Vital statistics

DNA microarrays have given geneticists and molecular biologists access to more data than ever before. But do these researchers have the statistical know-how to cope? Claire Tilstone investigates.

In June, Nick Fisher issued a dire-sounding warning. He is president of the Statistical Society of Australia, and fears that sloppy statistics could undermine the revolution promised by genomics and biotechnology. "If the collection, analysis and interpretation of the data are flawed then it may not only be a waste of a valuable resource — we could draw faulty conclusions and potentially risk our health and environment," Fisher claimed in a release to the media.

Fisher has a vested interest: he wants research organizations to employ professionally accredited statisticians — his society's members — to oversee the collection and analysis of genomic data. But he has a point. Technologies such as DNA microarrays have sent avalanches of data tumbling into labs where, previously, analysing the results of an experiment meant little more than glancing at an electrophoretic gel.

Over the past few years, microarrays, also known as DNA chips, have transformed molecular genetics, allowing researchers to study the activity of thousands of genes at a time. For example, you can compare healthy tissues with those that are cancerous, to identify genes that become more, or less, active in a developing tumour. But how do you separate significant differences in gene expression from background fluctuation?

This is where good experimental design and statistical analysis should come in. But there are no simple answers: interpreting microarray experiments is taxing the skills of even the most adept number-crunchers. "It's a technical and esoteric topic," says Paul Meltzer, a cancer geneticist and microarray specialist at the National Human Genome Research Institute in Bethesda, Maryland. If trying to make sense of microarray data has left you with spots circling before your eyes, you're in good company.

The problem is perverse: a typical microarray experiment provides both too much information, and too little. In most research projects, the idea is to study a small number of variables and repeat the measurements over and over again. Provided that you perform enough replicates, standard statistical tests can establish whether experimental results have real significance, or are more likely to be a consequence of random noise. Microarrays turn this approach on its head: there can be thousands of variables, corresponding to the number of individual genes being studied; but the high cost of the chips means that the number of repeated observations is usually very low.

In their initial forays into microarray research, many biologists didn't even try to use statistical methods. Some of the earliest papers simply recorded whether genes were active or not. Even when researchers began to use microarrays to measure levels of gene expression, they tended not to quote standard error values or confidence limits, and differences were judged to be meaningful if they exceeded some arbitrary level. Today, many manuscripts submitted to journals still focus their conclusions on genes that show a change in activity of, say, more than twofold.

But how do you know whether or not an apparent twofold change in gene expression is biologically meaningful? That's a difficult question to answer, because of the many sources of noise that can cloud the results.

The concept of microarray analysis is simple enough: messenger RNAs are extracted from a biological sample, converted into DNA, labelled with fluorescent dyes, and then washed over a glass slide bearing a grid spotted with DNA sequences from known genes. The labelled sequences bind to spots representing the genes from which the messenger RNAs were transcribed. So by analysing the location and intensity of the fluorescent signals, you can determine the level of activity for each gene. In some cases, it is possible to analyse two samples on the same chip by using different coloured dyes.

But noise creeps into microarray experiments at every stage, from the preparation of tissue samples to the extraction of data. Using different dyes can influence the results recorded by the lasers that measure the fluo-

AL GRANT

rescent signals, as can the location of the spots on the chip, or any unevenness or dust on the glass slide. Even using samples from the same piece of tissue, it is possible to get different profiles of gene expression using different microarray technologies[1].

Few researchers are in a position to repeat their experiments using various microarray systems. But it should, in theory, be possible to perform two other types of replication. First, each sample can be subdivided and the experiment repeated on several chips to assess fluctuation from array to array. You can also perform measurements on several different samples within each experimental condition. This replication is particularly important, as it is the only way to address fluctuations in gene expression between biological samples that have nothing to do with the issue under investigation.

The main deterrent is cost: commercial DNA chips retail for about US$1,000 each. And in some cases, for instance when studying rare diseases, it can be difficult to obtain enough samples to perform the desired replicates. One way round the cost issue is to collect several samples, pool them, and apply them to just one microarray. But the jury on pooling is still out. Many experts believe it is a bad idea because valuable information on sample-to-sample variation is obscured; others endorse the practice as a reasonable compromise that helps to smooth out background fluctuation.

R. PEDRONCELLI/AP

A sight for sore eyes? Statistical analysis of the data provided by DNA chips can be a major headache.

