MICROARRAY Technologies

Short Technical Report

ArrayExplorer©, a Program in Visual Basic for Robust and Accurate Filter cDNA Array Analysis

BioTechniques 31:862-872 (October 2001)

P.C. Patriotis[1,2], T.D. Querec[1], B.N. Gruver[1], T.R. Brown[3], and C. Patriotis[1]

[1]Fox Chase Cancer Center, Philadelphia, PA, [2]Penn State University, University Park, PA, and [3]Columbia University, NY, USA

ABSTRACT

Determining the dynamics in the global regulation of gene expression holds the promise of bringing a better understanding of the processes that govern physiological cell growth regulation and its disruption during the development of disease. The advent of cDNA arrays has created the possibility for the parallel analysis of expression of thousands of genes in a given cell population, simultaneously. The level of expression of a given set of genes within the studied tissue corresponds to the intensity of a labeled cDNA probe synthesized from the studied tissue RNA and bound specifically to the cDNAs of the genes spotted on the array. The accurate extraction of gene expression intensity values is essential for further data analysis and the interpretation of the obtained results. Here, we describe a new array image-processing software developed in Microsoft® Visual Basic, the ArrayExplorer©, which provides a user-friendly, multiple-window interface and a number of automatic and manual features that facilitate a reliable, robust, and accurate extraction of gene intensity values from filter-array images.

INTRODUCTION

The cDNA microarrays are an emerging new technology developing on the basis of recent technological achievements and the sequence information generated by the Human Genome Project. cDNA microarrays allow the parallel analysis of the expression of hundreds to thousands of genes in different cell populations, simultaneously (1–6). Gene expression arrays contain unique cDNA sequences corresponding to the genes of interest, which are spotted on the array substrate in a known configuration. Depending on the substrate used, cDNA arrays may be glass or filter arrays. Arrays may be purchased pre-made from a number of commercial suppliers (e.g., Affymetrix, Clontech Laboratories, Research Genetics, Genome Systems) or may be individually spotted by in-house facilities with custom configurations. In the case of filter macroarrays, equivalent amounts (about 5 ng) of PCR-amplified DNA sequences (0.5–2.0 kb) of specific cDNA clones are immobilized at a density of 10–20 cDNAs/cm^2 on positively charged nylon membranes (3,4).

Complex cDNA probes are derived in vitro by reverse transcription of purified total or poly(A)$^+$ RNAs and an incorporated radioactive label (^{32}P or ^{33}P) and are then hybridized to the DNAs immobilized on the arrays. The detection of the hybridized DNA probe is usually carried out by means of phosphoimage scanning or direct autoradiography. Once captured, the high-resolution digital images are subjected to processing by densitometric analysis. The accurate measurement of the gene intensities from the array-generated analog images is quite challenging. One of the issues is the proper placement of the reference grid over the array image relative to the location and size of the hybridization signals. Another issue is the extraction of the hybridization intensities, particularly when they are overlapping and in the presence of nonspecific, background noise. A final but very important point is to present the results in a way that facilitates their further analysis and easy comprehension. However, most of the commercially available array image processing and analysis programs address some but not all of these issues, and their resolution contributes to a significant improvement of the degree of accuracy of the extracted gene intensity values, including a more linear detection of signals within a wide dynamic range, and the detection of a greater number of genes on each array.

ArrayExplorer© contains routines for flexible grid placement. The variable diameter of the hybridization signals, a common characteristic of array images obtained with radiolabeled probes, is incorporated in the gene intensity calculation. The problem of high spot-to-spot signal interference or overlapping hybridization spots is also solved by the ArrayExplorer. Automatic and manual routines facilitate the easy separation of such signals from each other and allow the accurate estimation of their intensities. Another feature of ArrayExplorer is flagging of signals that are solely due to higher local background noise or when an automatic routine fails to converge. The combination of such novel automatic and manual features allows for the robust and accurate extraction of gene in-

tensity values, which are finally exported automatically in a text/spreadsheet format, where they are conveniently linked to information regarding array location, gene identity, and access to DNA and protein sequence databases.

