11. **Lopez-Cabrera, M., A.G. Santis, E. Fernandez-Ruiz, R. Blacher, F. Esch, P. Sanchez-Mateos, and F. Sanchez-Madrid.** 1993. Molecular cloning, expression, and chromosomal localization of the human earliest lymphocyte activation antigen AIM/CD69, a new member of the C-type animal lectin superfamily of signal-transmitting receptors. J. Exp. Med. *178*:537-547.
12. **Mages, H.W., O. Rilke, R. Bravo, G. Senger, and R.A. Kroczek.** 1994. NOT, a human immediate-early response gene closely related to the steroid/thyroid hormone receptor NAK1/TR3. Mol. Endocrinol. *8*:1583-1591.
13. **Marton, M.J., J.L. DeRisi, H.A. Bennett, V.R. Iyer, M.R. Meyer, C.J. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai, et al.** 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat. Med. *4*:1293-1301.
14. **McCullagh, P. and J. Nelder.** 1983. Generalized Linear Models. Chapman & Hall, London.
15. **Ruan, Y., J. Gilmore, and T. Conner.** 1998. Towards Arabidopsis genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. Plant J. *15*:821-833.
16. **Schena, M., D. Shalon, R.W. Davis, and P.O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray [see comments]. Science *270*:467-470.
17. **Schena, M., D. Shalon, R. Heller, A. Chai, P.O. Brown, and R.W. Davis.** 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA *93*:10614-10619.
18. **Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, et al.** 1996. A gene map of the human genome. Science *274*:540-546.
19. **Shalon, D., S.J. Smith, and P.O. Brown.** 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. *6*:639-645.
20. **Spellman, P.T., G. Sherlock , M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher.** 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell *9*:3273-3297.
21. **Wang, A., M. Doyle, and D. Mark.** 1998. Quantitation of mRNA by the polymerase chain reaction. Proc. Natl. Acad. Sci. USA *86*:9717-9721.
22. **Yssel, H., R. de Waal Malefyt, M.G. Roncarolo, J.S. Abrams, R. Lahesmaa, H. Spits, and J.E. de Vries.** 1992. IL-10 is produced by subsets of human CD4+ T cell clones and peripheral blood T cells. J. Immunol. *149*:2378-2384.

8 Information Processing Issues and Solutions Associated with Microarray Technology

Yi-Xiong Zhou, Peter Kalocsai, Jing-Ying Chen, and Soheil Shams
BioDiscovery, Inc., Los Angeles, CA, USA

INTRODUCTION

Microarray technology provides an unprecedented means for carrying out high-throughput gene expression analysis experiments. With today's technology, a single hybridization experiment using one chip can yield expression profiles for tens of thousands of genes simultaneously. A microarray project, as with most genetic research projects, usually involves acquisition and validation of large datasets. These datasets contain a variety of different information, ranging from sequence data on the genes or clones placed on each slide to quantified expression values for each gene under different experimental conditions. Furthermore, each project typically requires iterations of series of processes, which compounds the amount of data needed to be handled. Figure 1 depicts a typical processing scheme. The process starts from experiment design and array fabrication, through array scanning, image analysis, and finally gene expression data analysis. The expression data analysis will likely lead to a new hypothesis, which will in turn require another iteration of experiment design, array fabrication, and so on. During each iteration, a process generates new data, which in turn can be used by other processes within the overall system. For example, data associated with the array fabrication process, containing information that links gene sequences to spots, can be exploited by the image processing process to automate many aspects of the operation. Due to the large volume of data generated in microarray projects, information processing issues become prominent.

In the maturation process of microarray technology, there are two kinds of challenges. One is to develop the hardware for fabricating, hybridizing, and reading

Microarray Biochip Technology
Edited by Mark Schena

arrays. The other is to manage the massive amount of information associated with this technology, so that the results can yield understanding of genomic functions in biological systems. With the steady progress in hardware technology development, currently available equipment can reliably produce and image >10 000 spots on single microscope slides, and 100 000 spots will be possible in the near future (e.g., http://arrayit.com). In other words, the fundamental challenge of microarray hardware has been mostly resolved. On the other hand, the informatics challenge has just begun to be addressed. There are three major issues involved. The first is to keep track of the information generated at the stages of the chip fabrication and hybridization processes. The second is to process microarray images to obtain the quantified gene expression values from the arrays. The third is to mine the information from the gene expression data. The goal of this chapter is to present these issues in detail and show some of the possible software solutions.

This chapter is organized into four sections. First, we present the informatics problems emerging from the array fabrication stage. The second section describes the problem of signal detection in image processing of the arrays. The third section illustrates the variety of ways to quantify the signal and to perform quality measurements for assessing the reliability of the data and revealing problems during the array fabrication and hybridization processes. Finally, the fourth section describes some of the issues and solutions for mining and analysis of gene expression data.

ARRAY FABRICATION INFORMATICS

Array Fabrication and Hybridization Process

The array fabrication process may be conceptualized in four steps and their subprocesses, each of which has its own set of data and/or parameters, as shown in Figure 2. Once a gene expression experiment has been planned, the first step is to design the microarray chips that will contain the clones of interest. The chips are

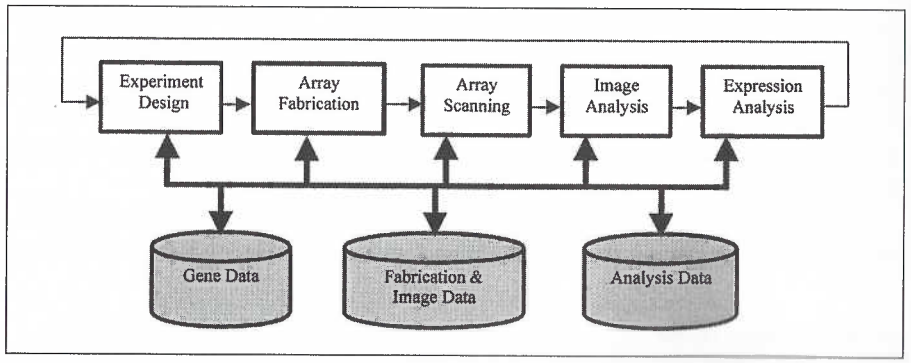

Figure 1. The overall microarray information processing system.

then manufactured using a robotic arrayer. Hybridizations are performed with either unamplified or amplified cDNA or RNA from sample tissues. Finally the florescence images of the hybridized chips are scanned and stored for image analysis.

In each of these four steps, a large amount of information is produced and needs to be handled appropriately. The use of this information is not limited to the array fabrication process. Some of it may be useful in the subsequent processes, such as the image processing and gene expression analysis. The informatics challenge in array fabrication is to maintain a vast amount of the information, organize it into a structured database for efficient access, and provide utilities for the design of microarray chips. These three informatics issues are the central theme of this section. After each of these issues are presented, solutions are proposed.

Keeping Information About the Process

The information generated in the stages of array fabrication and experimentation comes from multiple sources, as illustrated in Figure 2. It is essential to track the data as they pass through the various stages of the microarray fabrication process. Each step in this process has its own set of data and/or parameters. These are described below.

Plate data. Each plate may have a unique identification (ID), the user name, and the clone in each well.

Array data. Data related to an array may include the layout information such as the spot density, the spacing between adjacent spots, and, more importantly, information about the clones placed at each spot on the array.

Configuration of arrayer and array spotting. The arrayer configuration may

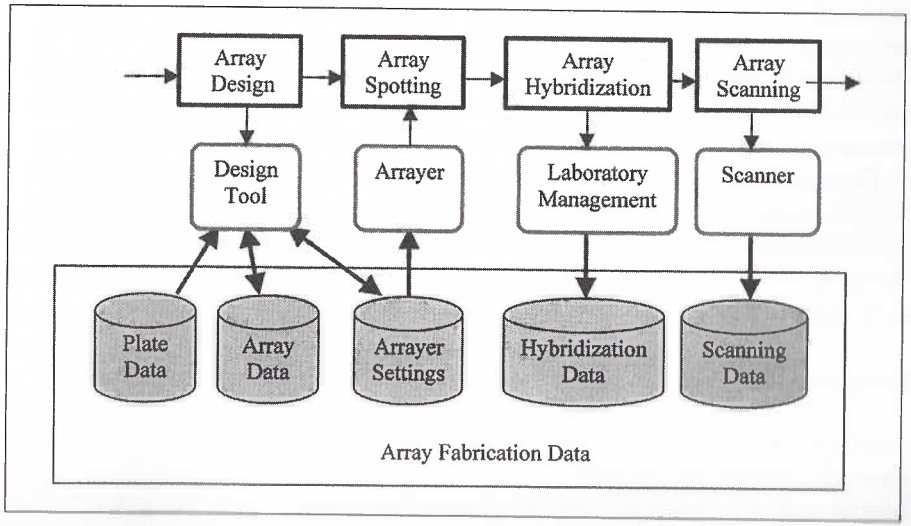

Figure 2. The microarray fabrication process.

include the number and type of pins, the number of pin cleaning cycles, and so forth. Some are parameters needed to configure the arrayer, and some are for tracking and logging purposes.

Array hybridization conditions. This may include the nature of the sample, the protocols, and other experimental conditions.

Array scanning conditions. This includes the type of scanner and the parameters used when scanning an array. From a conventional point of view, array scanning is not included in the array fabrication process, but from a laboratory information management point of view it makes sense to keep data regarding *physical* operations on arrays in the same place for tracking purposes.

The steady progress and increased density of microarray experiments is producing an explosion in the amount of data generated. The only solution is to employ an effective database management system.

Accessing the Array Information

The major purpose of building the database system for array informatics is to facilitate the access of information for experimental design, problem tracking and system tuning, replicating chips and experiments, and data sharing with other processes.

Experimental design. The ability to browse through a plate database or obtain spot densities of existing arrays can provide useful information for decision making, such as how many arrays to produce or the number of duplicates needed for each clone.

Problem tracking and system tuning. With an efficient database, the user can quickly identify a badly fabricated array or an erroneous data point, trace the source of the problem, and remedy it by fine tuning the array fabrication and experiment processes.

Replicating arrays and experiments. The user may want to replicate the same experiments on different arrays or vice versa. Having an efficient database system ensures that the information is available for replicating the processing steps.

Data sharing. The information in the database is useful not only for the array fabrication and experiments but also for subsequent processes, such as image processing and gene expression analysis. Image analysis software, such as ImaGene™ and AutoGene™, can take gene ID data as input and associate them with quantified values for each spot in the array.

Designing the Arrays

The informatics problem in the array design cannot be solved by a database alone. It involves the database infrastructure plus tools for navigating the database as well as software support for design of the array. The major components of this process are described below.

Laboratory information management. A management system is necessary that provides access interface to the underlying array fabrication database. This is used to monitor the overall fabrication process by letting the user query different aspects of the fabrication process, update data step by step according to the progress of the experiment, and track the various operations performed on each spot in each array.

Array design. An easy-to-use visual interface is needed that enables the user to design or adjust the array layout based on the type of mechanical arrayer used and the design objectives. The system could also export the various arrayer parameter settings, or even generate complete programs to control the arrayer robot directly.

Utilities for complex data manipulation. The information flow in the fabrication process involves several transformations that cannot be handled using simple database operations. One example is the transformation of clone data from a set of plates (plate data) into spot data on the array (array data). Generating the correct mapping is an essential component of array informatics since it relates the array data back to the plate data.

Solutions to Data Management

There are many solutions to the data management problem. Many companies and labs are developing their own specialized databases and user interfaces intended to work with internal projects. However, setting up an in-house microarray processing system is expensive and difficult to integrate with other modules such as image processing and gene expression data analysis. The ideal solution would be a standard, industry-wide scheme that is sufficiently flexible to handle diversities arising from the needs of different users. Such a solution would allow direct exchanges of microarray data between different labs and would foster collective efforts among multiple labs.

