(cDNA) Microarray Experimental Steps

(A) ‘Attach ’ single-strand DNA segments for each of $i = 1, \ldots, N$ (eg $= 1.e4$) Genes robotically to $N \times p$ distinct tiny ‘spots’ on treated membrane or glass slide (eg $p = 30$). These spots are grouped into ‘grids’ or ‘arrays’ sharing common ‘print-tip’ run.

(B) Extract tissue samples from one or two sources (eg healthy and cancerous organ cells); purify mRNA from each.

(C) Reverse-transcribe and copy (‘clone’ and/or amplify via Polymerase Chain Reaction) mRNA from tissue samples to cDNA in solutions where fluorescent dyes Cye3 (green) and Cye5 (red) will be chemically taken up in the phosphate/sugar backbone of newly formed cDNA.

(D) Apply cloned/amplified cDNA solution(s) labelled with fluorescent dye to (subset of) the $N \times p$ array of spots on the slide. For example each of the p columns might represent a distinct tissue-sample (eg from distinct individual or cell-line). In two-dye comparative setting, **both** cDNA preparations are applied to the same (columns of) spots.

Microarray Experimental Steps, cont'd

(E) 'Wash' the slide so that only hybridized (tightly chemically bound) cDNA-to-fixed DNA will remain attached on the slide.

(F) Image-processing consists of laser-light shone onto the slide which causes fluorescently labelled cDNA components to register green or red (according to Cye 3 or 5) dye. Greater length of bound segments 'should' correspond to greater intensities.

(G) Preprocessing involves 'subtracting out' background intensities, and in some cases, forming local contrasts to remove local spatial effects on registered intensities.

(H) Resulting (preprocessed) data consists of $N \times p$ matrix of intensities registered in 1 or two channels. Such data is sometimes replicated across additional slides and/or tissue-sources (cell-lines).

Sources of Variation in Intensities

(1). unequal purity of RNA in prepared tissue samples.

(2). very different mix of RNA's in distinct tissue samples unless very carefully controlled.

(3). size of spot; amount of attached material.

(4). variation of amount of material placed on spots by 'print-tip'.

(5). random process of uptake of fluorescent dye(s), including chemical variations in (and across) the dyes themselves.

(6). gene-specific variations in (Red vs Green) dye effects do occur.

(7). contamination of spot intensities by neighboring spots bleeding across spot-boundaries.

(8). lengths of attached/anchored (probe) cDNA or oligonucleotide pieces is not unform from spot to spot (even among multiple spots per gene).

(9). extent of repeat-sequences and consequent possibilities for short hybrids or hybridization across different 'genes' varies strongly from spot to spot, by probe genes.

(10). lengths of target segments matching probe-segment pieces error-free is largely uncontrolled.

(11). spatial variations on the array may be due to local inhomogeneities in target solutions and/or ‘print-tip effects’ (systematic early vs late printing differences).

(12). systematic variations in intensity ratio (Red/Green) as a function of average Red + Green (log)-intensities.

TERMINOLOGY

Probe: immobilized cDNA on array

Target: labelled DNA in solution.

This terminology is uniform for cDNA arrays, but often **backwards** *for ‘oligonucleotide’ arrays (on membrane backing).*

Gene Expression Analysis – MicroArray Data

Multiple objectives:

- screening for 'active' (over- or under-expressed) genes
- grouping of genes by similar effect across tissue samples
- grouping/classification of tissue samples/cell lines/patients by similar profiles of (subsets of) gene intensities.

Excitement over the possibilities of microarrays come from all three objectives:

(a). multiple genes suggested friom other organisms (mice, fruit-flies, bacteria) can be screened simultaneously for relevance to human health;

(b). not-yet-understood genes can be partially classified by similarities with groupings of known related function.

(c). new patients can by grouped with others of better-understood prognosis.

Statistical Methods & Talk-topics

Key problem separate arrays, with separate tissue-samples, are independent statistically, but even ignoring dependencies across tissue-samples within arrays we have data structure of only p independent N-vectors, where p counts samples and N genes.

For each of the following topics, a talk could consist of a brief tutorial of a statistical topic and explanation of microarray application via a journal article. I can suggest one or more journal articles for each.

(I). Preprocessing, 'normalization'. Statistical techniques involve

(a) estimation of effects in simple linear (ANOVA) models for data cross-classified by gene, array, print-tip group, and replicate;

(b) nonparametric-smoothing or 'locally linear model' fitting to find (and subtract away) systematic curves connecting intensity ratios with average over dye of log intensities.

(II). Thresholding to find 'significant' single genes. Statistical technique: multiple comparisons or bootstrapping.

(III). Simultaneous Modelling of Sample-by-Gene intensities.

(a) Mixture models.
(b) Superposition ('plaid') models.

(IV). Unsupervised Learning/Clustering approaches. **Particular interest centers on hierarchical clustering of genes and of samples, preferably simultaneously.**

(V). Supervised Learning approaches: Discriminant analysis, classification methods, plus more sophisticated algorithmic approaches such as neural networks and 'support vector machines'.

(VI). Recall other (non-microarray) talk-topics mentioned, including: sequence algorithms and DNA sequence alignment and matching; Hidden-Markov Models for classification of genes and other DNA features; phylogeny and evolutionary distance via molecular 'similarity measures'.

General Resource

Encyclopedia of Statistical Sciences, ed. S. Kotz. Multivolume reference work with 1–6 page articles and many further background references.