MATERIALS AND METHODS

Program Design

ArrayExplorer is coded in the Microsoft® Visual Basic programming language. The interface is developed and compiled using Microsoft Visual Basic 6.0 (Service Pack 3) Professional Edition for 32-bit Windows™ applications. The code is optimized for execution speed and structured into object-classes for handling the storage, management, and analysis of gene information. These classes are subdivided into "whole grid", "individual grid region", "individual gene", and "individual spot". Each class contains flags and locations of the regions that it affects. The presentation of the data is reflected by a variety of flagsets that give the hybridization signals different colors on the array image.

Many features set ArrayExplorer apart from other programs dealing with array image processing. The program uses circular scanning regions (CSRs), which may vary in location and size. The intensity of each array spot is calculated by integrating the pixel intensities within each CSR, which are corrected for the local background. For 8-bit images, the intensity value in each pixel may vary within a range of 0–255. The background is determined globally for the entire gridded array image or for each individual grid section. In either case, the background value is determined as the average intensity per pixel within regions that do not contain any spotted cDNAs and, therefore, with no hybridization signals. A number of regions have been incorporated as a default option in the program for background determination; however, the user may manually customize the size and location of these regions. Another important feature of ArrayExplorer is the ability to subdivide the CSR into selectable and sumable 1/8th-sectors, or "pie-sectors". This procedure is used at instances when signal bleeding from adjacent hybridization spots causes signal overlapping or when there is a distortion in the signal caused by hybridization artifacts. The user selects the pie-sector(s) that are free of signal interference at a given spot. The intensity of the selected pie-sector(s) is determined automatically and, since the hybridization spots are for the most part symmetrical and circular, the program calculates the intensity of the entire spot, assuming that it is proportional to the intensity determined within the selected pie-sector(s).

The most original feature of ArrayExplorer is its ability to locate and expand the CSRs to encompass the entire hybridization signal at each spot. Following grid alignment, a gradient method is used to locate the center of the hybridization signal within each square and align a CSR: two sums are taken by calculating the average intensity in the