There are two kinds of diversities among users. One is due to the size of the labs. Users in small labs may have limited or no database tools in place, whereas more sophisticated users have already invested in database systems and are only interested in incorporating some useful functionality into their systems. The other kind of diversity is related to differences in the type of data being stored and tracked. Different users may use different kinds of fabrication methods or hybridization protocols, or they may store the plate/array data in different tables or formats (which may depend on the manufacturer). To deal effectively with these diversities, we have proposed and implemented a simple, yet complete, relational database schema that is itself immediately useable but can also be customized to suit specific needs of different users.

As shown in Figure 3, our scheme classifies the data into categories according to different stages in the experimental process. At the top level, there is a global structure recording management information. In each category, there is a *skeleton* table that links to the data in that category, and the top-level structure refers only to these skeleton tables. For each data category a set of tables is provided, but the

user may augment or replace them as desired.

Software support is needed to provide simple navigation and management interfaces to the database. When browsing user-defined data, special interfaces are required to ask the user for the tables information.

Tools for Array Design

One of the major challenges in the design of high-density arrays is establishing a map between the wells in the plates to the spots on the chip. We have developed a software system, called CloneTracker™, to address this problem. The software provides several tools to help the user in different stages of the fabrication process. In addition, the software provides database maintenance and query tools for tracking and maintaining information about plates, arrays, and sets of arrays. All the tools are integrated into a central user interface, as shown in Figure 4. The key features of this system are described below.

Visual array design interface. The layout of an array can usually be characterized by a handful of parameters. That is, it can be determined given the chip size,

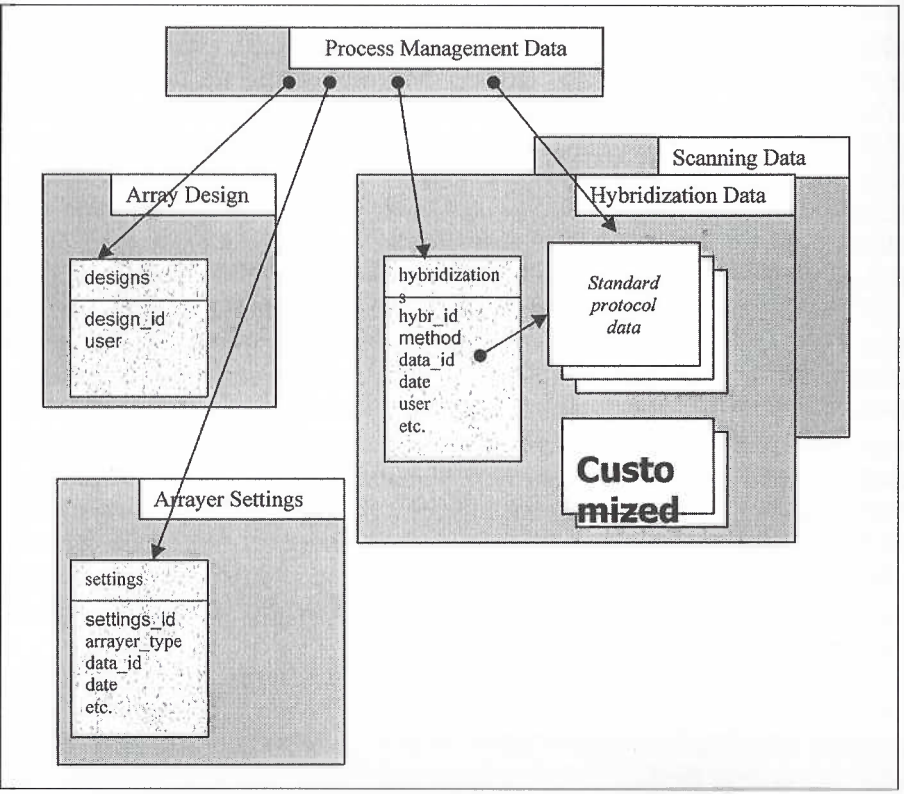

Figure 3. The informatics database.

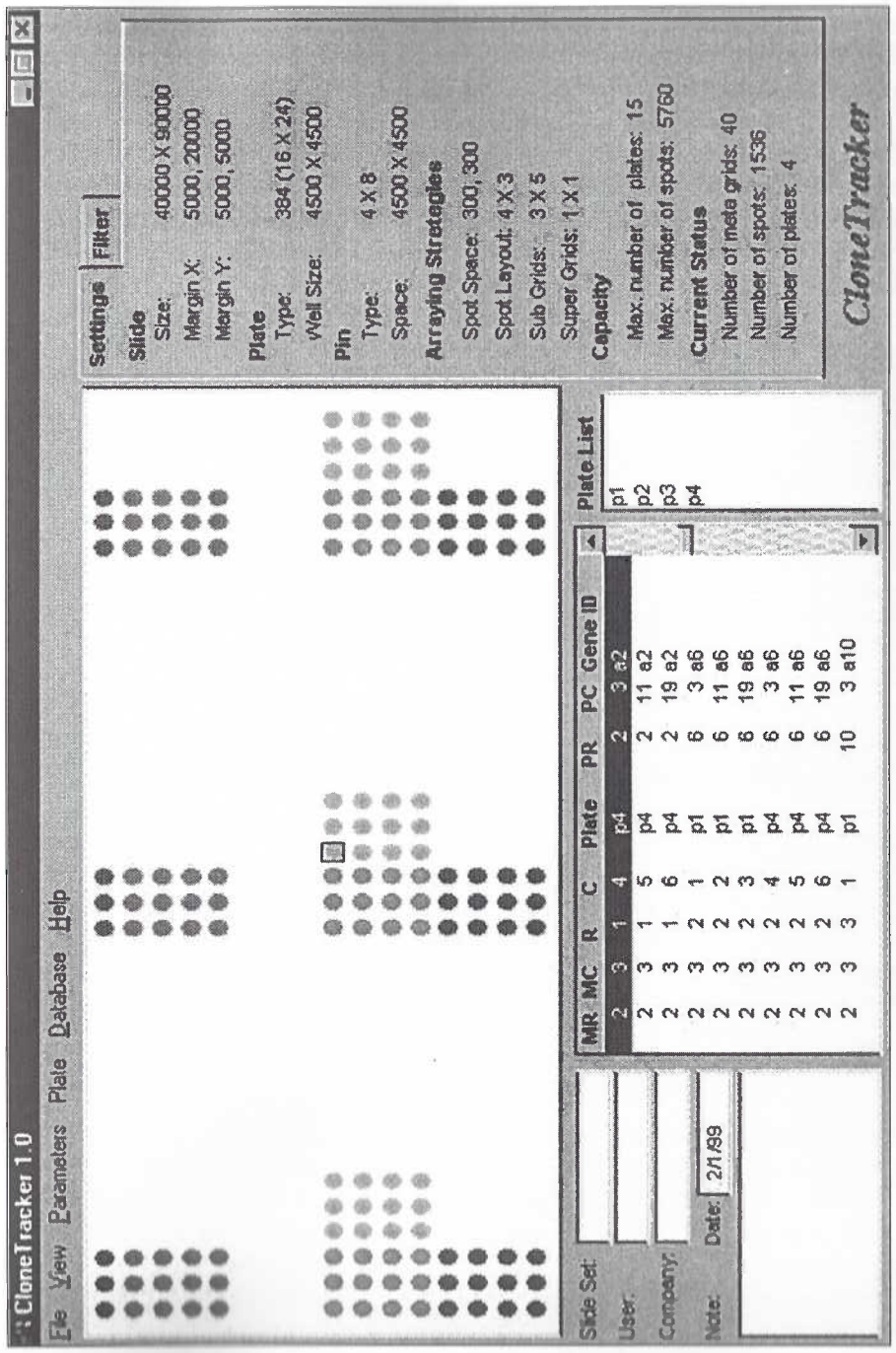

Figure 4. Array fabrication application user interface. (See color plate A14.)

pot spacing, pin number, arraying orientations, etc., regardless of the actual ar-ayer that manufactures it. By identifying a set of parameters, we can provide a impler interface for array design without digging down too much into arrayer programming details. This has two different uses. First, multiple arrays can be "virtually" designed and visualized before a specific arrayer program is written. The user can adjust different parameters, see the effects, and obtain useful information such as the spot density or the numbers of rows and columns. The parameters used in the final "optimum" design can then be used to set up the arrayer. Second, if the user already has a prearrayed slide, the software can assist in establishing the necessary association of gene IDs to each spot.

Generation of plate/array mapping. Given an array design, a map between vells of microtiter plates to spots on an array can be generated. With this mapping, it is straightforward to derive the plate and well location from a spot on a given array and vice versa, hence making the tracking of plate/array data easy.

Database maintenance and tracking. Data about all available plates, including plate numbers, size, and IDs for each clone in each well, as well as information about different arrays produced, are all maintained in the database for tracking and archival purposes.

Generation of programs for specific arrayers. Given a generic arrayer settings plus some arrayer-specific parameters, it is possible to generate programs for different arrayers. This is highly arrayer specific and also requires support from vendors. (They need to provide public interfaces.)

ARRAY IMAGE ANALYSIS I: SPOT DETECTION

The fundamental goal of array image processing is to measure the intensity of he arrayed spots and quantify their expression levels based on these intensities. n a more sophisticated and complete approach, the array image processing will lso assess the reliability of the quantified spot data and generate warnings for possible problems during the array production and/or hybridization phases. Before these operations can be performed, the spots in the images must be identified. This is the theme of this section.

We will start with enumerating the properties of DNA spots and issues in the microarray images. They are the central concerns in the design of image processing systems for microarray images. These concerns can be seen throughout this ection and the next. In fact, any effective image processing system must be designed based on the specific properties of the images to be dealt with.

This section proceeds with a discussion of a range of solutions to the spot detection problem, describing their advantages and disadvantages under various conditions of the images. The section ends with a brief description of the existing approaches used in software systems currently available. The issues of signal intensity measurements and quality measurements are presented in the next section.

Properties and Problems of Microarray Images

Microarray images consist of arrays of spots arranged in grids. All the grids have the same number of rows and columns of spots. These grids, called subgrids, are arranged with relatively equal spacing to each other, forming a complete array. Figure 5 shows an example of a complete grid structure in a microarray image. There is a 2×2 array in this image, with each subgrid having 20×16, or 320, total spots. This microarray structure is formed with a robotic arraying system that employs pins to create the array. Each pin deposits DNA material in each of the subgrids. The microarray shown in Figure 5 has been produced with

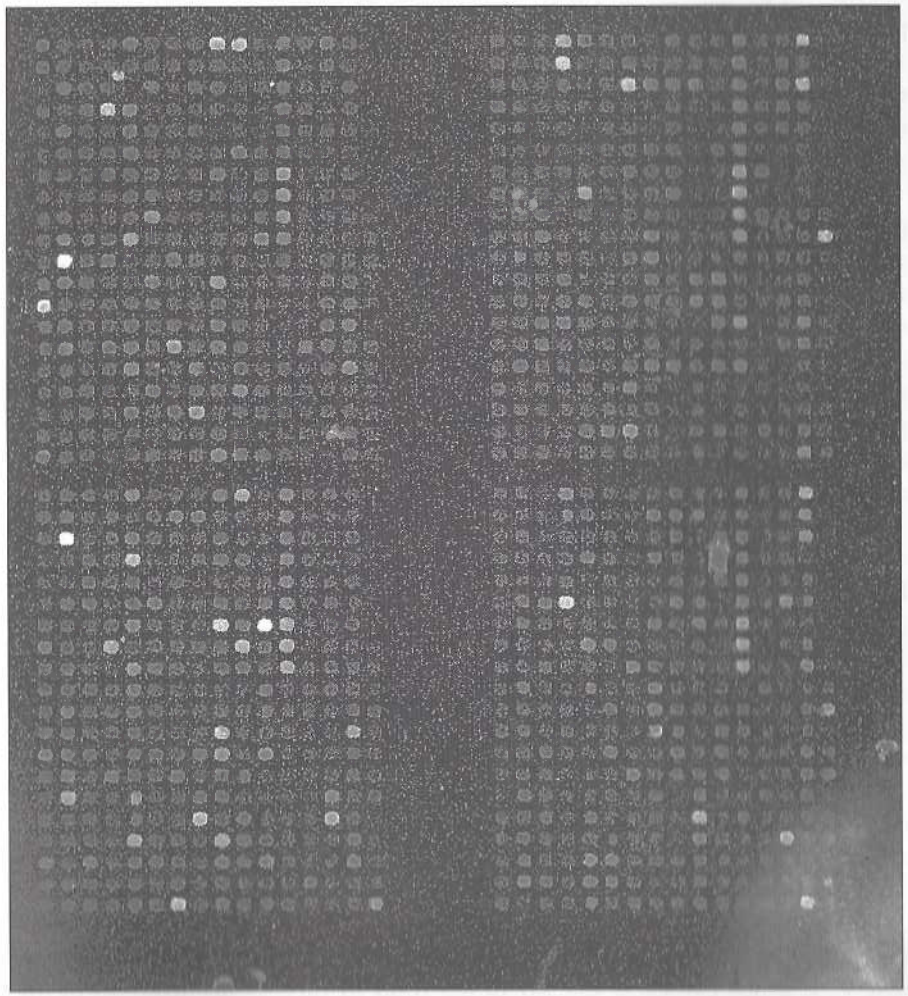

Figure 5. Microarray scanned image containing four subgrids.

a four-pin assembly organized into a 2 × 2 configuration.

