'It is essential that the data extracted from microarray experiments is subjected to the same high standards of rigour as data collected from traditional platforms'

editorial

David Chrimes
Business Development manager, BlueGnome

How can data quality and automation enhance confidence in microarray data?

Microarray technology is being used for many research applications including drug discovery in the pharmaceutical industry [1]. Information used to aid the validation of potential drugs needs to be highly accurate in order for companies to proceed to trials and ultimately licensing.

Accuracy and repeatability of this technology becomes evermore essential as we move into a world where arrays are being studied as a potential tool for clinical diagnostics. Therefore, it is essential that the data extracted from microarray experiments is subjected to the same high standards of rigour as data collected from traditional platforms. This rigour can be achieved through the automation of microarray experiments, in this article I will be focusing on the image analysis.

Reproducibility will provide the confidence to enable the full power of these experiments to be exploited. It is worth noting that whilst the specific aspects covered in this article relate to spotted arrays, the analysis is applicable to many other high throughput platforms including NMR and MS.

Automation removes the need for human intervention and hence human subjectivity

The process of identifying which spots to analyse or exclude, and also how to analyse these spots, has traditionally been a manual procedure. It has relied on the person(s) doing the analysis setting 'user thresholds' that they feel provide optimal results for the experiment.

Variability between users makes comparing the results within laboratories and between laboratories problematic. An automated system that can assess the data on a spot-by-spot basis is essential to remove this variability and provide cross-user comparisons.

A further problem with analysing microarray slides is that they are becoming more densely packed with gene-probes. Spotted arrays routinely use 30,000 spots, with some groups looking at 50,000 spots per slide. For an individual, manually checking each slide, looking at in excess of 30,000 spots, becomes a highly repetitive task. In such cases it is inevitable that human error impacts the final results.

An automated system will always follow a predefined set of rules and can provide a log of events, detailing each of the processes carried out and also if one of them fails. So if at any stage there is a failure, it can be traced back. By contrast, if an individual makes an error or does something unusual, it is often very hard to trace and place a 'value' on how this has altered the data.

A pioneer of the automated approach is Nigel Saunders, The Sir William Dunn School of Pathology, University of Oxford.

1359-6446/05/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)034641

He says: 'For data quality reasons, and to maximize the comparability of results, the operator dependent processes of image acquisition and data extraction have to be as consistent and reproducible as possible.

Optimal consistency is not achievable without manual manipulations at these stages, particularly when more than one individual is involved with the process. Automation offers a means to overcome this process bottleneck and remove an important source of operator-generated variability from experiments.'

An automated system allows the user quality assurance of the whole process

It may seem obvious, but the final quality of the data is dependent on the quality of the original slide and this can be a time consuming task to establish.

Anthony Brown, Dept of Pathology, University of Cambridge, uses an automated approach to speed up and improve slide production and quality assessment.

He said: 'This enables rapid analysis of a large matrix of spotting possibilities offered by the broad range of probe types: long/short oligos; cDNAs; BAC products etc; slide chemistries ; slide manufacturers ; spotting buffers (aqueous buffered, organic solvents, surfactant additives etc) and spotting technologies (such as split pins, internal capillary pins, non-contact). Spot parameters of interest are deposition accuracy, spot diameter, circularity, uniformity, reproducibility and signal. This is important for tracing problems in spotting processes because uniform, reproducible spotting performance enables uniform, reproducible array performance.

The result is that the end user has confidence that the new slides they are using will provide comparable results to the last batch, which is vital when comparing cross-experiment results.'

As well as providing quality assurance of the slide production process, a robust automated system provides the ability to place a confidence rating for each spot. This approach allows the quality of each spot on the array to be compared and also provides a picture of the overall quality of the array.

Automation saves time

As well as an increase in slide complexity, there has also been an increase in the number of slides that are used per experiment.