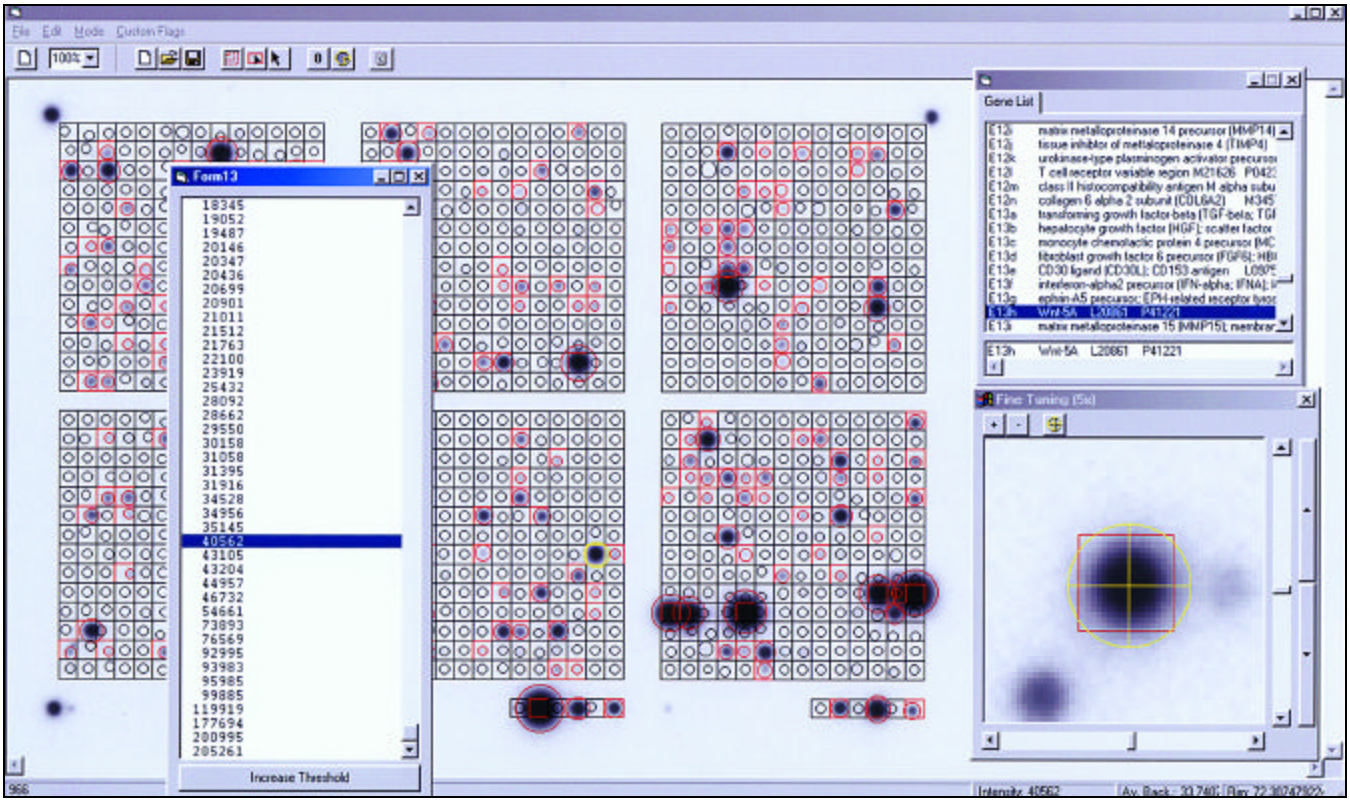

Figure 1. ArrayExplorer screen display. ArrayExplorer has four linked operational windows. Each window can be resized and repositioned on the screen. The main window displays the entire array image with the overlaid reference grid. Red boxes indicate signals above the background threshold; black boxes are below this threshold. The *Fine Tuning*, *Gene List*, and *Threshold* windows allow easy navigation and manual adjustment of the grid. Operations are facilitated through the menu items, the toolbar buttons, and the designated shortcut keys.

rim of the CSR, along the X and Y axes. As these sums are weighed trigonometrically, the program changes the coordinates of the CSR in the direction(s) that results in a greater sum. The difference of the sums between the top and bottom and the left and right sides of the CSR rim is calculated. The magnitude of the calculated sums in the CSR rim will instruct the program toward which direction and how much to move the CSR in each axis, for the difference in the sums between the left and right and the top and bottom of the rim to be close to zero. Once CSR rim intensity symmetry has been achieved, the CSR is expanded by default until the average intensity value in the rim reaches the sum of the average background value plus three standard deviations of the background. Additionally, the user may expand or reduce the size of the CSR using the operational window for manual fine-tuning. An array spot is automatically flagged when the automatic routines fail to detect the spot center or the CSR is expanded beyond the limits of the grid square.

ArrayExplorer exports the extracted intensity values in a text file, which also contains other relevant information for each gene included in the array. This information includes: (*i*) the gene's array location; (*ii*) where available, the gene's name, family, and functional classification; (*iii*) database access information, including GenBank® and SwissProt access numbers; (*iv*) the gene intensity value; and (*v*) custom or automatic flags. Each of these pieces of information is tab-delimited, and each of the genes is separated by carriage-return line-feed. This format allows for easy import into a custom program or a professional spreadsheet/data-analysis program such as Microsoft Excel®.

Simulated Array Images for Program Testing

Simulated array images were created with a custom program in Visual Basic that was designed to generate a bitmap image of a 1200-gene array. To reproduce the variability in the actual hybridization signals, the size, position within the grid square, and spot intensities were generated at random within the ranges of the observed values from actual arrays. The low-intensity hybridization spots were simulated by generating spots with a fixed diameter and variable height; the high-intensity spots were simulated by varying the diameter of the spot while keeping a constant, maximum height. Some of the high-intensity spots overlapped to varying degrees. The borders of the simulated spots were gradually blending into the background, much like the true hybridization signals. Furthermore, the simulated array gene signals were randomly offset from their center to test the accuracy of ArrayExplorer for CSR autoplacement. Finally, a global background and random noise was added to the entire image.

Hardware and System Requirements and Program Availability

ArrayExplorer requires an Intel® Pentium® II or later CPU (200 MHz or greater) with at least 64 MB RAM and 10 MB available disk space for installation. The program is compatible with Microsoft Windows 95, 98, NT4.0, or Windows ME, and it may be obtained upon request.