Ideally, the microarray images should have the following properties: *(i)* all the subgrids are the same size; *(ii)* the spacing between the subgrids is regular; *(iii)* the distance between rows and columns should be the same; *(iv)* the location of the spots should be centered on the intersections of the lines of the rows and columns (the canonical position); *(v)* the size and shape of the spots should be perfectly circular and the same for all the spots; *(vi)* the location of the grids is fixed in images for a given type of slides; *(vii)* no dust or other contamination is on the slide; and *(viii)* minimal and uniform background intensity across the image.

If all these idealized conditions were satisfied, the image processing task could be easily accomplished by a simple computer program. An array of circles with the defined dimensions and spacing could be superimposed on the image. The pixels falling inside these circles would be considered signal and those outside would be background. However, some of these idealized conditions are often violated in a real-world setting. These violations may be conceptualized in four categories: spot position variation, spot shape and size irregularity, contamination, and global problems that affect multiple spots.

Spot variation is caused, in part, by the mechanical limitations in the arraying process. Researchers are motivated to put as many spots as possible onto a single array to increase the data content of a given hybridization. The fundamental concern in the fabrication of the array is to make sure that each spot is spatially distinct within the arrays. Whether the spots are spaced perfectly is a secondary concern. There are several sources contributing to the nonuniform positioning of the spots within the array, including inaccuracies in robotic systems, the printing apparatus, and the platter that holds the slides. The spatial mapping between slides and scanned images may not be perfectly linear, causing the spacing between the adjacent columns to be unequal across the arrays. Some of the spots may be "missing" from the grid due to lower than measurable expression level or empty wells in the plates used for array fabrication. Some of these issues are observed in Figure 5. Because the spot sizes are on the scale of 100 to 200 μm, fixing the location of the quantitation grids on each image means the location of spots would have to be accurate to ±10 μm (condition *vi* above). This requirement is unrealistic to expect from many research laboratories. Also, because the spots in a subgrid are typically printed with multiple pins, aligning the subgrids accurately would mean that the pin's spacing would have to be ±10 μm as well (conditions *i–iv* above), which is another difficult requirement for many research laboratories. The existence of these positional variations in microarray images demands an intelligent spot localization operation, which must be done either by human intervention or by computer vision algorithms.

In addition to the spot position variation, the shape and size of the spots may fluctuate significantly across the array. The sizes of the droplets of DNA solution may vary, which can cause the size of the spots to vary. Concentrations of the DNA and salt in solution may change over time, making the shape of spots

deviate from a typical circle; furthermore, the density of DNA can vary within a given spot. The surface properties of low-quality slides can vary across the surface. All these factors perturb the shape and size of the spots from the ideal. Figure 6 illustrates some of the irregularities of the spots, such as the shape and intensity variation inside of the spots. To obtain accurate measurements of the hybridization signals, quantitation of pixels needs to be done intelligently to take account of these problems.

Contamination is also a source of difficulty in microarray image processing. From the procedure of spotting to the hybridization step, airborne dust can contaminate the array and produce fluorescent signal in the scanned images. Impurities on the glass surface can also cause speckles in the image. Uneven drying of the arrayed samples on the surface may cause unevenness in the images. Figure 7 shows an example of contamination in both the background areas and inside the spots, some of which produces intense fluorescence. The solution to these problems is to identify the contaminants and remove them from the image before applying measurements.

There are also other factors affecting the quality of data in a larger area of the slides, such as glass quality, temperature nonuniformity across the slide, scratches, or bubbles that occur during hybridization. Figure 5 demonstrates a few global problems that can occur in microarray images. The two lower corners of the image have elevated background, some of which obscures the array. Also, three spots in the fourth subgrid are smeared into each other.

All these global and local quality problems need to be handled either by human inspection or by software systems. In this section, we present a number of solutions to spot localization and the methods of signal pixel determination.

Design of the Signal Detection Process for Microarray Technology

The above four issues in microarray image analysis may be resolved in two steps. The first step is to localize the position and size of each spot, which addresses the spot position variation problem. The second step addresses the shape irregularity and contamination problems through a signal pixel identification process. The last step is to deal with the spatially global problems. The first two steps are discussed in this section. The process in the third step requires global quality information and is discussed in "Importance of Quality Measurements" in the next section.

Spot finding methods for resolving the spot localization errors. The goal of the spot finding operation is to locate the signal spots in images and estimate the size of each spot. There are three different levels of sophistication in the algorithms for spot finding, corresponding to the degree of human intervention in the process. These are described below in order of most to least amount of manual intervention.

Manual spot finding. This method is essentially a computer-aided image

processing approach. The computer itself does not have any visual capabilities to "see" the spots, rather, it provides tools to allow users to tell the computer the location of each signal spot in the image. Typically, a grid frame is given that the user can place manually on the image and then manipulate to fit the spots in the image. Because the spots in the image may not be spaced evenly, the user may need to adjust the grid lines individually to align with the arrayed spots. The size of each circle may also need manual adjustment to fit the size of a particular spot. To conduct an accurate measurement, the manual spot finding method is prohibitively time-consuming and labor-intensive for microarray images with

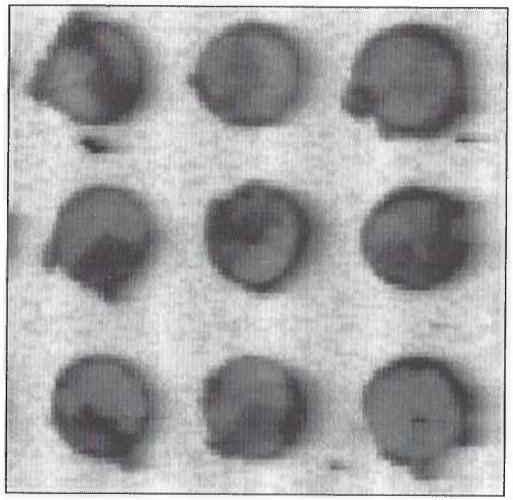

Figure 6. Irregularity of the spot shape. The shape of the spot may deviate from a perfect circle, and the intensity inside the spot may vary considerably.

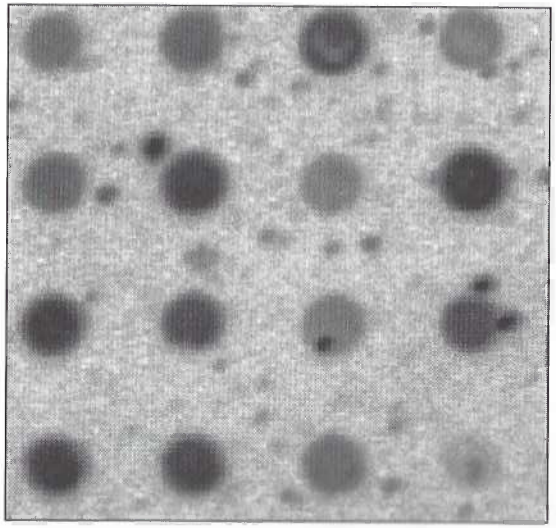

Figure 7. Contamination in microarray images. Punctate contamination is visible both in the background and inside of the spot signal regions.

thousands of spots. Thus, considerable imprecision may be introduced due to human errors, particularly with arrays having irregular spacing between the spots and large variation in spot sizes.

Semiautomatic spot finding. The semiautomatic method requires some level of user interaction. This approach typically uses algorithms for automatically adjusting the location of the grid lines, or individual grid points after the user has specified the approximate location of the grid. What the user needs to do is to tell the program where the outline of the grid is in the image. For example, the user may need to put down a grid and adjust the size of it to fit on the array of the spots, or to tell the program the location of the corners of the subgrids in the images. Then the spot finding algorithm adjusts the location of the grid lines, or grid points, to locate the arrayed spots in the image. User interface tools are usually provided by the software to allow for manual adjustment of the grid points if the automatic spot finding method has not correctly identified each spot.

This approach could potentially offer great time savings over the manual spot finding method since the user needs only to identify a few points in the image and make minor adjustments to a few spot locations if required. The key issue for spot finding algorithms here is to identify the spots correctly even at very low levels of intensity. In addition, the algorithm must deal effectively with two opposing criteria. First, due to variation in spot position, as described earlier, the algorithm must tolerate a certain degree of irregularity in the spot spacing. At the same time, the algorithm must not be "distracted" by contaminants that could be adjacent to a true arrayed spot. Such an algorithm has been developed and implemented in BioDiscovery s ImaGene™ software.

Automatic spot finding. The ultimate goal of array image processing is to build an automatic system that utilizes advanced computer vision algorithms, to find the spots without the need for any human intervention. Such a method would greatly reduce human intervention, minimize the potential for human error, and offer a great deal of consistency in the quality of data. One such system, called AutoGene™, has been developed by BioDiscovery. This tool only requires the user to specify the configuration of the array (i.e., number of rows and columns), and will automatically search the image for the correct grid position. In many tests, the software has been able to locate grids as accurately as those obtained by manual placement of the grid.

Methods of spatial segmentation of signal and background pixels. After the spot location is determined in the image, a small patch around that location (target region) can be used to quantitate the spot intensity level. The next step is to determine which pixels in the target region are signal and which are background. This operation is called signal or image segmentation in computer vision terminology. At this stage, size and shape irregularities of the spots and any contamination problem in the images are the major concerns to the algorithm design. A number of methods have been developed with different levels of sophistication. Their advantages and disadvantages are described below.

Pure space-based signal segmentation. Methods of this class use purely spatial information from the result of spot finding to segment out signal pixels. After the spot finding operation has been completed, the location and size of the spot are determined. A circular mask is placed in the image at the determined position and size to separate the signal from background. It is assumed that the pixels inside the circle are due to the true signal and that those outside are background. Measurements are then performed on these classified pixels. These types of methods are optimal when the spot finding operation is effective (i.e., when spots have been correctly located and sized, the spot shapes are close to perfect circles, and no contamination is present).

