G Wright, Y Zhao. Springer (2003)

An Overview of Microarrays



1. **Spotting of cDNA clones or oligonucleotides on solid support**
2. **Extraction of RNA from specimen**
3. **Reverse transcription and labeling**
4. **Scanning of hybridization Quantification of signal intensity**
5. **Data analysis**

Lancet Oncol. 2:674, 2001.

Myth

- That the greatest challenge is managing the mass of microarray data
- Greater challenges are:
  - Effectively designing and properly analyzing experiments that utilize microarray technology
    - Distinguishing hype and misinformation from sound methodology
    - Avoiding software developed by individuals with no qualifications for determining valid methodology
  - Organizing and facilitating effective interdisciplinary collaboration with statisticians, clinicians & biologists

Myth

- That data mining is an appropriate paradigm for analysis of microarray data
  - find interesting patterns that give clear answers to questions that were never asked
- That planning microarray investigations does not require “hypotheses” or clear objectives

- Good microarray studies have clear objectives, but not generally gene specific mechanistic hypotheses

Design and Analysis Methods Should Be Tailored to Study Objectives

- Class Comparison
  - For predetermined classes, establish whether gene expression profiles differ, and identify genes responsible for differences
- Class Prediction
  - Prediction of predetermined class (phenotype) using information from gene expression profile
    - Response vs non-response
    - 5-yr RFS vs relapse
- Class Discovery
  - Discover clusters among specimens or among genes

- Cluster analysis is appropriate only for class discovery
  - Cluster analysis is subjective and always produces clusters
  - Cluster analysis is frequently used in a misleading way
- *Supervised methods* are more appropriate for class comparison and class prediction

Erroneous Use of Cluster Analysis

- In comparing two classes of specimens which are actually the same with regard to gene expression profiles except for random variation
  - Arrays with 10,000 genes represented
  - Number of genes whose mean expression differs between the classes to an extent that is statistically significant at $p<0.05$
    - 500 (false positives)
- If you cluster the samples with regard to these 500 false positive genes you will get good separation of the classes

Myth

- That comparing tissues or experimental conditions is based on looking for red or green spots on a single array or by comparing results from two arrays

- Comparing expression in two RNA samples tells you (at most) only about those two samples and may relate more to sample handling and assay artifacts than to biology. Robust knowledge requires multiple samples that reflect biological variability.
- Many software packages and algorithms provide inappropriate tools for data analysis

Components of Class Prediction

- Feature (gene) selection
  - Which genes will be included in the model
- Select model type
  - E.g. Diagonal linear discriminant analysis, Nearest-Neighbor, …
- Fitting parameters (regression coefficients) for model

Feature Selection

- Genes that are differentially expressed among the classes at a significance level α (e.g. 0.01)
  - The α level is selected to control the number of genes in the model

Myth

- Complex classification algorithms such as neural networks perform better than simpler methods for class prediction.

- Artificial intelligence sells to journal reviewers and peers who cannot distinguish hype from substance when it comes to microarray data analysis.
- Comparative studies have shown that simpler methods work as well or better for microarray problems because they avoid overfitting the data.

Effective Simple Classifiers for Microarray Studies

- Linear classifiers on informative genes
  - Diagonal Linear Discriminant Analysis
  - Compound Covariate Predictor
  - Golub's Weighted Voting Classifier
  - Support Vector Machines with inner-product kernel
- Nearest neighbor or nearest centroid classifiers on informative genes

- Fitting complex classifiers to training data results in unstable models unless the training data set is huge
- Unstable models have large error rates for independent data

Evaluating a Classifier

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data.
- When the number of candidate predictors (p) exceeds the number of cases (n), perfect prediction on the same data used to create the predictor is always possible

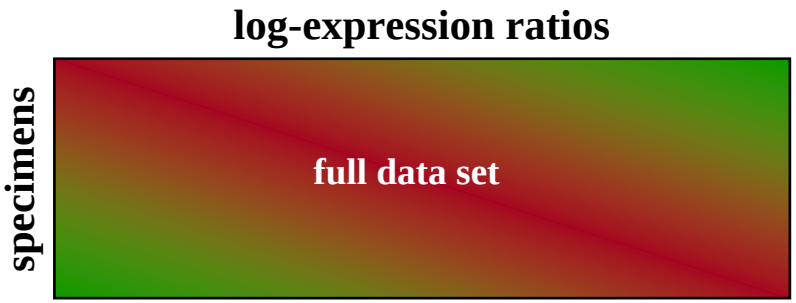
Validation of a Predictor

- In-study validation
  - Re-substitution estimate
    - Horribly biased
  - Split-sample validation
    - Often applied with too small a validation set
  - Cross-validation
    - Often mis-used
- Independent data validation

