Molecular Genetics: Challenges for Statistical Practice

J.K. Lindsey

1. What is a Microarray?

2. Design Questions

3. Modelling Questions

4. Longitudinal Data

5. Conclusions

1. What is a microarray?

A microarray looks similar to a computer chip:

a glass slide systematically covered with tens or hundreds of thousands of pieces of DNA, called probes, arranged in known positions.

Two main kinds of probes: (spotted) cDNA and (high density) oligonucleotides.

DNA microarrays allow biologists to conduct large-scale experiments yielding massive quantities of data.

One microarray experiment produces hundreds of thousands of observations.

Experimental size is growing faster than computing power.

Roughly, the experimental goals may be classified as

- identifying new genes,
- assigning function to genes,
- determining when genes are expressed,
- studying interactions among genes,
- detecting gene mutations,
- elucidating mechanisms of disease.

among others.

Most current microarrays are used to study gene expression.

This means the translation of DNA code into a protein (or RNA).

mRNA from an individual to be studied is prepared and then labelled with a fluorescent dye and applied to the microarray.

This preparation combines (hybridises) with certain probes on the microarray and the rest is then washed away.

The microarray is scanned to make it fluorescent and the intensity of each spot (probe) is recorded.

The results are then preprocessed, often with black-box commercial software:

filtering, transformation, normalisation.

This is a complex, multi-step procedure with variability at every stage:

background fluorescent, dye effects, spatial artifacts, etc.

yielding multi-level (probe, array, batch...) measurement errors,

as well as biological variability.

In more complex cases, two or more samples may be applied to the same array.

For example, in a two-channel microarray, matched samples labelled with two different dyes are competitively hybridised on the same array.

Ultimate goal: detecting 'real' variations in intensities indicating differential gene expression.

A simple, often used, criterion is "fold change": an arbitrarily chosen difference in level of expression.

For example twofold change between experimental and control cases.

However, this does not allow for sampling variation.

2. Design questions

High quality microarrays are expensive, often more expensive than the cost of obtaining the material to be studied.

Thus, there will be few individual sampling units with enormous quantities of data on each one.

Power and Sample Size

Thus, classical power and sample size calculations are difficult because there are many outcomes of interest per slide and they are not independent.

New methods, specifically for microarrays, are being developed.

Replications

Technical replication: same sample on different microarrays.

Results from the same RNA sample may differ

- among identical probes on the same microarray,
- among (supposedly) identical microarrays,
- among labs using the same procedures,
- among different types of arrays.

all sources of measurement, and perhaps systematic, error.

Whenever possible, arrays should be from a single batch, processed by the same technician on the same day.

Differences among labs is of special concern for reproducibility of results, the foundation of science!

Some work has been done on standard benchmark slides for interlab comparisons.

Biological replication: samples from different individuals.

Variation in technical replication complicates comparisons among different individuals.

Treatment Differences

By the use of more than one dye, certain microarray procedures allow comparisons of two (or more) RNA samples within the same slide:

Then, for example, compare

- matched pairs,
- each pair of two samples on two chips,
- each sample to a reference sample,
- in a chain or loop, with each sample on two different microarrays.

As far as possible, comparisons of most interest should be on the same slide.

Probe Layout

Design of the layout of the probes on the microarray involves:

- position of control probes,
- physical separation among replicate probes,
- accounting for systematic patterns due to the manufacturing process (such as variation in printing pins).

3. Modelling Questions

Gene expression levels are usually not normally distributed, with different probes often having different distributions.

However, there is so much noise in microarrays that, if the design and many nuisance covariates are not accounted for, the normal distribution may be reasonable.

A standard method is to fit the same statistical model independently to all probes on a microarray.

This involves making thousands of similar inferences and ignores similarities among probes.

Especially for simple models, many probes may yield virtually identical results.

Consider construction of gene profiles to classify patients into treatment groups.

Suppose 150 individuals are used to construct the profiles based on 20,000 genes.

A logistic regression is to be developed to predict whether or not to treat new patients, depending on the pattern of gene levels.

At most, there only 150 different gene patterns: on average, sets of about 133 genes have the same pattern.

How, then, can relevant models be constructed?

The number of different types of slides is rapidly growing:

promoter chips, protein chips (using antibodies), chip-on-chip, exon chips, etc.

Each requires the development of specific analysis techniques.

4. Longitudinal Data

Samples may be taken at several points in time from the same organism or cell culture.

Thousands of genes may be monitored over time.

The phenomenon under study may be a trend or periodic.

Trends: development of an organism, effect of some treatment such as a drug...

Periodic: cell cycles, circadian rhythms...

How can reasonable models be fitted when there are

- few time points: 5 - 20,
- many probes: 10-20,000,
- perhaps also under several biological conditions,
- with no or very few replications??

Biologists are looking for genes varying over time and, possibly, how they vary.

5. Conclusions

Microarray experiments provide statistical challenges in a wide variety of areas:

- design
- exploratory analysis
- testing
- modelling

How would that eminent geneticist, R.A. Fisher, have approached all of these problems?