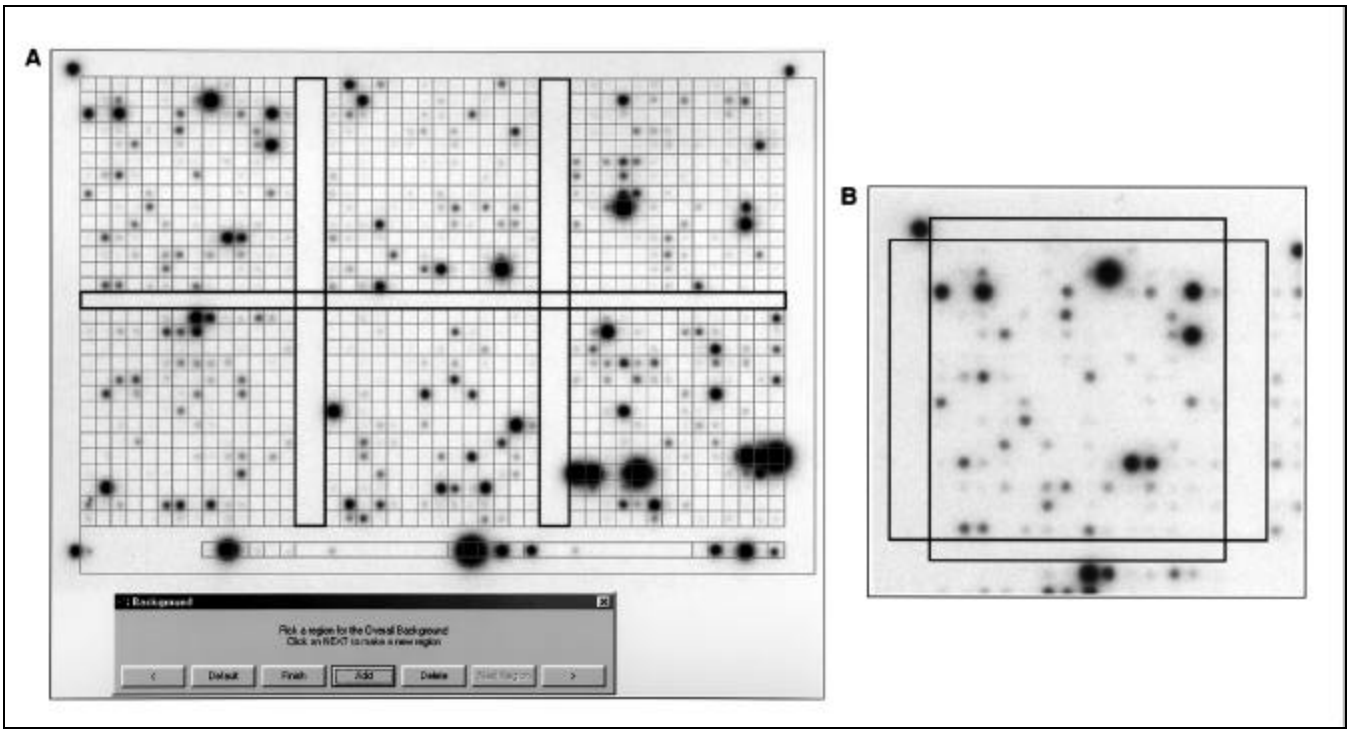

Figure 2. Background determination. Background can be calculated with ArrayExplorer in two ways: (A) The default overall background is calculated as the average intensity value per pixel in the central array areas, boxed in black, which are devoid of spotted cDNA/hybridization signals. (B) The default sectional background is calculated as above within signal-free regions along the borders of each array section. In both cases, the user can modify the size and location of the default background calculation regions.

RESULTS AND DISCUSSION

General Program Layout

ArrayExplorer uses a multi-window interface to analyze array images in bitmap format (Figure 1). The main window displays the array image and its corresponding reference grid. Drop-down menu options *File*, *Edit*, *Mode*, and *Help* are displayed at the top of the main window. The menu options allow the user to initiate all available commands operating in the main window, as well as for opening other available operational windows. A selection of buttons and shortcut keys are available for commonly used commands, which increases the functionality of the program. Finally, the background and net intensity values for each selected grid location are displayed at the bottom right of the main window.