Split-Sample Evaluation

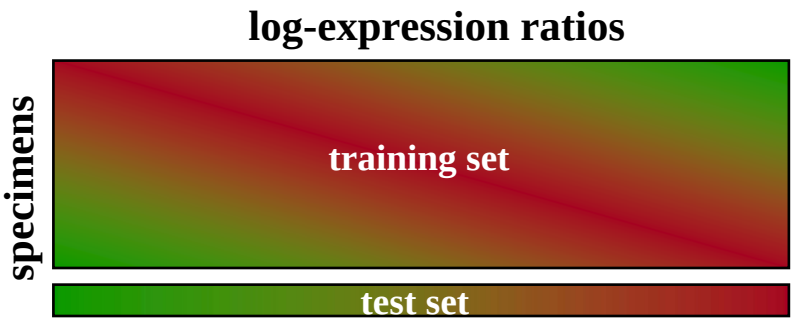
- Used for Rosenwald et al. study of prognosis in DLBL lymphoma.
  - 200 cases training-set
  - 100 cases test-set

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)



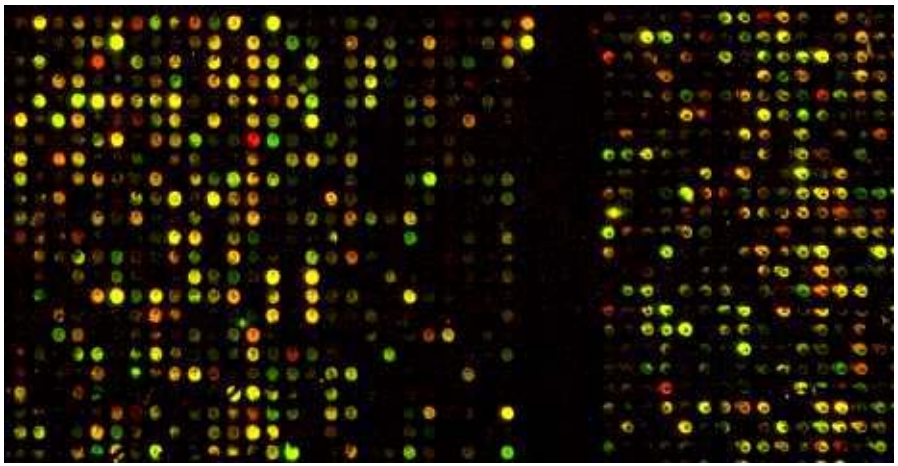
1. Full data set is divided into training and test sets (test set contains 1 specimen).
2. Prediction rule is built from scratch using the training set.
3. Rule is applied to the specimen in the test set for class prediction.
4. Process is repeated until each specimen has appeared once in the test set.

- Cross validation is only valid if the test set is not used in any way in the development of the model. Using the complete set of samples to select genes violates this assumption and invalidates cross-validation.
- With proper cross-validation, the model must be developed from scratch for each leave-one-out training set. This means that feature selection must be repeated for each leave-one-out training set.
- The cross-validated estimate of misclassification error is an estimate of the prediction error for model fit using specified algorithm to full dataset

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

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

BRCA1

α_g	# of significant genes	# of misclassified samples (m)	% of random permutations with *m* or fewer misclassifications
10^{-2}	182	3	0.4
10^{-3}	53	2	1.0
10^{-4}	9	1	0.2

BRCA2

α_g	# of significant genes	m = # of misclassified elements (misclassified samples)	% of random permutations with m or fewer misclassifications
10^{-2}	212	4 (s11900, s14486, s14572, s14324)	0.8
10^{-3}	49	3 (s11900, s14486, s14324)	2.2
10^{-4}	11	4 (s11900, s14486, s14616, s14324)	6.6

