Microarrays: handling the deluge of data and extracting reliable information

Kenneth R. Hess, Wei Zhang, Keith A. Baggerly, David N. Stivers and Kevin R. Coombes

Application of powerful, high-throughput genomics technologies is becoming more common and these technologies are evolving at a rapid pace. Genomics facilities are being established in major research institutions to produce inexpensive, customized cDNA microarrays that are accessible to researchers in a broad range of fields. These high-throughput platforms have generated a massive onslaught of data, which threatens to overwhelm researchers. Although microarrays show great promise, the technology has not matured to the point of consistently generating robust and reliable data when used in the average laboratory. This article addresses several aspects related to the handling of the deluge of microarray data and extracting reliable information from these data. We review the essential elements of data acquisition, data processing and data analysis, and briefly discuss issues related to the quality, validation and storage of data. Our goal is to point out some of the problems that must be overcome before this promising technology can achieve its full potential.

A series of recent articles has demonstrated the use of microarray technology in gene identification and disease classification. Following the success of such proof-of-principle studies and the announcement of a draft of the human genome, application of genomics technologies is becoming more frequent. The innovative integration of automation, molecular biology and informatics has revolutionized the analysis of gene expression. These high-throughput platforms have generated a massive onslaught of data that threatens to overwhelm researchers.

Data acquisition

Two major array platforms are chemically coated glass slides and nylon membranes (filters). Both involve printing thousands of cDNA fragments (probes) onto a substrate. To profile gene expression in a given cell population, mRNAs are isolated and reverse-transcribed to cDNA, at which step fluorescent dyes (for glass arrays) or radioactive isotopes (for membrane arrays) can be incorporated. The labeled cDNAs (targets) are then hybridized to the arrays. Hybridization signals are detected using a laser scanner (for fluorescence) or a phosphorimager (for radioactive materials), resulting in the production of digital images.

The images from the hybridized arrays constitute the essential raw data for microarrays. The intensities of the signals are measured using specialized imaging software. The micro-spotting robot that is used to print the arrays lays down the probe spots in a regular pattern. The pattern is usually a rectangular grid and often has a hierarchical arrangement of subgrids. Unfortunately, printing problems, membrane distortion or image registration problems require that the software accurately locates each individual spot before quantification; commercial quantification software is available for this. However, it is often necessary to manually adjust subgrid and/or spot alignment – this is particularly true for poor quality images. Because problems at this stage will introduce unrecoverable and often undetectable artifacts in the data, it is important to visually inspect the computer-generated alignment and establish a reproducible protocol that optimizes the software parameter settings for a given platform and array layout.

Most quantification programs offer several options for measuring spot signals and related data on the image. Most allow selection of total spot intensity (sometimes labeled ‘volume’), background level, detected or measured area of spot and variance of pixel intensity within the detected spot. Because the signal intensity is not uniform over the spot (Fig. 3), careful consideration must be given to the best way to capture both the information relating to the amount of labeled cDNA on a given spot and also some measure of the quality of this information. Background measurement techniques also vary; typically, both local and global background measurements are available. The choice of how to estimate background levels is more crucial for membranes because of the effect of ‘bleeding’; which is when intense spots spread into adjacent spots. This can also be an issue with slides that have a very tight pitch (inter-spot distance) relative to spot size. For standard glass slides, background levels can vary considerably over the area of the array, and so local background measurement is preferred. When taking both spot and background measurements, contamination (e.g. with dust) is a real concern (Fig. 4), and so robust metrics are preferred. We recommend choosing the median of the background pixels for analysis.

Background correction is crucial to accurately estimate true expression levels. Usually we assume that the observed signal is a combination of the true
signal from the specific hybridization of interest, and background signal due to nonspecific hybridization and contamination. In fact, analysis of background levels can also be diagnostic for artifact contamination problems. Several situations lead to increased background. These include dust, fibers, auto-fluorescence of the coated glass, fingerprints and hybridization misadventures such as residue from inadequate washing and dehydration near the edge of coverslips. The standard approach is to simply subtract the background estimate directly from the spot intensity. Fluorescence is affected by the presence of bound DNA on the glass and the background, measurement in regions between spots might yield poor estimates of spot background.