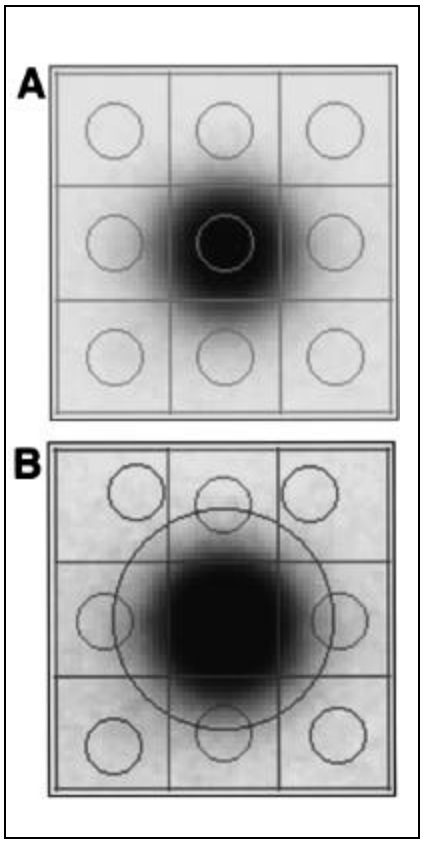

Figure 3. Automatic CSR expansion. The ArrayExplorer *AutoPlacement* option allows automatic, differential expansion of the grid CSRs to encompass hybridization signals with variable diameters. (A) Before CSR expansion, only a portion of the total gene hybridization signal is registered. (B) Automatic expansion of the CSR allows intensity calculation over the entire hybridization signal. Grid boxes flagged in gray represent areas that are manually zeroed by the user to avoid registration of false-positive signals caused by signal bleeding from adjacent boxes.

Four additional operational windows increase the flexibility of ArrayExplorer. The *Gene List* displays a searchable list of the genes spotted on the array and their grid locations. The *Fine Tuning* window allows the user to adjust manually the size and coordinates of the CSR around individual gene spots. The *Threshold* window provides a list of the gene intensities sorted in ascending order. This window allows the user to zero all points that have intensities below a given threshold value, which is caused by background noise. These operational windows are linked together such that selecting a gene at a grid location in the main window will display a magnified image of it in the *Fine Tuning* window, list its location number and name in the *Gene List*, and show its intensity in the *Threshold* list of intensities. A fourth window allows the user to look at the 2-D profile of the intensity values for a manually selected set of genes.

Input

ArrayExplorer uses bitmap-formatted images generated with a phosphorimager-linked application, Adobe® PhotoShop®, or other graphics-editing programs. Phosphorimager scanning at variable exposure times is carried out at 16-bit/50-µm resolution. Alternatively, autoradiographic images are obtained on BioMax® MS films (Eastman Kodak, Rochester, NY, USA) for variable exposure times and scanned at 16 bit/1200 dpi (25 µm) grayscale resolution using a flatbed scanner. The resolution of the obtained images is then converted to 8 bit/200 dpi using a standard graphics-editing program. Processing of higher-resolution images (400 or 600 dpi) with ArrayExplorer indicated that an increase in image resolution does not yield a higher degree of accuracy (data not shown). For a more precise grid alignment, before processing with ArrayExplorer, array images may be subjected to rotation and cropping using standard graphics-editing programs.

Features

Grid selection and placement. ArrayExplorer allows users to select from a list of standard reference grids for commercially available filter arrays. These grids are linked to corresponding files that contain gene identity and array location, as well as database access information (GenBank and SwissProt). Currently, the grids available with ArrayExplorer include all human, murine, and rat Atlas™ array series from Clontech Laboratories (Palo Alto, CA, USA). However, the versatility of ArrayExplorer allows users to design and save grids for arrays with custom configurations and size. Reference grids that correspond to new arrays can be easily incorporated in the program.

Once an appropriate configuration for the analyzed array is selected, the grid is aligned over the array image. This feature is quite flexible and allows the user to fine-tune the alignment of the grid over the array image. To address the inherent variability of the spotting process, ArrayExplorer contains three routines that facilitate rapid and accurate positioning of the grid: (*i*) the user positions the grid based on the known location of two genes in the grid; (*ii*) the grid can be stretched and dragged over the image with the mouse and cursor arrows; and (*iii*) individual grid sections can be selected and realigned, which allows for potential variability in the spacing between individual array sectors, as well as stretching of the filters after repeated hybridization. Since each gene intensity is calculated within a CSR that is automatically aligned with the hybridization signal and further fine-tuned if necessary, the required level of precision for grid alignment is such that each gene spot grossly falls within a grid square.