Sources of Bias in Estimation of Error Rates

- Confounding by sample handling or assay effects
  - Cases collected and assayed at different times than controls
- Failure to incorporate important sources of future variability
  - Assay drift
- Change in distribution of un-modeled variables

Independent Data Validation

- From different clinical centers
- Specimens assayed at different time from training data
- Positive and negative samples collected in the same way
- Study sufficiently large to give precise estimate of sensitivity and specificity of the multivariate classifier
- The validation study is prospectively planned
  - patient selection pre-specified to address a therapeutically relevant question
  - endpoints and hypotheses pre-specified
  - predictor fully pre-specified
  - Study addresses assay reproducibility
  - Specimens may be either prospective or archived

Applications of Microarrays To Clinical Trials

Many Prognostic Factor Studies

- Have no impact
- Are not reproducible

To Have Impact Diagnostic Categories Should Be

- Therapeutically relevant
- Biologically plausible

Traditional Approach for Marker Development

- Focus on candidate protein involved in disease pathogenesis
- Develop assay
- Conduct retrospective study of whether marker is prognostic using available specimens
- Marker dies because
  - Therapeutic relevance not established
  - Adequate validation study not performed
  - Inter-laboratory reproducibility not established

Using DNA Microarrays to Select Patients for Phase III Trials

- Perform microarray gene expression profiling on patients in phase II trials of new drug E
- Develop gene expression based predictor of responsiveness to E
- Select patients for phase III trial based on predicted responsiveness to E

Randomized Clinical Trials Targeted to Patients Predicted to be Responsive to the New Treatment Can Be Much More Efficient than Traditional Untargeted Designs

- For a drug like Errisa in lung cancer
  - 10% response rate
  - Traditional untargeted designs are hopelessly inefficient, even with 1000 patients randomized
  - More effort should be placed in finding predictors of response based on phase II data
    - Sequencing key genes
    - Expression profiling

- For Herceptin, even a relatively poor assay enabled conduct of a targeted phase III trial which was crucial for establishing effectiveness
- In many cases, the assay based on the presumed mechanism of action will not correlate with response and it may be more effective to let the data develop the assay via expression profiling

Genomic Approach to Diagnostic/Prognostic Marker Development

- Select therapeutically relevant population
  - Node negative, ER+, well staged breast cancer patients who have received Tam alone and have long follow-up
- Perform genome wide expression profiling
- Develop multi-gene/protein predictor of outcome
- Obtain unbiased within-sample estimate of prediction accuracy
- Adapt platform to clinical application
- Establish assay reproducibility
- Conduct prospective study to confirm results

Limitations of Genomic Approach

- It's difficult to identify real molecular targets from microarray profiling of human tumors
  - But it is very feasible to identify prognostic indicators or treatment selection indicators
- Availibility of fresh/frozen tissue for therapeutically relevant questions
  - Many studies address toy problems or heterogeneous non-therapeutically relevant populations
    - Comparing AML to ALL
    - Inclusion of N+, N-, ER+ and ER- patients

Limitations of Genomic Approach

- Difficulty in performing proper prospectively planned validation studies
- Lack of inter-laboratory reproducibility evaluations
- Applicability of a classifier that requires fresh tissue

Validation Study
Node negative Breast Cancer

- Prospective study design
- Samples collected and archived from patients with node negative ER+ breast cancer receiving TAM
- Apply single, fully specified multi-gene predictor of outcome to samples and categorize each patient as good or poor prognosis
- Are long-term outcomes for patients in good prognosis group sufficiently good to withhold chemotherapy?
- Are long-term outcomes for patients in poor prognosis group sufficiently poor to indicate chemotherapy?

Validation Study for Identifying Node Positive Patients Who Benefit from a Specific Regimen

- Standard treatment C
- New treatment E
- Predictor based on previous data for identifying patients who benefit from E but not C
- Randomized study of E vs C
- Measure markers on all patients
- Compare E vs C separately within groups predicted to benefit from E and those not predicted to benefit from E
- Two clinical trials worth of patients

Acknowledgements

- Kevin Dobbin
- Sudhir Varma
- Ed Korn
- Lisa McShane
- Michael Radmacher
- Joanna Shih
- George Wright
- Yingdong Zhao