The classical 'treatment and control' slide experiments are being replaced by time-course experiments, all of which require biological replicates per time point. Also, there are an expanding number of groups undertaking large array comparative genomic hybridization (CGH) classifications that require 200+ samples.

This has led to a massive production of data with a bottleneck in data processing. For this reason, the speed of data acquisition needs to be increased and again, automation is the key. In the early days of microarray experiments, people were running only a few slides per experiment and people could afford to spend the time manually checking each slide. Now that tens to hundreds of slides are run per experiment, a great number of person-hours are required to analyse the data, and this has a great financial implication.

This has been borne out by a group at the Institute of Food Research headed by Jay Hinton: 'Automation enables our computers to analyse three times more microarrays per day. Last year, we used some 3000 microarrays in the Institute of Food Research. We estimate that this would have saved us six months of a scientist's time, two months in my laboratory alone'.

Automation and data quality

It is essential that any automated approach does not swap repeatability for accuracy. The biggest problem with trying to automate any procedure is the ability to cope with variables. It is known, for example, that the grid is rarely, if ever, perfect and therefore a robust grid and spot finding algorithm is required. Also, there is the question of automating the extraction of signal from noise; non-uniform background across a slide due to dust-spots and smearing, caused by hybridization or washing problems, significantly complicates signal extraction.

Traditional approaches to the analysis of images have required several major manual interventions. The first of these has been to ensure that all the spots have been found. The second intervention was to assess whether the shapes around the spots accurately contain the spots and provide a strong separation of signal against background.

This is compounded as traditional programs have tended to use strong geometric shapes with hard edges with everything inside the edge called as 'signal' and everything outside classified as 'background'. This kind of approach does not deal well with irregular shaped spots, for example comets, where incorrect definition of the edges of signal and noise actually introduce unnecessary noise to the experiments. Often it is up to the user to define thresholds to determine signal from noise and hence to distinguish between which results are to be kept and which discarded. If these thresholds are tightened, you tend to improve the data quality but decrease the quantity and *vice-versa*. This means that there is a great deal of operator skill required to optimize the results for each and every slide

The key to automation is the fact that as these processes are well understood, they can be modelled. One approach that has been successfully applied has been to use the Bayesian statistical theory [2,3]. The Bayesian approach provides a rigorous mathematical framework using prior knowledge that is available about the system to inform the way new data should be analysed. In practice this prior knowledge varies from that in which there is a high degree of confidence, such as the array layout, to the less predictable, such as the washing effects.

Using the Bayesian approach, it is possible to automatically combine all grades of prior knowledge with the data to deliver improved results. This is radically different from traditional methods that use simplistic thresholds to divide the data into prefect or to be discarded. The other main advantages of using a Bayesian approach to spot finding and quantification are that there are no hard assumptions made of circularity or hard edges as the image is assessed on a statistical pixel by pixel basis.

Conclusions

The solutions to high-quality data collection and extraction from high-throughput experiments often lag behind the scientific knowledge of how to generate the data, and in many ways microarrays proved to be no exception.

There is a lot of ongoing work that is closing this gap because, although the initial investment might appear high, the returns are significant.

Greatly improved spot-finding and image analysis has brought about intra and inter-laboratory reproducibility with a resultant increase in the confidence people are prepared to place on microarray data. The most important aspects of the ability to extract highly reproducible data is that it has led to a greater understanding of the underlying biology than had previously been possible.

References

1 Howbrook, D.N. *et al*. (2003) Developments in microarrays. *Drug Discov. Today* 8, 642–651
2 Zou M, Conzen SD *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data*. *Bioinformatics* 21, 71–79
3 Bhattacharjee M. *et al*. (2004) Bayesian integrated functional analysis of microarray data. *Bioinformatics* 20, 2943–2953

David Chrimes
Blue Gnome,
Breaks House,
Mill Court,
Great Shelford,
Cambridge,
UK, CB2 5LD
email: david.chrimes@cambridgebluegnome.com