Background estimation. ArrayExplorer has two different options for calculating the background around a particular gene. When the general background over the entire array image appears uniform, a global average background value can be determined automatically from areas within the borders of the grid. More often, however, the background varies between the different sections of the array image. In such cases, a sectional average back-

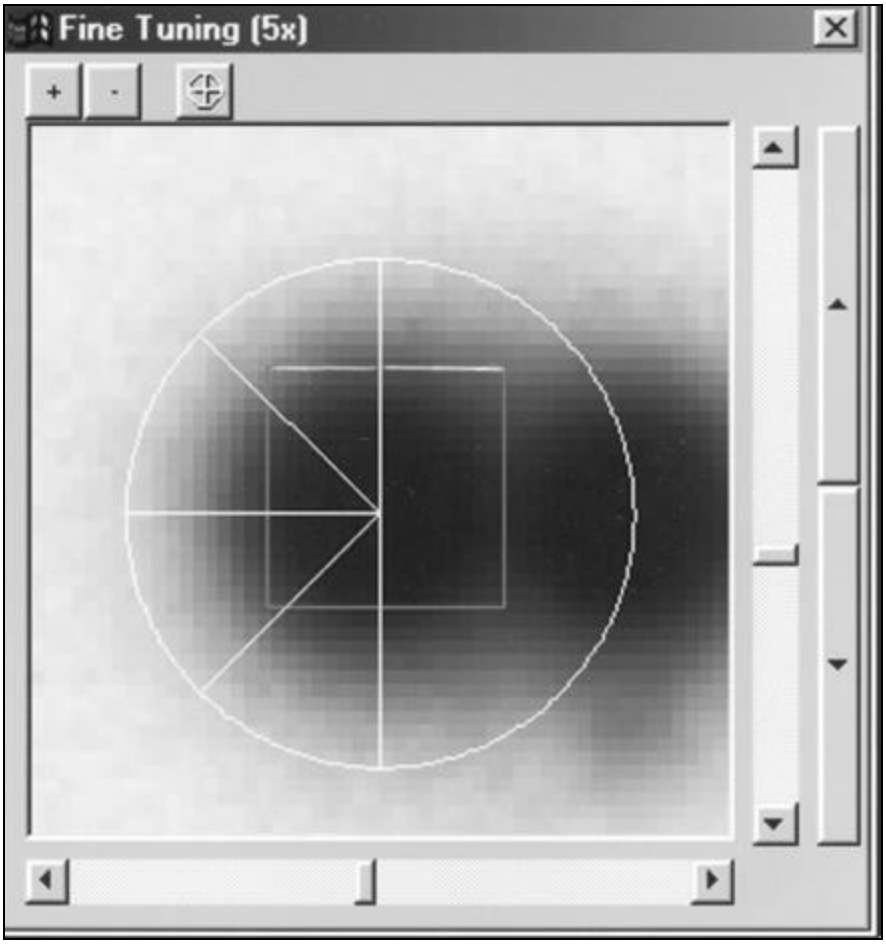

Figure 4. Fine Tuning and Pie-sectoring options. The Fine Tuning window displays an image centered on the selected gene spot. In this operational window, the user is able to manually adjust the diameter and positioning of the CSR. The Pie-sectoring option is used to determine the gene intensity value extrapolated from the selected portion (here from the left half) of the hybridization signal, which is free of signal interference from adjacent spots. The user may decide how many consecutive CSR sectors should be included for calculation of the total signal.

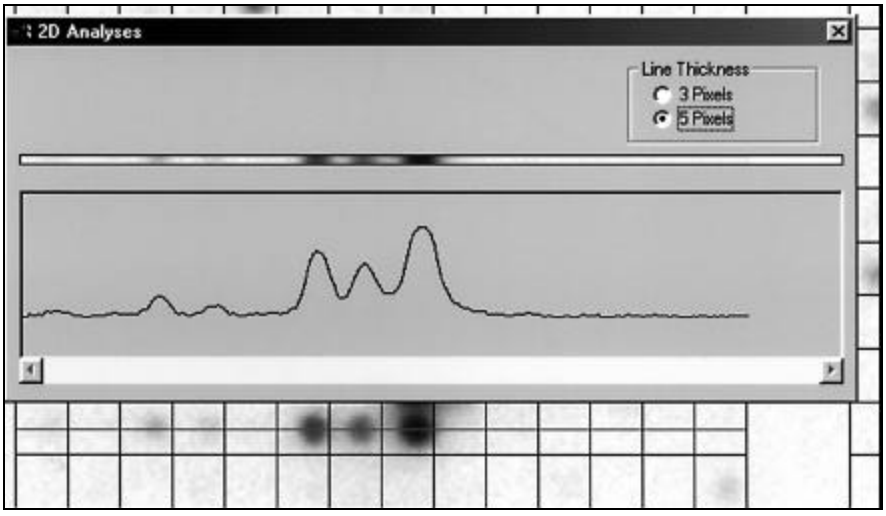

Figure 5. 2-D profile analysis. The user may choose sections of the gridded array image to obtain a 2-D display of the hybridization intensities. This may prove helpful during manual manipulation and fine-tuning of the grid, and for data processing.