However, irregularities of the spot shape and sizes are rules rather than exceptions in many microarray images. Whenever these conditions occur, the accuracy of the measurements is largely compromised. In addition, spot contamination is still an issue in many microarray images. An example is shown in Figure 8 to illustrate the problem encountered with this method. In this example, nearly 50% difference is observed between mean measurements made before and after the removal of contamination from the image.

In addition to quantification inaccuracies generated due to the presence of contaminants in the image, problems arise due to variation in the shapes of the spots. There are two broad classes of spot shapes that will cause problems with pure space-based signal segmentation. First, doughnut-shaped spots, which are often seen with many arraying systems, contain many nonhybridized pixels within the circular spot area. These pixels will be mistakenly classified as signal pixels using this segmentation approach. Second, noncircular spots (e.g., more elliptical spots) cannot be fit perfectly with a round circle, thus causing some signal pixels to be considered as background and vice versa.

This method of quantification is used by a number of available microarray image analysis software tools and would be appropriate for "quick and dirty" look at the image data. However, it is not sufficiently sophisticated to ensure accuracy of the quantified data in most cases.

Pure intensity-based signal segmentation. Methods of this class use intensity information exclusively to segment out signal pixels from background. They assume that the brightness intensities of signal pixels are statistically higher than that of the background pixels. As an example, suppose that the target region around the spot taken from the image consists of 40×40 pixels and that the spot is about 20 pixels in diameter. From a total of 1600 pixels in the region, about 314 ($\pi \times 10^2$) pixels or approximately 20% are signal pixels that are expected to have intensity values higher than the background pixels. To identify the signal pixels, all the pixels from the target region are ordered in a one-dimensional array from the lowest intensity pixel to the highest one $\{p_1, p_2, p_3, \ldots p_{2500}\}$, where p_i is the intensity value of the pixel of the i-th lowest intensity among all the pixels. If there are no contaminants in the target region, the top 20% of the pixels in the intensity rank are classified as signal pixels.

The advantage of this method is its simplicity and speed. The method works well on computers with moderate computing speed and when the spots are high intensity relative to the background pixels, but it works less well on spots of low intensities or noisy images.

When the signal intensity is low, the intensity distribution of the signal overlaps largely with background. The signal and background pixels are not separable based on their intensity values alone. Applying this method will produce biased estimates of the signal and background intensity values. The shortcoming of the intensity-based segmentation method can be remedied by exploiting spatial information. One simple method of obtaining spatial information is to segregate the target region into potential signal and background regions before conduct-

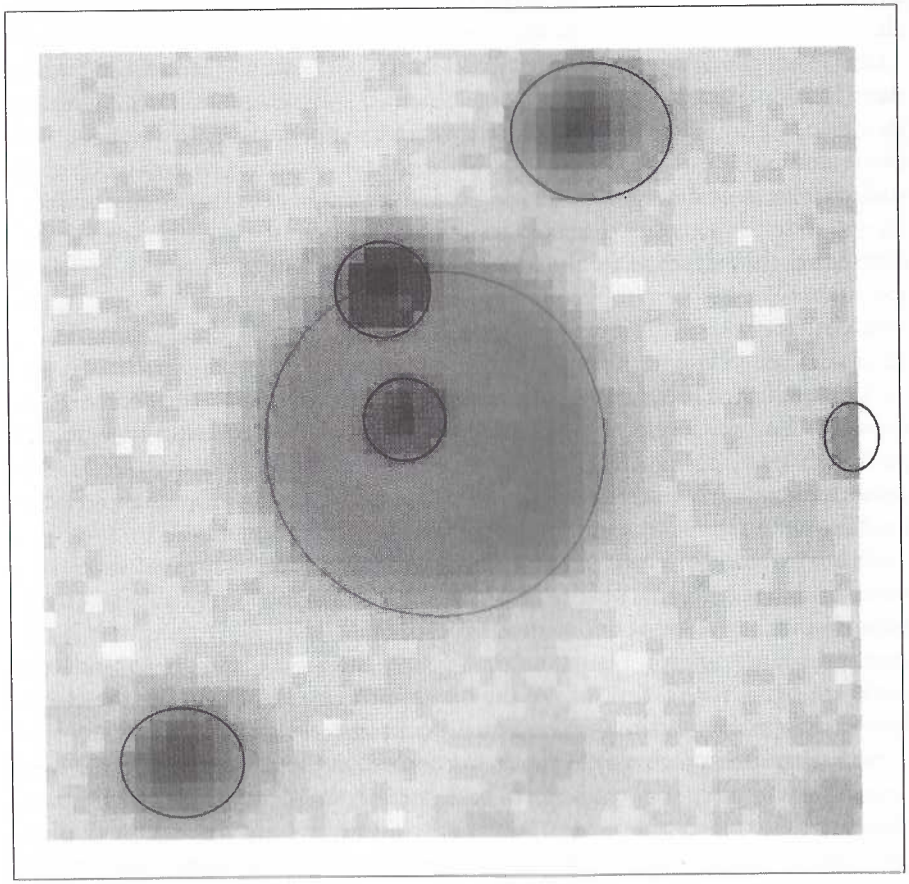

Figure 8. Signal identification from the spot finding operation. All the pixels inside the large circle are considered signal according to this method. When contaminants exist (small circles), measurement errors are introduced into both the signal and background values. The mean signal intensity using purely spatial segmentation is 1230. After the contaminant pixels are removed, the signal mean is 834 or a 48% difference. The mean background values are 113 and 28, respectively, or a 300% difference before and after the contamination removal.

ing intensity-based segmentation. A circle can be placed in the target region based on the result of the spot finding. Pixels inside the circle are then temporarily classified as signals, and those outside are background pixels. Statistical differences in the two regions are then assessed and compared for signal segmentation. In the following discussion, we describe two such methods: the Mann-Whitney test and the trimmed measurement method.

Mann-Whitney segmentation. Chen et al. (1) introduced a nonparametric statistical method, the Mann-Whitney test, to segment out signal pixels from target regions. After the spot finding operation, a circle is placed in the target region to demarcate the spatial region of the spot. Because the pixels outside the circle are assumed to be background, the statistical properties of these background pixels can be used to determine which pixels inside the circle are signal pixels. The Mann-Whitney test is the method used to obtain a threshold intensity level. Pixels inside the circle that exceed the threshold intensity are identified as signal. This method works very well when the spot location is found correctly and when there is no contamination in the image. However, when contaminated pixels exist inside of the circle, they are incorrectly scored as signal pixels. If there are contaminated pixels outside the circle or if the spot location is not found correctly such that some of the signal pixels are outside the circle, these high-intensity pixels will raise the intensity threshold level; consequently, signal pixels with intensities lower than the threshold will be incorrectly scored as background. Applying the Mann-Whitney test to the example in Figure 8, using α value of 0.0001, the mean intensities are 1230 for signal pixels and 33 for background pixels. These values deviate 50% and 20% from the true values, respectively.

This method also has its limitations when dealing with weak signals and noisy images. When the intensity distribution functions of the signal and background are largely overlapping, classification of pixels based on an intensity threshold is prone to errors, resulting in measurement biases. This scenario is similar to what has been discussed in the pure intensity-based segmentation method.

Trimmed measurements method. This approach is another method that combines both spatial and intensity information in segmenting signal pixels from background. The logic of this method proceeds as follows. After the spot is localized and a target circle is placed in the target region, *most* of the pixels inside the circle are signal pixels, and *most* of the background pixels are outside the circle. Due to shape irregularities, some signal pixels may lie outside the circle, and some background pixels may lie inside the circle. These pixels are considered outliers in the sampling of the signal and background pixels. Similarly, contaminant pixels can also be considered as outliers in the intensity domain. These outliers will severely change the measurement of the mean and total signal intensity. To remove the impact of the outliers on these measurements, one may simply "trim off" a certain percentage of signal and background pixels.

Typically, a certain percentage of pixels from the high end of the intensity distribution, say 10%, is trimmed off from the pixels inside the circle, due to the

high possibility that such pixels may be contaminants. A certain percent of pixels at the low end of the intensity distribution, say 20%, is trimmed off from the pixels inside the circle, due to the high possibility that such pixels may be background. Whatever remains inside of the circle after trimming is used for quantification. The exact amount to be trimmed depends on the effectiveness of the spot finding process and the quality of the image, such as the extent of size and shape of irregularities. A good estimate for these thresholds can be determined empirically. As long as the image quality does not significantly change (e.g., major increase in contamination), this method is effective.

The major advantage of the trimmed measurement method is the robustness of the measurement against outliers, at the expense of accuracy. When there are no outliers, trimming will reduce the measurement accuracy by reducing the number of pixels used in the calculation of the mean signal value. However, if there is a significant number of pixels per spot and only a small percentage of the pixels are removed, this method can yield highly accurate values. Furthermore, by averaging the measurements from replicate spots, this method should achieve very good performance. Figure 9 illustrates the application of such a method used on the data in Figure 8. The measurements derived from this method deviate about 30% from the true value, which is less than the other methods discussed above.

The major limitation of this method is the loss of the geometric information of the spots due to trimming. However, this information is mainly important if it is used as part of the quality measurement score associated with a spot. Its considerable computational needs demand a fully automated system for practical implementation. In semiautomatic systems, computing speed is a major concern. The trimmed measurement method provides an optimal blend of speed and accuracy.

Integrating spatial and intensity information for signal segmentation. The two methods discussed above use a minimal amount of spatial information. They use target circle derived from the spot localization process to improve the detection of signal pixels. Their design priority is to conduct the spot intensity measurements with minimal computation. These methods are useful in semiautomatic image processing because the speed has high priority and the user can inspect the data visually.

In a fully automated image processing system, the accuracy of the signal pixel classification becomes a central concern. Not only does the correct segmentation of signal pixels offer accurate measurement of signal intensity, but it also permits multiple quality measurements based on the geometric properties of the spots. These quality measures can be used to ask for human intervention for spots of questionable quality that are flagged after the completion of an automated processing run. The correct classification of the signal pixels can be realized by algorithms that use information from both spatial and intensity domains. Such an approach has been implemented in the AutoGene software mentioned earlier.

ARRAY IMAGE ANALYSIS II: DATA QUANTIFICATION AND QUALITY MEASUREMENT

Methods of Quantitation

On a single microarray chip, the expression levels of many genes are in parallel. Under the proper conditions, the total fluorescence intensity of a spot is proportional to the expression level of a gene. These conditions are as follows:

1. The preparation of the probe cDNA (through reverse transcription of the extracted mRNA) solution is performed such that the probe cDNA concentration in the solution is proportional to the mRNA in the tissue.
2. The hybridization experiment is performed such that the amount of cDNA binding to each spot is proportional to the partial concentration of each cDNA species in the probe solution.
3. The amount of cDNA target deposited at each spot during the chip fabrication is constant and in approximately 10-fold excess relative to the most abundant species in the probe solution.
4. There is no contamination on the spots.
5. The signal pixels are correctly identified by the image processing.