**Data quality**

The large number of measurements that result from microarray experiments makes data quality assurance a crucial issue. The individual bioassays can no longer be validated or conditions be optimized for each particular gene of interest. With respect to microarrays, all genes are hybridized under the same conditions (which might be optimal for very few, if any). Quality assurance is made difficult by both the large number of potential sources of variation and the large number of measurements. It is important that each individual laboratory runs control experiments to identify and quantify the relevant sources of variation. In addition to the contaminations mentioned previously, other potential sources include: steps in the sample collection and preparation (RNA extractions, labeling and amplification), elements of the hybridization process (e.g. container, solution, temperature, time and the extent of cross-hybridization), elements of the array printing process (e.g. purity of probe solutions, print head status, robot movement, quality of substrate and uniformity of substrate coating), elements of scanning (e.g. type of optics, laser cross-talk, dye quenching, laser power and length of exposure for radioactive labeling), elements of spot identification and quantification as described later. Dilution experiments are also necessary to assess the relative accuracy of the gene-expression quantification. Individual experiments should be conducted to optimize the experimental conditions. The combination of conditions that yield optimal results is dependent on many factors, and many different combinations might yield comparable and reliable results. However, a small modification in only one parameter might drive the system well out of the optimal range and therefore produce unreliable results. For example, insufficient amounts of blocking agents in the hybridization solution results in a decreased signal-to-noise ratio. It also appears that hybridization temperature might affect the results. Of course, the importance of these hybridization conditions might depend on the particular specifications of other parts of the whole microarray analysis.

**Data analysis**

The appropriate analysis depends on the experimental design and the study objectives. A well-planned design is more likely to lead to interpretable results and appropriate designs are dictated by particular study objectives. For microarrays, these include gene discovery, pathway elucidation and molecular classification (i.e. class prediction and class discovery). Simple experiments can be conducted on single slides. Two samples are hybridized such that genes that are differentially expressed between the samples are interpretable from the experimental design. One example would be a tissue sample from a particular tumor, and adjacent, normal tissue of the same cell type. Analysis of data from a single slide is typically conducted in terms of the ratios of the measurements from the two samples. The statistical behavior of these ratios, and statistically reliable methods of assessing significant differential expression is the subject of current research. It is worth noting that low relative variation in the gene expression ratios that correspond to individual genes is far more important for gene discovery than for class discovery or class prediction, in which the entire genomic signature is used.

Our current approach to the statistical analysis of microarray data from slides involves working with the variation between replicate spots on the array. The use of replicate spots is crucial but not often reported in
the literature. Thus we describe our method in some detail. Genes are spotted on arrays in replicate pairs so that there are two measurements for each gene. The data are the base 2 logarithmic transformations of the background-subtracted spot pixel volumes. Rather than plot the values for the two replicates against each other, we plot their difference against their average (Fig. 5). It is clear from this plot that the variability of the differences is higher for spots with a lower average intensity. To model this variability as a function of the average intensity, we use a locally smoothed (i.e. moving-average-type) estimate of the interquartile range (IQR). We superimpose curves on the plot that correspond to ±3 × IQR. These curves can be used as a rough guide for identifying data artifacts by noting replicate differences that are much larger than would be expected based on their average intensity. Similar analysis of replicate pairs is performed on the ratios of spot intensities. Here, the ratios are taken between the images corresponding to the two separately labeled samples that have been co-hybridized to the slide. The variability of the log ratios depends on the strength of the signal at the spot; weaker signals lead to greater variability. Thus, we model the variability of replicate log ratios as a function of the log signal intensity. We then use this estimate of expected variability in our analysis of the log ratios. The individual log ratios are scaled by our estimate of the variability. Genes with the largest scaled values are those most likely to be truly differentially expressed. Care must be taken with low-intensity signals, where one or both of the spot readings have been replaced with a threshold value (to avoid taking logarithms of negative values), because they can bias the results.

The results are normalized or adjusted to a common standard so that results from different experiments can be compared. Normalization is required because the expression levels are measured relative to the amount of labeled mRNA in the hybridization process. There are several approaches to normalization of expression levels. One method is to normalize against a putative housekeeping gene or a group of such genes. This approach assumes that the concentrations of these positive-control gene products is relatively constant from cell to cell but there are little data to support this assumption. We prefer to divide the observed (background subtracted) expression values by the median value for the spots measured (above background measurement). To normalize the ratios of the red-channel intensities to the green-channel intensities, we recommend dividing them by the median ratios for spots (above background measurement). This approach avoids the vagaries of putative housekeeping genes, but is only appropriate if less than half of the spots on the array will be affected by differences between arrayed samples. An alternative approach is to use a large number (≥75) of housekeeping genes.

For experiments involving multiple arrays, between-array normalization is required to avoid artifacts in the data caused by differences in experimental and processing conditions. For membranes, the expression levels are normalized, whereas for slides, the ratios are usually normalized. For multi-slide experiments, one approach is to have a common reference sample (often a cell line or mixture of cell lines) to act as a control for normalizing the ratios. The reference sample must be chosen carefully, because only genes expressed at least at a moderate level in the sample will yield reliable ratios for assessment.