ground value can be calculated and then subtracted from the intensities of the genes included in each corresponding array section. In both cases, users can either choose a default background setting (Figure 2) or create a custom pattern of areas within the array image that are devoid of hybridization signals, which can then be used to calculate the average background value.

Intensity determination. Accurate extraction of the gene intensities is quite challenging because of irregularities in DNA spotting, signal bleeding beyond the coordinates of the grid square, and interference with adjacent gene signals.

ArrayExplorer addresses these problems through a combination of innovative features for flexible CSR placement and size. Once the CSRs are placed around each hybridization signal, their radius is increased until the circle encompasses the entire signal. This allows extraction of the intensity values from hybridization signals with variable levels of bleeding beyond the circumference of the spotted cDNAs (Figure 3). Gene spots whose hybridization signals cannot be determined automatically by ArrayExplorer because of signal interference and overlapping between adjacent gene spots are flagged as green boxes on the grid. Such spots can be easily selected and their CSR positioning can be adjusted using the Fine Tuning window (Figures 1 and 4). Two methods in ArrayExplorer allow the user to deal with intense hybridization signals that leak into adjacent grid squares. In simple cases, the program identifies and flags adjacent squares that do not contain signal from their own spotted gene cDNA. The user can manually select each of these boxes and annul the corresponding gene intensity. The result is color-coded in the output data file. A more complex situation occurs when one hybridization signal partially overlaps with another. To solve this problem, the *Pie-sectoring* option has been incorporated in ArrayExplorer, which allows the user to divide each of the overlapping hybridization signals into 1/8-slices and use only those slices that are free of signal interference (Figure 4). *Pie Auto* then readjusts the radius of the corre-