Even if you decide to perform proper, non-pooled replicates, it is difficult to know how many to do. One paper investigating the topic, published in 2000, recommended at least three replicates[2]. But among statisticians who have considered the issue, there is no clear consensus. "The number of replicates really depends on the kind of accuracy one wants to achieve," says Ernst Wit, a statistical geneticist at the University of Glasgow, UK. "It is impossible to come up with a single recommendation."

Biologists are now trying to use replication and statistical analysis to separate meaningful changes in gene expression from background noise, often using software packages produced by academic researchers and available for free download, or marketed by chip manufacturers. But in many cases, say experts, the statistics aren't being used correctly. "The majority of microarray papers are analysed with substandard methods," claims David Allison, a biostatistician at the University of Alabama at Birmingham.

Noise abatement

One source of confusion is how to correct for the 'false positive' results that are a consequence of the multiple comparisons inherent to microarray analysis. The results for an individual gene may suggest that there is only a 5% chance that recorded differences in its activity are the result of chance fluctuations; but if you repeat the same test across 10,000 genes, you are likely to get 500 'significant' results even if there is no real difference in gene expression. For simpler data sets, established methods exist for dealing with multiple comparisons. But opinions differ on how these should be applied to microarray data, says John Quackenbush, a specialist in genomic data analysis at The Institute for Genomic Research in Rockville, Maryland.

For many biologists running microarray experiments, comparing the activity of individual genes across different experimental conditions is only the start. The true power of the technology, Quackenbush argues, is its ability to reveal common patterns of gene expression across different samples. Very often, genes with similar profiles of activity will have related functions or be regulated by common mechanisms.

Grouping genes in this way involves statistical techniques collectively known as cluster analysis. The first step is to draw up a gene-expression matrix in which the rows represent individual genes, the columns are individual samples, and each cell contains a measure of

the gene's activity. From this matrix, each gene can be given a coordinate called its 'expression vector', which represents a point in an *n*-dimensional mathematical space, where *n* is the number of samples. The precise position of each gene in each dimension depends on its level of activity in the sample concerned.

The genes are then clustered into groups by methods that 'measure' the distances between their respective expression vectors. The difficulty is that there are various ways of doing this, which delight in exotic names such as '*k*-means clustering' or 'self-organizing maps'[3]. Statistical experts struggle to agree over which clustering method should be applied under what circumstances. And for most biologists — even those whose eyes haven't glazed over at the initial mention of an *n*-dimensional space — the detailed workings of these methods remain a mystery.

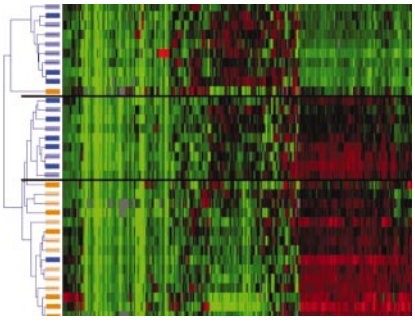
Nevertheless, cluster analysis has caught on, and the most common approach is a technique called hierarchical clustering, first applied to microarray data in a paper published in 1998 by David Botstein and other microarray pioneers at Stanford University in California[4]. Here, an iterative algorithm draws up a tree in which the lengths of the branches correspond to the degree of similarity between genes.

General clusters

Hierarchical clustering has helped flustered researchers to make sense of what would otherwise be unmanageable spreadsheets of data. But statistical purists point to problems with the way in which it is being applied. Some argue that the technique is best suited to determining relationships between a small number of variables, rather than deriving patterns involving thousands of genes across a huge data set. "Hierarchical trees are famously unreliable for good high-level clusters," explains Wit.

This means that biologists can be sent down blind alleys, if they see clusters that reinforce their own assumptions about the relationships between individual genes. "People just tend to pick their favourite cluster for further study," says Allison, who likes to play a trick on biologists: he presents them with two trees drawn up from simulated microarray data, one representing genuine clusters of genes with similar expression profiles, the other in which the genes have been clustered randomly. When he shows this slide, Allison is greeted by gasps and chuckles, as those in the audience realize that they don't have an intuitive ability to recognize a 'true' cluster.