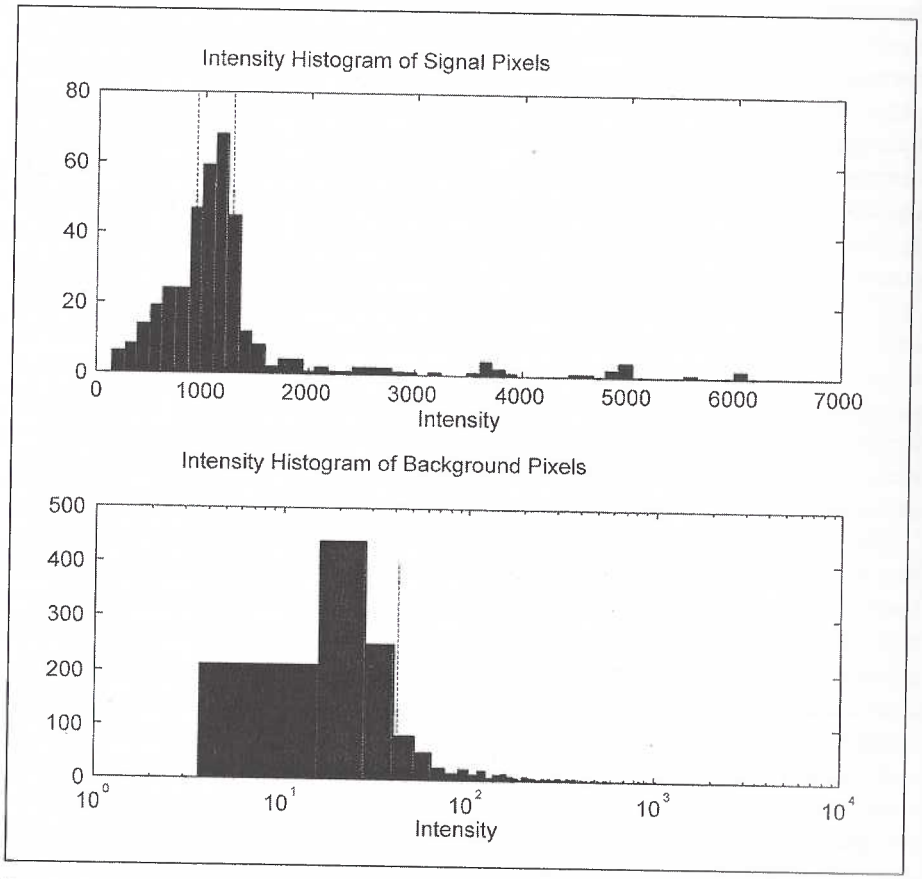

Figure 9. Trimmed measurement method for signal pixel identification. Top, plot of the intensity histogram of the pixels inside the target circle. Thirty percent of the pixels are trimmed off at the lower intensity side (left of the left vertical line). Twenty percent of the pixels are trimmed off the high-intensity side (right of the right vertical line). Bottom, plot of the histogram of the pixels outside the target circle. Thirty percent of pixels are trimmed off at the high-intensity side of the histogram (right of the vertical line). The abscissa is a log scale to show the detail on the lower intensity side. The mean intensities are 1100 for signal and 20 for background. They differ by 32% and -29% from the true value. (See color plate A15.)

In the following discussion, we assume that conditions 1 and 2 are satisfied. Whether these two conditions are truly satisfied is determined through the design of the experiments. For the quantitation measurements, the more closely conditions 1 to 5 are followed the better. Often, conditions 3, 4, and 5 are violated to varying degrees. The DNA concentrations in the spotting procedure may vary from time to time and spot to spot. Higher or lower concentrations may result in altered signals. When adjacent spots overlap, the signal intensity corresponding to the contaminated region is not measurable. The image processing may not correctly identify all the signal pixels; thus, the quantification methods should be designed to address these problems. The commonly used methods are total, mean, median, mode, volume, intensity ratio, and the correlation ratio across two channels. The underlying principle for judging which one is the best method is based on how well each of these measurements correlates to the amount of the DNA probe hybridized to each spot location.

Total. The total signal intensity is the sum of the intensity values of all the pixels in the signal region. As has been indicated above, total intensity is sensitive to variations in the amount of DNA deposited on the surface, the existence of contamination, and anomalies in the image processing operation. Because these problems occur frequently, the *total* may not be an accurate measurement.

Mean. The mean signal intensity is the average intensity of the signal pixels. This method has certain advantages over the *total*. Very often the spot size correlates with the samples and pins used in the arraying step. Measuring the mean will reduce the error caused by the variation of the amount of DNA deposited on the spot. With advanced image processing allowing for accurate segmentation of contaminated pixels, the mean is perhaps the best measurement method.

Median. The median of the signal intensity is the intensity value that splits the distribution of the signal pixels such that the number of pixels above the median intensity is the same as the number below the median intensity. The median is a landmark in the intensity distribution profile. The advantage of choosing this landmark as the measurement derives from the resistance of the median value to outliers. As discussed in the last section, contamination and problems in the image processing operation introduce outliers in the sample of signal pixels. The mean measurement is very vulnerable to these outliers. When the distribution

profile is unimodal, the median intensity value is very stable and close to the mean. In fact, if the distribution is symmetric in both high and low intensity sides, the median is equal to the mean. Thus, if the image processing operation is not sophisticated enough to ensure the correct identification of signal, background, and contaminated pixels, the *median* is a better choice than the *mean*. An alternative to the median measurement is to use a trimmed mean, as was discussed in the of "Trimmed measurements methods" section above. The trimmed mean estimate is obtained by trimming a certain percentage of pixels from the high- and low-intensity sides of the distribution.

Mode. The mode of the signal intensity is the "most likely" intensity value and can be measured as the intensity level corresponding to the peak of the intensity histogram. It is also a landmark in the intensity distribution. Thus, it enjoys the same robustness against outliers offered by the median. The tradeoff is that the mode will be more unstable than the median when the distribution is multimodal. This is because the mode value will be equal to one of the modals in the distribution, depending on which one is the highest. When the distribution is unimodal and symmetric, mean, median, and mode measurements are equal. Often the difference between mode and median values can be used as an indicator of the degree to which a distribution is multimodal.

Volume. The volume of signal intensity is the sum of the signal intensity above the background intensity. It may be computed as:

$$(\text{signal mean} - \text{background mean}) \times \text{signal area}.$$

This method adopts the argument that the measured signal intensity has an additive component due to the nonspecific binding, and this additive component is the same as that of the background. This argument may not be valid because the nonspecific binding in the background is different from that in the spot. Perhaps the better way is to use blank spots for measuring the strength of nonspecific binding inside spots. It has been shown that the intensity on the spots may be lower than it is on the background, indicating that the nature of the nonspecific binding is different between what is on the background and what is inside of spots. Perhaps the background intensity should be used for quality control rather than signal measurement.

Intensity ratio. If the hybridization is two-color and the scanning measurements are taken in two channels, then the intensity ratio between the channels is often an important quantified value of interest. This value will be insensitive to variations in the exact amount of DNA spotted since the ratio between the two channels is being measured. This ratio can be obtained from the mean, median, or mode of the intensity measurement, obtained as discussed above, for each channel.

Correlation ratio. Another way of computing the intensity ratio is to perform a correlation analysis across the corresponding pixels in two channels on the same slide. It computes the ratio between the pixels in two channels by fitting a straight line through a scatter plot of intensities of individual pixels. This line must pass through the origin, and the slope is the intensity ratio between the two

Table 1. Measurements Using Different Quantification and Image Processing Methods

Method	Mean	Median	Mode	Total
Full AutoGene	834	935	998	384 450
Trimmed	1103	1109	1109	207 375
Pure spatial	1233	1076	1063	464 765

The pure spatial method is discussed in "Pure space based signal segmentation." It places a circle in the signal region and identifies the pixels inside of the circle as signal and outside as background. The trimmed method in the table is discussed in "Method of trimmed measurements." Thirty percent of the signal pixels were trimmed from the low-intensity side of the distribution and 20% from the high-intensity side. For background pixel identification, pixels in a 3-pixel-wide ring outside the signal circle were discarded due to their potential of containing signal pixels "bleeding" into this region. In addition, 30% of the remaining background pixels were trimmed from the high-intensity side. The measurements were performed after trimming. Full AutoGene identified the signal, background, and contamination pixels automatically, which are considered the true values based on visual inspection of the segmentation results. The measurements were performed after the regions were segregated.

channels. This method may be effective when the signal intensity is much higher than the background intensity. The motivation behind using this method is to avoid the signal pixel identification process. However, for spots of moderate to low intensities, the background pixels may severely bias the ratio estimation of the signal toward the ratio of the background intensity. The only remedy to this problem is to identify the signal pixels prior to performing correlation analysis. Then the advantage of applying this method becomes unclear, and the procedure suffers the same complications encountered in the signal pixel identification methods discussed above. Thus its advantage over the intensity ratio method may not be warranted.

Table 1 lists the results of mean, median, mode, and total quantification measurements using three signal pixel identification methods. The image data are displayed in Figure 8. Because of the existence of contamination in the image, the mean method with full AutoGene is optimal. The median and mode measurements with full AutoGene are about 10% to 15% deviated from the mean. With the trimmed method, the three measurements are very similar and represent about 30% deviation from the mean with full AutoGene. Without trimming, the mean measurement produces a 50% deviation; however, median and mode estimations are essentially unchanged.

We derive the following recipe for selecting the measurement methods using different signal detection processes. With full AutoGene and the trimmed method, the mean is the best measurement. Without trimming, the median is the best choice. Although mode provides the same value as median in this case, because it is not stable when the intensity distribution is multimodal, median is a better choice.

Importance of Quality Measurements

An important step in processing microarray images is to assess the reliability of the data obtained and report the problems in the images that may be arising from array fabrication process and experiments. These tasks may be conducted by human intervention or by image processing software. The importance of these tasks cannot be overstated. Without assessing the reliability of the data, the conclusions drawn from analyzing such data may be misleading. Very often such errors can lead to a false hypothesis or overlooking important findings.

The reliability of the data is affected by multiple factors, ranging from problems in array fabrication to problems in image processing. In a high-throughput gene expression analysis system, the reliability assessment needs to be done with a software system. Figure 10 shows an example in which quality measurements can be used to improve the reliability of the data. The data were obtained from an

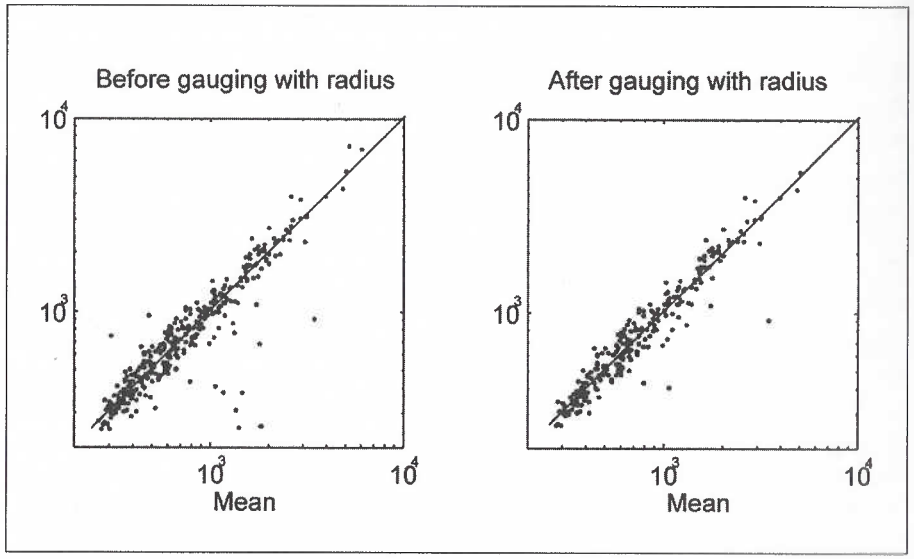

Figure 10. Reliability improvement following the quality control. Scatter plots of the mean values obtained from duplicated spots. The axes represent the mean pixel intensity values. The image was processed with AutoGene. Left, data from all the spots in the image. Right, plot of 75% of the spots in the left plot. The 25% discarded spots have an excessively large or small radius. The coefficient of variance is 27% for the left plot and 19% for the right, a reduction of 30%.

image file processed with AutoGene. The coefficient of variance [(standard deviation)/mean] was 27%. Using spot size as the quality measurement, we discarded about 25% of the spots that had excessively large or small diameters; the coefficient of variance was then reduced to 18.5%, suggesting an improvement in the reliability of the data after discarding low-quality spots.