Multi-array experiments, such as time-series and classification studies, often involve data filtering – spots exhibiting little variation across arrays are omitted from the analysis. These studies often have multifaceted goals that include pattern and feature recognition. The patterns are typically assessed using cluster analysis (i.e. unsupervised classification). Clustering is often accompanied by dimension reduction methods (such as multidimensional scaling [MDS] or principal components analysis [PCA]) that facilitate visualization, and provide new smaller sets of independent dimensions. These new sets of dimensions contain most of the information from the original data.
The most widely used clustering algorithms are hierarchical methods that produce nested sequences of clusters by repeatedly aggregating smaller clusters or disaggregating larger clusters. The results are plotted as a dendrogram. Partitioning algorithms (such as k-means) produce clustering by iteratively reassigning elements to clusters to optimize some pre-selected criteria (often to maximize the between-cluster variation and minimize the within-cluster variation). These methods produce a single clustering result, as opposed to a nested hierarchy of results, but the number of clusters must be pre-specified. Comparing results for a range of numbers of clusters is often informative. Most clustering algorithms are not driven by the original data directly, but by a matrix of distances (or similarities) between the objects to be clustered (computed from the original data). Common choices are the Euclidean distance and the linear correlation coefficient. Results can change with the choice of metric and the choice of clustering algorithm. It is advisable to try several approaches to assess the robustness of the observed clusters. Statistical resampling methods can also be helpful for assessing the robustness of the clusters. MDS analysis also uses the matrix of distances (similarities) but seeks to find an arrangement in low dimensional space (two or three dimensions) that preserves the relative distances between objects. New coordinates (dimensions) are formed for graphing the new arrangements. These facilitate visual pattern recognition and cluster identification. Often, clustering is observed using multiple dimensions of MDS that can be missed by the mainly unidimensional approach of hierarchical or partitioning cluster analysis. By comparing the distances between objects in the new arrangement to those from the original data, the extent of preservation of the original arrangement can be assessed. A variety of MDS algorithms are available, and results can vary with different approaches. PCA produces a similar result, but the new dimensions are computed as mutually independent linear combinations of the original variables with maximum variance. PCA can be made more robust by rescaling the original variables to have equal variance, or by replacing the measured values by their relative ranks. A self-organizing map is a form of neural-network-type clustering that has features of both MDS and partitioning clustering. Clusters are formed in such a way that distances between neighboring objects are preserved. Linear discriminant analysis, recursive partitioning analysis (i.e. classification trees) and artificial neural networks are methods of supervised learning used for class prediction. The first two methods can also be used for gene discovery, where they can find combinations of genes that together can predict class membership, but individually are not predictive, and therefore would be missed in the usual one-variable-at-a-time analysis.

It is important to distinguish clustering of the samples and clustering of the genes. The former will identify samples that have similar gene expression profiles and so might constitute some new, previously unidentified subgroups (such as class discovery for molecular classification in pathology). The latter will identify genes that behave similarly across various experimental conditions, and thus might be related in some meaningful way (such as belonging on the same biological pathway). It is important to remember that any clustering of samples will depend on both the genes included and the control samples used for normalization. Similarly, any clustering of genes will depend on the samples included in the analysis. Several recent publications have discussed clustering methods for microarray data22–28. One promising avenue of research involves weighting the expression data according to quality metrics, such that more reliable data are given relatively more ‘weight’ in the analysis (e.g. the replication-based method described above can be used to estimate the SD of the log ratios, and these estimates can be used to construct weights for use in computing a correlation coefficient by taking the inverse of the estimates of SDs and scaling these by the sum over all the samples [so the weights add up to one]).

Feature recognition (i.e. ‘cherry picking’) is a common goal in large-scale microarray experiments. Finding the few genes that are most responsible for the observed patterns in the data is a well-studied, but still unsolved, statistical problem. Recent advances in statistical pattern recognition and data mining are increasingly being applied to microarray data29–32. However, careful consideration of data quality and experimental design issues is required for
interpretable results. Too often, data mining activities are simply large-scale applications of poorly understood methods to poorly understood data.

Data storage
Because microarray data typically consist of thousands to tens of thousands of measurements for each array, effective storage is essential but challenging. Because the images contain the original data for the microarray, they are the most important aspect of the data and thus it is essential that they are stored and made readily available for comparison and visual verification with the derived results. (Some programs include simultaneous visualization of the original image, and a pseudoimage constructed from the idealized array layout and the quantified measures.) Along with the image, the experimental details and the quantification results need to be stored.