Cluster analysis can also group samples that show similar patterns of genome-wide gene expression. The power of this approach was demonstrated in 2000 by a team led by Louis Staudt of the National Cancer Institute in Bethesda, Maryland, which used hierarchical clustering to group B-cell lymphomas into two distinct classes. These correlated with differences in patient survival, and seemed from their gene-expression profiles to be related to the stage of development of the cell from which the cancer originated[5]. But last year another group, using a different clustering method, failed to find this association[6].



Branching out: cluster analysis can group samples that show similar patterns of gene expression.

Experts argue that such disagreements are only to be expected, given that microarray analysis is a young and fast-moving field. They point to efforts to refine the statistical methods applied to microarray data, such as the annual Critical Assessment of Microarray Data Analysis (CAMDA) meeting, now in its fourth year. Organized by Simon Lin and Kimberly Johnson of Duke University in Durham, North Carolina, CAMDA culminates in the award of a prize to the group judged to have conducted the best analysis of real data sets posted before the meeting on CAMDA's website.

Quackenbush, who is giving a seminar at the next CAMDA gathering in November, suspects that there may never be agreement on the 'right' statistical techniques to deploy on microarray data. The future, he suggests, lies in incorporating additional biological information into the analysis. He points to a paper published in March, in which microarray data and information on the chromosomal location of genes that influence obesity were considered together to identify two subtypes of obesity in mice[7].

Array of hope

While microarray experts push back the frontiers, efforts are being made to improve standards of experimental design, data presentation and analysis among rank-and-file users of the technology. Particularly useful is a set of guidelines called minimum information about a microarray experiment (MIAME), laid down by the international Microarray Gene Expression Data Society. These include specific statements about experimental design, including the number of replicates done, allowing researchers to interpret one another's data more easily.

Many journals are now toughening up their criteria for accepting papers describing microarray experiments. Since December 2002, for instance, *Nature* and its sister research journals have required authors of such papers to complete a MIAME checklist; data must also be submitted before publication to one of the two main public repositories for microarray data[8]. One of the most explicit statements has come from the journal *Arthritis & Rheumatism*, which in April last year published guidelines stressing the importance of appropriate statistical analysis[9]. "It is not sufficient to say that the expression of a particular gene is twofold greater in a sample than in a control," the journal's editors stated.

Chip manufacturers and specialists in academia are also trying to help microarray users become more rigorous. Last year, for instance, the chip-making firm Affymetrix of Santa Clara, California, began a series of worldwide workshops on experimental design and the use of statistical software. Meanwhile, the Bioconductor project, run by the Dana Farber Cancer Institute in Boston, is backing its programs for genomic data analysis with short courses in their use. By the end of 2003, these will have taken place in the United States, Europe and Taiwan.

Standards should improve further as statistics and bioinformatics become more prominent components in the training of molecular biologists, and as a growing number of statistical experts gets to grips with the complexities of microarrays. "There are lots of good statisticians out there, but not as many of them have been exposed to microarray data as are needed," says Meltzer.

In the meantime, the message to biologists is clear: if you want to work with microarrays, you need to find yourself one of these precious experts — and don't wait until after you've collected your data. The following advice, from pioneering British geneticist and statistician Ronald Fisher, rings even more true today than when he uttered it, back in 1938: "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." ■

Claire Tilstone is a postgraduate at the University of Bath, UK, studying science communication; further reporting by Peter Aldhous, *Nature*'s chief news and features editor.

1. Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L. & Kohane, I. S. *Bioinformatics* **18,** 405–412 (2002).
2. Lee, M.-L. T., Kou, F. C., Whitmore, G. A. & Sklar, J. *Proc. Natl Acad. Sci. USA* **97,** 9834–9839 (2000).
3. Quackenbush, J. *Nature Rev. Genet.* **2,** 418–427 (2001).
4. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
5. Alizadeh, A. A. *et al. Nature* **403,** 503–511 (2000).
6. Shipp, M. A. *et al. Nature Med.* **8,** 68–74 (2002).
7. Schadt, E. E. *et al. Nature* **422,** 297–302 (2003).
8. *Nature* **419,** 323 (2002).
9. Firestein, G. S. & Pisetsky, D. S. *Arthritis Rheum.* **46,** 859–861 (2002).

Critical Assessment of Microarray Data Analysis
www.camda.duke.edu
Microarray Gene Expression Data Society
www.mged.org
Affymetrix
www.affymetrix.com
Bioconductor
www.bioconductor.org

REF: 7