DATA ANALYSIS AND VISUALIZATION TECHNIQUES

The image processing and analysis steps produce a large number of quantified gene expression values. The data typically represent thousands or tens of thousands of gene expression levels across multiple experiments. To make sense of this many data, it is essential to use multiple visualization and statistical analysis techniques. One of the most typical microarray data analysis goals is to find statistically significant up- and down-regulated genes. Another goal is to find functional groupings of genes by discovering a similarity or dissimilarity among gene expression profiles, or predicting the biochemical and physiological pathways of previously uncharacterized genes.

In the following sections, visualization approaches and algorithms implementing various multivariant techniques are discussed that could help realize the above-mentioned goals and provide solutions to the microarray research community.

Scatter Plot

Probably the simplest analysis tool for microarray data visualization is the scatter plot. In a scatter plot each data point represents the expression value of a gene in two experiments, one assigned to the x-axis and the other to the y-axis (Figure 11). In such a plot, genes with equal expression values in both experiments fall along a diagonal "identity" line. Genes that are differentially expressed fall off the diagonal. The larger the deviation from the identity line, the larger the difference in the expression of a given gene between two samples. Absolute expression levels can also be readily visualized in this plot in that the greater the expression of a given gene the further the data point lies from the origin.

Principal Component Analysis

It is easy to see how the scatter plot is an excellent tool for comparing the expression profile of genes in two samples. Three samples could be plotted and compared in a three-dimensional scatter plot. Three-dimensional plots can be rendered and manipulated on a computer screen. However, when more than three samples are analyzed and compared, the scatter plot cannot be used. In the case of 20 samples, for example, we cannot draw a 20-dimensional plot. Fortu-

nately, there are mathematical techniques available for dimensionality reduction, such as principal component analysis (PCA), in which high-dimensional space can be reduced down to two or three dimensions, which can be plotted. The goal of all dimensionality reduction methods is to perform this operation while

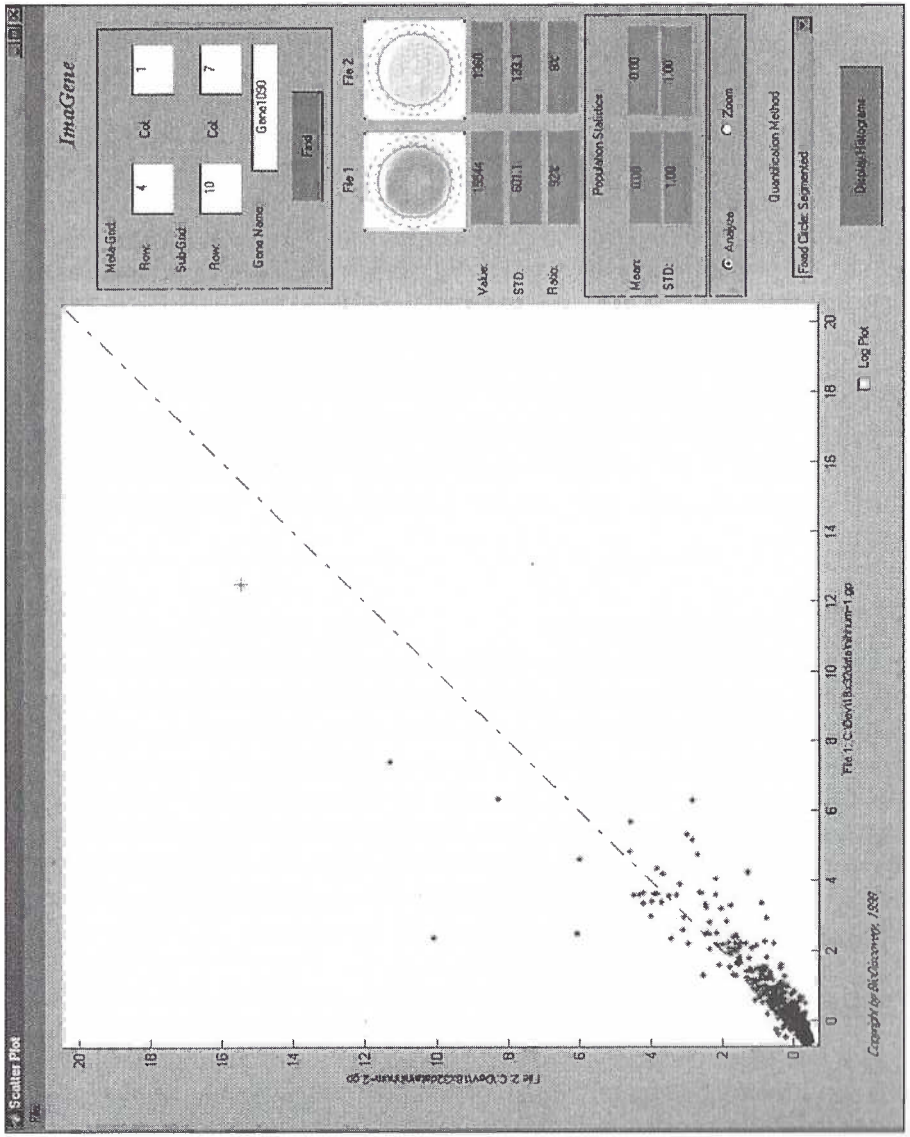

Figure 11. Scatter plot of two experiments. Each point in the plot represents a single gene. The position of the point on the graph reveals the absolute expression level of the gene, such that highly expressed genes lie far from the origin. Genes that are expressed at the same level in two samples fall along the identity (45° angle) line. (See color plate A16.)

preserving most or all the variances of the original dataset. In other words, if two points are relatively "close" to one another in the high-dimensional space, they should be relatively close in the lower dimensional space as well. In general, it is not possible to maintain this relationship perfectly. Imagine if a three-dimensional spherical orange peel were rendered flat on the two-dimensional surface of a table. Most of the points lying on the surface of the orange peel would keep their relationship to their neighbors. However, we must tear the orange peel in several places, thus breaking the neighboring relationship between some points on the orange peel to lay it flat.

PCA is a method that attempts to preserve the neighboring relationships as much as possible. Figure 12 is a three-dimensional plot of 600 genes in 21 experiments indicating the position of all 600 genes with respect to the first three principal components. Each "principal component" is made up of a linear combination (summation) of all the 600 genes each weighted by a different value. This multivariate technique is frequently used to provide a compact representation of large amounts of data by finding the axes (principal components) on which the data vary the most. In PCA the coefficients for the variables are chosen such that the first component explains the maximal amount of variance in the data. The second principal component is perpendicular to the first one and explains the maximum of the residual variance. The third component is perpendicular to the first two and explains the maximum of the still remaining variance. This process is continued until all the variance in the data is explained. The linear combination of gene expression levels on the first three principal components could easily be visualized in a three dimensional plot (Figure 12). This method, just like the scatter plot, provides an easy way of finding outliers in the data such as genes that behave differently than most of the genes across a set of experiments. It can also reveal clusters of genes that behave similarly across different experiments. PCA can also be performed on the experiments to find out possible groupings and/or outliers of experiments. In this case, every point plotted in the three-dimensional graph would represent a unique microarray experiment. Points that are placed close to one another represent experiments that have similar expression patterns. Recent findings show that this method should be able to detect moderate-sized alterations in gene expression (3). In general, PCA provides a rather practical approach to data reduction, visualization, and identification of outlier genes and/or experiments.

Parallel Coordinate Planes

Two- and three-dimensional scatter plots and PCA plots are ideal for detecting significantly up- or down-regulated genes across a set of experiments. These methods, however, do not provide an easy way of visualizing the progression of gene expression over several experiments. These types of questions usually come up in time series experiments in which, for instance, gene expression is measured at 2-hour intervals. The important question in this case is how gene expression

values vary over the duration of the entire experiment. The parallel coordinate planes plotting technique is an excellent visualization tool to answer these types of questions. By this method, experiments are ordered on the *x*-axis and expression values plotted on the *y*-axis. All genes in a given experiment are plotted at the same location on the *x*-axis; only their *y* location varies. Another experiment is plotted at another *x* location in the plane. Typically the progression of time

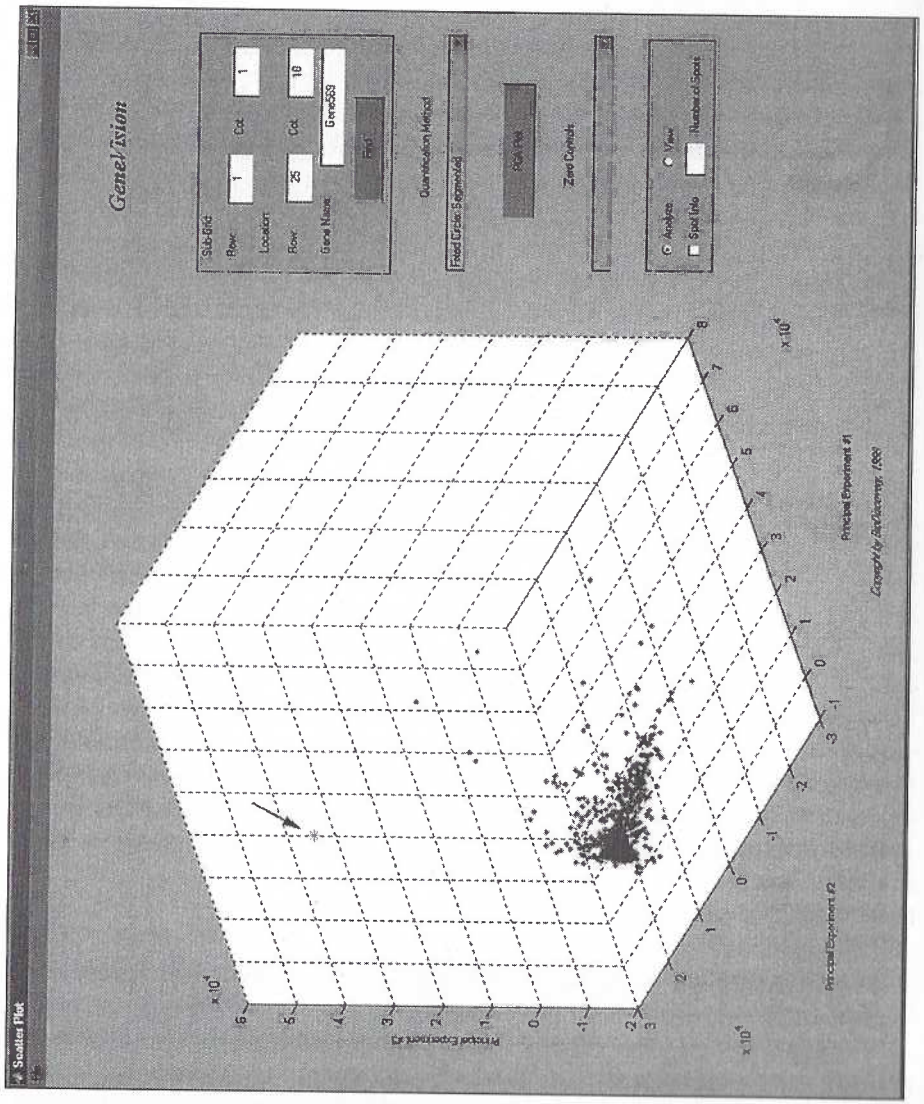

Figure 12. Principal component analysis on 600 genes across 21 experiments. Gene scores are plotted on the first three principal components. The arrow indicates a possible outlier. (See color plate A17.)

would be mapped into the *x*-axis by having higher *x* values for experiments done at a later time. By connecting the expression values for the same genes in the different experiments, one obtains an intuitive depiction of the progression of gene expression (Figure 13).