Microarray databases must also deal with the typical database issues that arise in any large-scale data enterprise; for example, data organization, user interface, security, maintenance, software and so on. The massive amount of data that arise from microarray experiments leads to problems similar to those seen for databases in large-scale commercial enterprises and scientific projects such as remote-sensing satellite imaging. Issues related to database standards, although essential for interaction among researchers, are not unique to microarrays and it is essential that microarray researchers take advantage of the advances in database technologies that are currently available.

Because of the large number of potential sources of variation, and the dependence of the accuracy and precision of the results of gene expression experiments on so many elements, at present it does not seem appropriate to postulate a public gene expression database. Rigorous standardization and validation of microarray data is required before such a data repository would be useful. Although such a database is appropriate for genomic sequence data, these data are not subject to the same level of perturbations as gene expression data. Recent progress has been made in formulating standards for microarray data representation and description. However, many obstacles must be overcome before microarray data are sufficiently robust to justify aggregating data across laboratories and platforms.

Data validation
In light of the high degree of variation within microarray data, it is essential that results are validated independently. We point out that not all the validation experiments need to be carried out using microarrays. Indeed, because of the variability associated with microarray technologies, an independent assay such as northern blotting or real-time PCR is often performed to confirm the interesting findings from microarray assays. Toward this end, high-throughput real-time PCR assays are being established to screen multiple samples for a single gene for diagnosis and prognosis purposes.

Conclusion
Considerable research during the past several years has demonstrated the power of genomic technologies, and although microarrays show great promise, they are susceptible to many problems, some of which have been identified and solved. Many unresolved issues remain to be tackled before the promise of this fascinating technology will be realized.

References
Manipulating redox systems: application to nanotechnology

Gianfranco Gilardi and Andrea Fantuzzi

Redox proteins and enzymes are attractive targets for nanobiotechnology. The theoretical framework of biological electron transfer is increasingly well-understood, and several properties make redox centres good systems for exploitation: many can be detected both electrochemically and optically; they can perform specific reactions; they are capable of self-assembly; and their dimensions are in the nanoscale. Great progress has been made with the two main approaches of protein engineering: rational design and combinatorial synthesis. Rational design has put our understanding of the structure–function relationship to the test, whereas combinatorial synthesis provided examples of novel approaches where redox proteins are ‘wired up’ in efficient electron-transfer chains, in artificial multidomain structures (molecular Lego), are ‘linked’ to surfaces in nanodevices for biosensing and nanobiotechnological applications.

As the understanding of the structure–function relationship of redox proteins and enzymes is increasingly understood and new means are developed to tailor their properties for nanodevices, a new era is approaching in nanobiotechnology for exploiting these systems. The state-of-the-art research in this field is at a stage where domains that traditionally belong to the physical sciences (atomic-scale microscopy), chemistry (electrochemistry and electron transfer) and biology (enzymology, protein design and molecular biology) are coming together to offer new synergetic opportunities for nanobiotechnology.

Redox proteins in nanobiotechnology

Redox proteins and enzymes carry out many key reactions of biological and technological importance (Fig. 1). The underlying process essential for these reactions is electron transfer. Protein-mediated electron transfer is a key phenomenon, not only in cellular processes (e.g. respiration and photosynthesis), but also in reactions of biotechnological interest (e.g. degradation of pollutants and biomass, and drug and food processing). Much progress has been made over the past ten years in understanding how the protein matrix finely tunes the parameters that are central to the regulation of biological electron transfer. R.A. Marcus’s theory of biological electron transfer (Box 1) gained him the 1992 Nobel Prize in Chemistry and fuelled many studies that attempted to unravel the details of key biological functions.

Undoubtedly, the protein matrix has a key role in regulating redox functions – indeed, few cofactors can perform the plethora of functions ascribed to redox proteins and enzymes. Even in simple electron-transfer proteins, such as b-type and c-type cytochromes that contain the same haem iron, a relatively simple parameter such as the redox potential varies over a range of 800 mV, from −400 mV for cytochrome c, to +400 mV for cytochrome b. This range highlights the power of the protein matrix in tuning function. More complex systems, such as the cytochrome P450 enzymes, gate electron transfer through a spin-state change associated with substrate binding. Such a mechanism allows variations of >100 mV to occur within the same protein and to allow the flow of electrons from the reductase only in the presence of the substrate.

Much research has focused on ‘analytical protein engineering’, which involves the mutation of existing redox proteins to test our understanding of protein structure and function. A more recent trend is the so-called ‘de novo protein engineering’, which aims to create novel redox proteins from first principles.