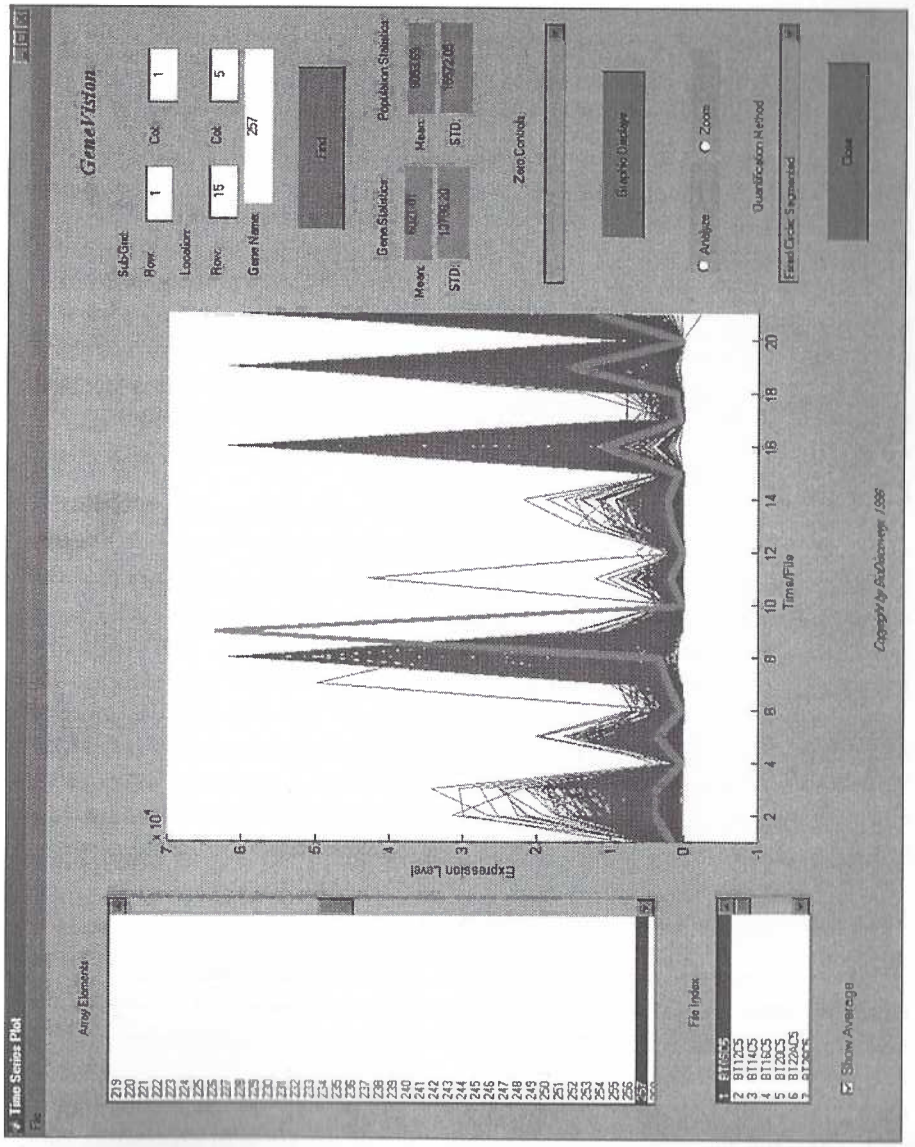

Figure 13. The parallel coordinate plot. The plot displays the expression levels of all genes across all experiments/files in the analysis. On the *x*-axis, the experiments or experimental files are ordered. The *y* axis reveals the expression level of all genes across all experiments. (See color plate A18.)

Cluster Analysis

Another common microarray issue involves finding groups of genes with similar expression profiles across a number of experiments. The most commonly used multivariate technique to find these groups is cluster analysis. Cluster analysis orders the data by grouping genes with similar expression patterns close to each other. Because genes are often expressed or induced only when they are required for a given process, cluster analysis can help establish functionally related groups of genes or predict the biochemical and physiological roles of previously uncharacterized genes.

The clustering method that is most frequently used in the microarray literature is hierarchical clustering. This method attempts to group genes and/or experiments in small clusters and then group these clusters into higher level clusters and so on. As a result of this clustering or grouping process, a tree of connectivity of observations emerges that can easily be visualized as dendrograms. For gene expression data, not only the grouping of genes, but also the grouping of experiments is important. When both are considered, it becomes easy to search for patterns simultaneously in gene expression profiles across many different experimental conditions, as shown in Figure 14. For example, a group of genes behaving similarly (e.g., all up-regulated) can be seen in a particular group of experiments (Figure 14).

Every colored block in the middle panel of Figure 14 represents the expression value of a gene in an experiment. The 600 genes are plotted horizontally, and the 21 experiments are plotted vertically. The color code is located in the lower right corner. The dendrogram for the genes is located just above the color-coded expression values, with one arm connected to every gene in the study. The dendrogram for experiments on the left shows the grouping of the 21 experiments in the study.

Although hierarchical clustering is a commonly employed approach presently, other nonhierarchical (k-means) methods are likely to gain popularity in the future with the rapidly growing volume of data and the expanding density of microarray assays. Nonhierarchical approaches provide sufficient clustering without having to create the full distance or similarity matrix or scan the whole dataset excessively.

Data Normalization and Transformation

There are several visualization and statistical techniques (described above) that could be useful for the analysis of microarray data. However, it is important to realize that even with the most powerful statistical methods, the success of the analysis is crucially dependent on the "cleanness" and statistical properties of the data. Essentially, the two questions to ask before starting any analysis are:

1. Does the variation in the data represent true variation in expression values or is it contaminated by experimental variability?
2. Are the data "well behaved" in terms of meeting the underlying assumptions of the statistical analysis techniques that are being applied?

It is easy to appreciate the importance of the first point. The significance of the second one derives from the fact that most multivariate analysis techniques are based on underlying assumptions such as normality and homoscedasticity. If these assumptions are not met, at least approximately, then the entire statistical analysis could be distorted, and statistical tests might be invalid. Fortunately, a variety of statistical techniques are available to help us answer "yes" to the above

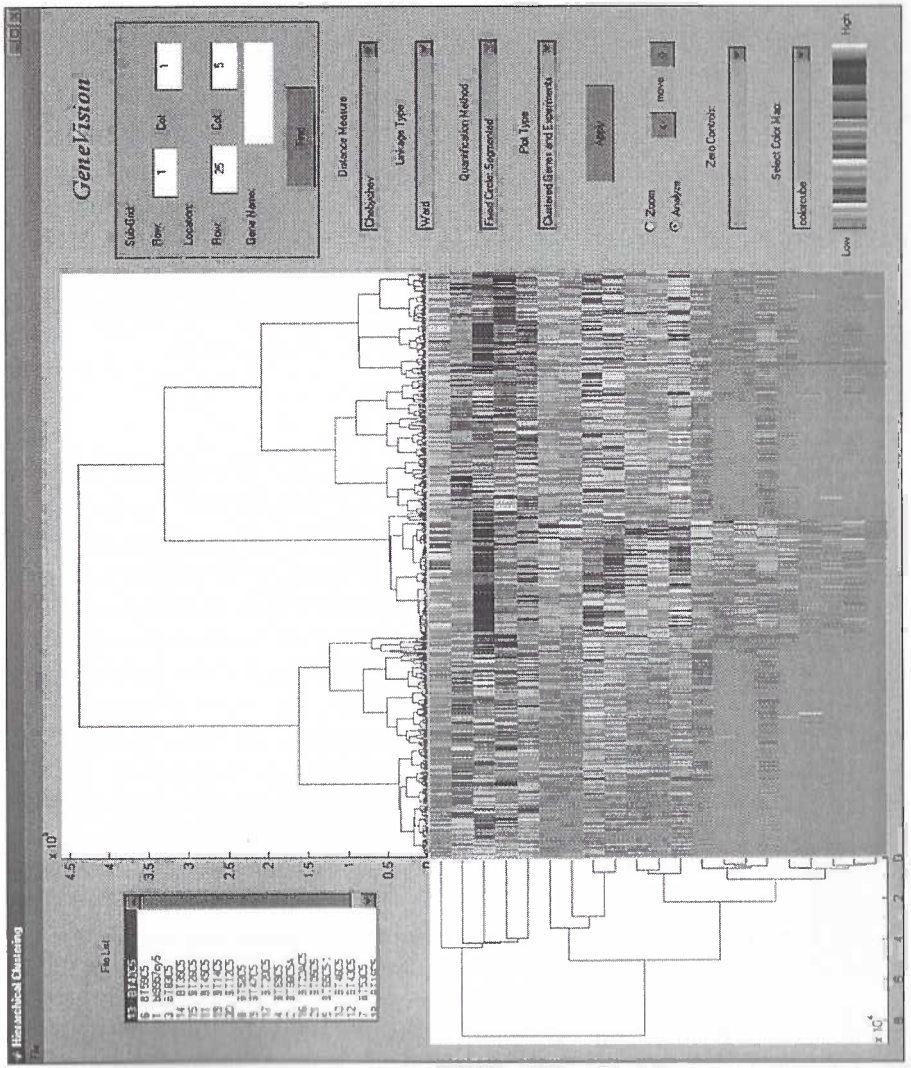

Figure 14. Color-coded gene expression values. Values are given for 600 genes (displayed horizontally) over 21 experiments (displayed vertically). Simultaneous clustering of genes and experiments is visualized by the top and left side dendrograms, respectively. (See color plate A19.)

questions. These are normalization (standardization) and transformation.

Normalization, and standardization, as a special form of normalization, can help us separate true variation in expression values from differences due to experimental variability. This step is necessary since it is quite possible that due to the complexity of manufacturing, hybridizing, scanning, and quantifying microarrays, variation originating from the experimental process may contaminate the data. During a typical microarray experiment, many different variables and parameters can affect the measured expression levels. Among these are slide quality, pin quality, amount of DNA spotted, accuracy of arraying device, dye characteristics, scanner quality, and quantification software characteristics, just to name a few. The various methods of normalization aim at removing or minimizing expression differences due to variability in these parameters.

As is discussed later, the various transformation methods all aim at changing the variance and distribution properties of the data in such a way so as to meet the underlying assumptions of the statistical techniques applied to it in the analysis phase. The most common requirements of statistical techniques are for the data to have homologous variance (homoscedasticity) and normal distribution (normality). Several popular ways of normalizing and transforming microarray data are discussed below.

Normalization of the Data

One of the most popular ways to control for spotted DNA quantity and surface chemistry anomalies involves the use of two-color fluorescence. For example, a Cy5 (red)-labeled probe prepared from healthy tissue could be used as control to examine expression profiles in a Cy3 (green)-labeled probe prepared from a tumor tissue. The normalized expression values for every gene would then be calculated as the ratio of experimental and control expression. This method can obviously eliminate much (but not all!) experimental variation by allowing two samples to be compared on the same chip. Even more sophisticated three-color experiments are under way in which one channel serves as a control for the amount of spotted DNA, and channels two and three allow two samples to be compared.

In addition to the local normalization method described above, global methods are also available in the form of "control" spots on the slide. With a set of control spots, it is possible to control for global variation in overall slide quality or scanning differences.

The above procedures describe some array-based measures one can use to normalize microarray data. However, even with multiple color fluorescence and control spots, undesired experimental variation can contaminate expression data. It is also possible that all or some of these physical normalization techniques are missing from the experiment, in which case it is even more important to find additional ways of normalization. Fortunately, additional statistically based normalization methods are available to add additional precision to the data.

For example, expression values for a given set of genes in one experiment may be consistently and significantly different from another experiment due to quality differences among the slides, printing, scanning, or some other factor (Figure 15). It might be very misleading to compare the expression values of the two files plotted in Figure 15 without normalization. The same data plotted after normal-

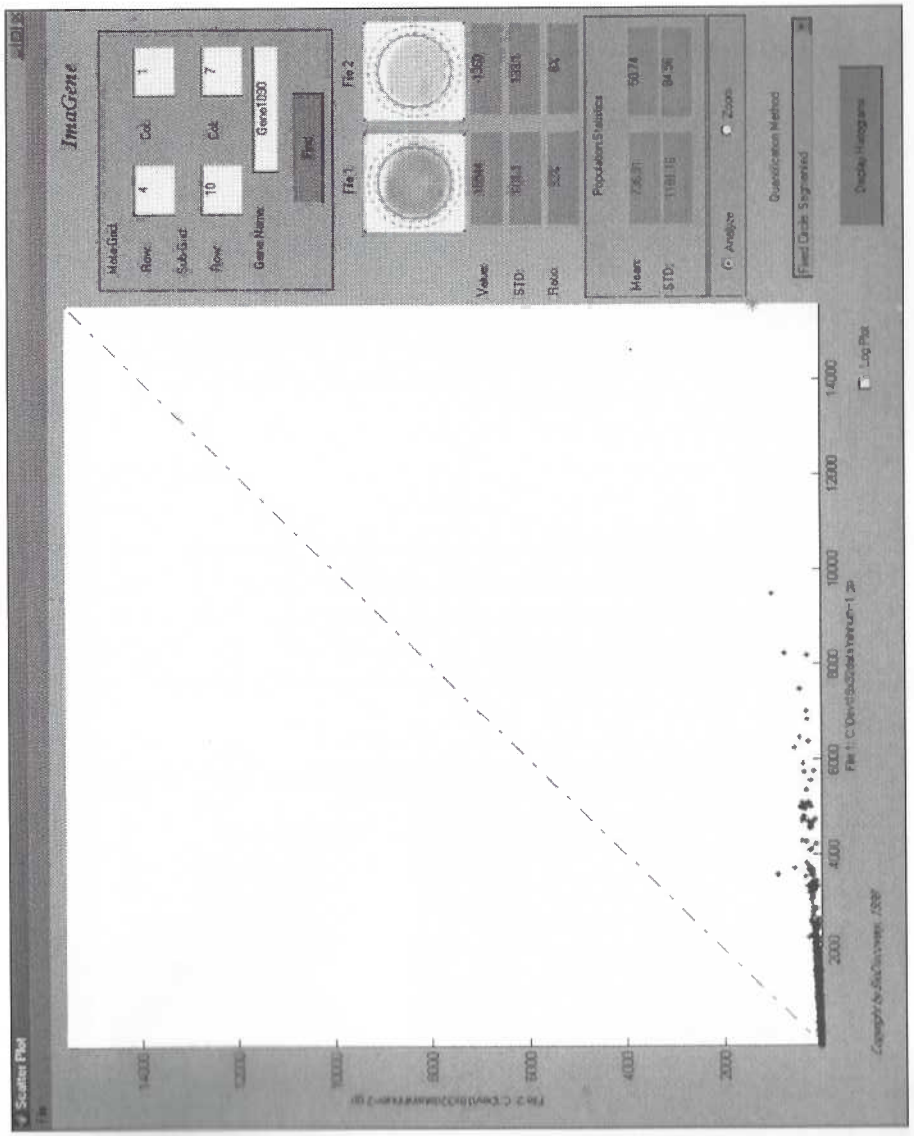

Figure 15. Comparison of fluorescent signals. Scatter plot of two experiments in which the overall fluorescent signal for one file is significantly stronger than the other. (See color plate A20.)

ization, however, reveal that most of the genes fall on the identity line, as would be expected for two experiments with the same set of genes (Figure 16).

Transformation of the Data

Although there are many different ways to transform data, the most frequently used procedure in the microarray literature is to take the logarithm of the

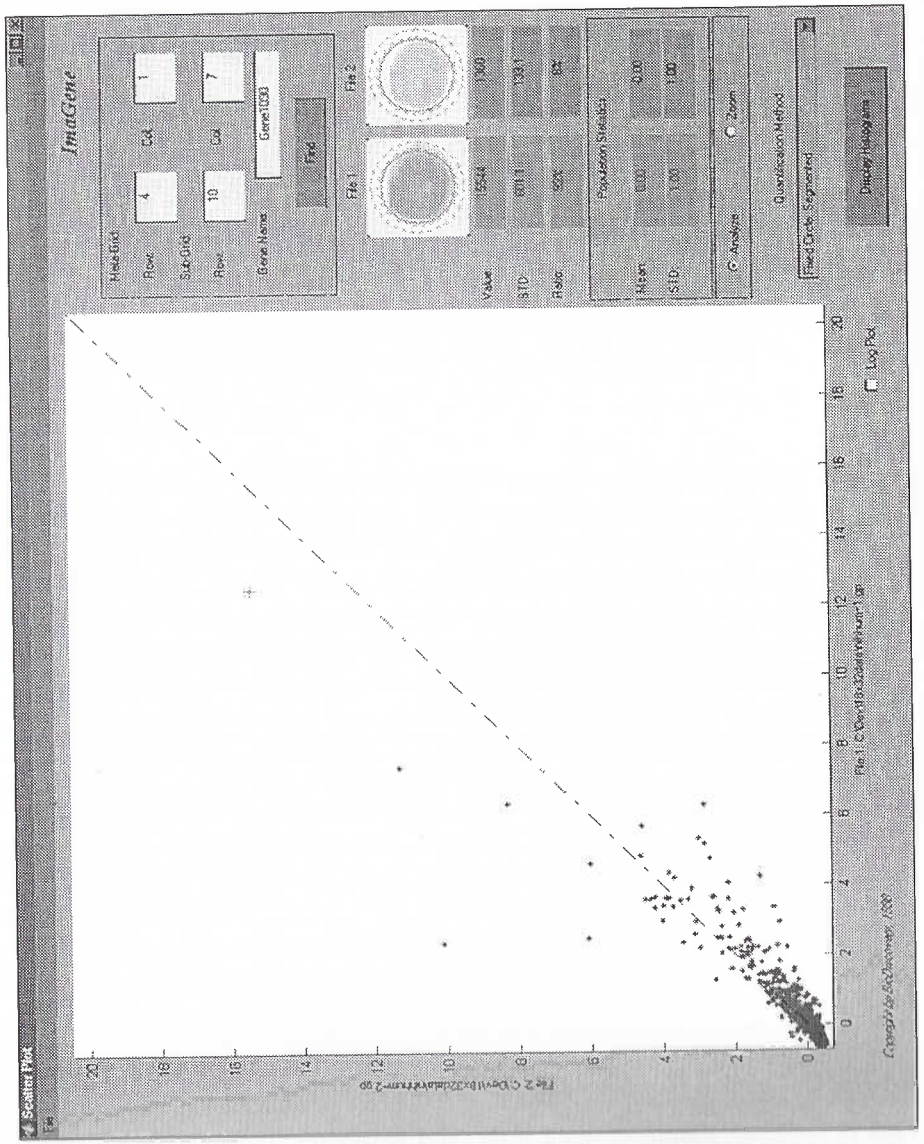

Figure 16. The same data as in Figure 15, but after normalization. (See color plate A21.)

quantified expression values (2,3). An often cited reason for applying such a transform is to equalize variability in the wildly variable raw expression scores. If the expression values were calculated as a ratio of experimental over control, then an additional effect of the log transform would be to equate up- and down-regulation by the same amount in absolute value scores ($\log_{10} 2 = 0.3$ and $\log_{10} 0.5 = -0.3$). Another important consequence of the log transform is bringing the distribution of the data closer to normal. Because the log transform has a good chance of meeting the normality and homoscedasticity assumptions, the use of a variety of parametric statistical analysis methods is also better justified.

As shown in Figure 17 (left panel), without a log transform the data distribution appears as a sharp peak with a very long tail. This distribution is far from normal, which violates the assumptions made by many standard parametric statistical analysis methods. The distribution of the log-transformed data (right panel) is clearly much closer to that of the normal distribution, which could also be verified by a simple normality test. Certainly, other types of transforms could also be explored. In fact, the choice of transformation should be dependent on which one brings the data closest to the requirements of homogeneity of variance and normal distribution (4). For most microarray expression data, the log transform typically provides the best solution.

CONCLUSIONS

Microarray technology has become a powerful tool for genetic research. From an information processing perspective, microarray technology aids the researcher in transforming and supplementing data available on genes and cells into useful information about gene expression, and ultimately, a broader biological understanding. In this chapter, many issues involved in data management and track-

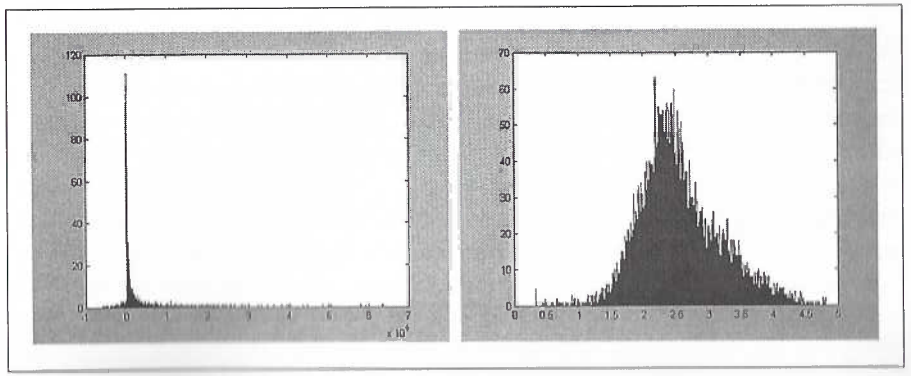

Figure17. Histograms of the previously cited gene expression data. Six hundred genes × 21 experiments. The *x*-axis indicates the expression values, and the *y* axis shows the number of genes with a particular gene expression level. The left and right panels show the data before and after the log transform, respectively.

ing, array production, analysis of high-density array images, and manipulation of massive data sets are described. Some methods and tools available for addressing many of these issues are also presented. Many challenges lie ahead as software tools for microarray analysis become increasingly powerful.

REFERENCES

1. **Chen, Y., E.R. Dougherty, and M.L. Bittner.** 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics *2*:364-374.
2. **Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*:14863-14868.
3. **Hilsenbeck, S.G., W.E. Friedrichs, R. Schiff, P. O'Connell, R.K. Hansen, C.K. Osborne, and S.A.W. Fuqua.** 1999. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. J. Natl. Cancer Inst. *91*:453-459.
4. **Johnson, R.A. and D.W. Wichern.** 1998. Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River.

9 Microarray Tools, Kits, Reagents, and Services

Todd Martinsky[1] and Paul Haje[2]
[1]TeleChem International, Inc. and [2]arrayit.com, Sunnyvale, CA, USA

INTRODUCTION

This chapter describes some of the tools, kits, reagents and services we have developed at TeleChem and arrayit.com (http://arrayit.com), and the science that underlies these products and services. TeleChem's products for the microarray industry are sold under the ArrayIt™ brand name. ArrayIt products are mainly consumables to help scientists make microarrays of high quality and obtain the finest scientific results with their biological "chips." Our microarray products currently fall into three categories: DNA purification systems, microarray manufacturing technology, and surface and hybridization chemistry. Additional custom microarray services are provided by arrayit.com, a new internet-based microarray company offering Flex-Chips™ and eChips™, as well as DiscoverChips™, which contain segments of genes from a diverse set of organisms including yeast, *Drosophila, Caenorhabditis elegans*, mouse, rat and human.

Microarray Experimental Cycle

Microarray analysis is revolutionizing biological research (5,9,10,18–21,23, 26, 29,30,34,39). Each microarray experiment contains five separate steps: biological question, sample preparation, microarray reaction, detection, and data analysis and modeling (33). This methodological life cycle provides the foundation for the microarray industry. Current products and services from TeleChem and arrayit.com fall into four of the five phases of the life cycle: biological question, sample preparation, microarray reaction, and microarray detection phases (Figure 1). These products and services and the underlying science are described below.

Microarray Biochip Technology
Edited by Mark Schena