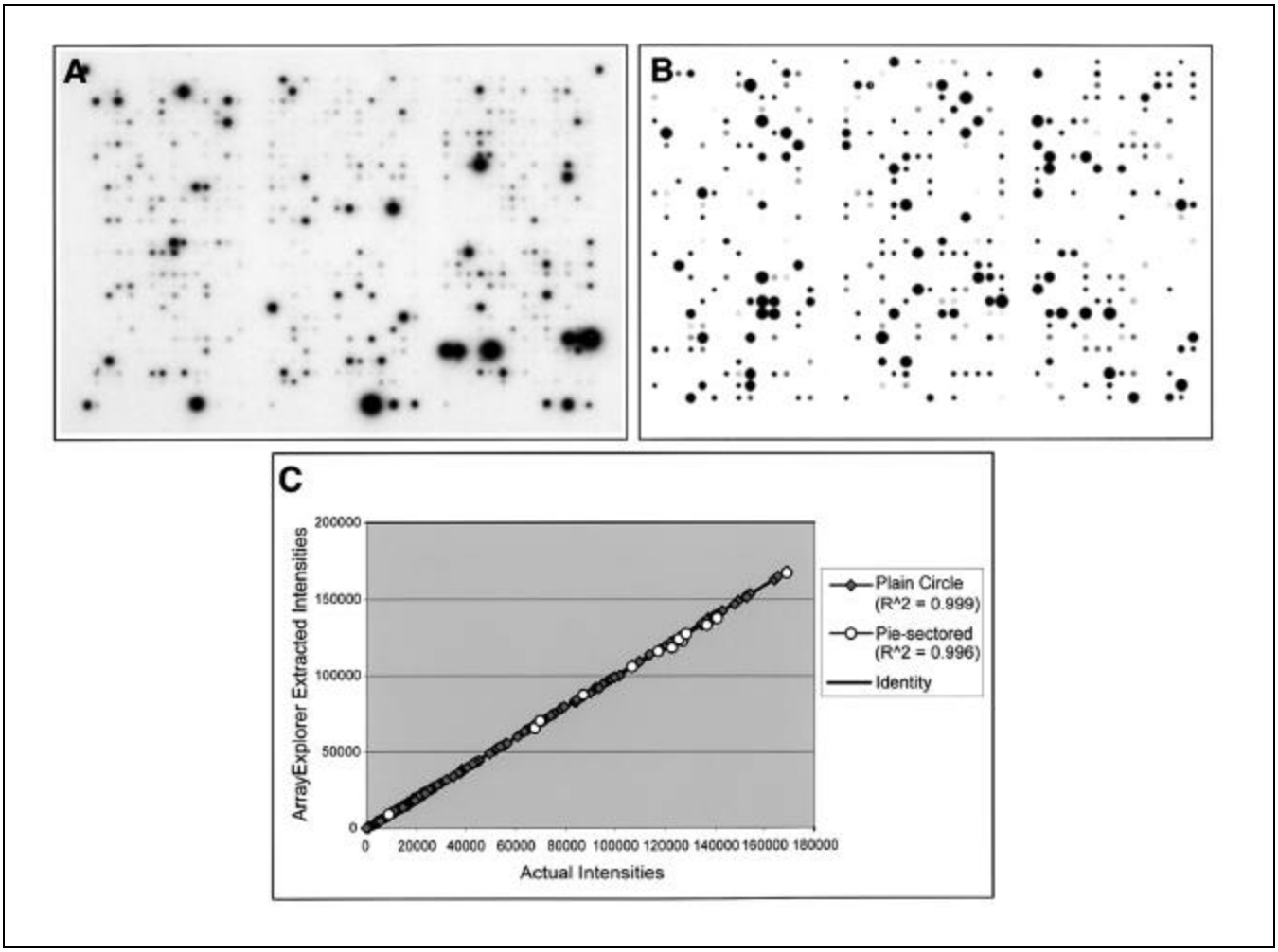

Figure 6. Accuracy of data extraction with ArrayExplorer. (A) Actual autoradiographic image of a hybridized Atlas 1.2 Human Cancer array. (B) Simulated array image: spots were generated randomly using a custom algorithm developed in the laboratory. The range of diameters and intensities found in the actual array images is duplicated in the simulated array. Plain spots in the simulated array were analyzed using a combination of automatic and manual CSR adjustment, while overlapping spots were analyzed using the Pie-sectoring option. (C) Correlation plot of the actual input intensity values versus the values extracted with ArrayExplorer from the simulated array image.

sponding CSR on the basis of the calculated average intensity value per pixel in the selected set of 1/8-slices.

Four other options in ArrayExplorer increase its utility for image processing and further data analysis: The *Threshold* feature allows the user to determine a cutoff value for true hybridization signals and noise-related signals. With this option, all detected hybridization signals are automatically sorted in a descending order, and the threshold signal value is being determined through direct inspection of grid squares linked to corresponding values in the *Threshold* list. All signals below the threshold value are then automatically zeroed and specifically flagged in the output data file. Another option allows the user to select multiple grid squares that require the same operation so that they are manipulated simultaneously. The user also has the option to custom design up to five flags for labeling selected gene signals on the processed array image. Each flag is color-coded on the grid and has a designated label that appears adjacent to the corresponding gene intensity value in the output data file. These flags can help the user to easily locate selected groups of genes during the later stages of data processing. Finally, ArrayExplorer has the ability to display a 2-D profile of manually selected array areas (Figure 5). This option is particularly important for determining whether a detected intensity value is indeed due to a true hybridization signal or if a particular signal has reached its saturation point, therefore requiring processing from images acquired at shorter exposure times.

Output

In addition to the gene intensity values extracted with ArrayExplorer, the generated output data file contains important relevant information regarding the location, identity, and database access entries for each gene and can be opened by a number of spreadsheet applications, including Microsoft Excel.

Simulated Images

The simulated array image (Figure 6B) was processed with ArrayExplorer similarly to actual array images (Figure 6A) obtained through autoradiography and scanning or with a phosphorimager. The results show that the intensity values extracted from the simulated array image with ArrayExplorer using standard automatic and manually adjusted CSRs are highly correlated to the actual input numbers, with an r^2 = 0.999 (Figure 6C). The flagged overlapping spots were processed using the *Pie-sectoring* option, and, again, the estimated values were in excellent agreement with the actual (r^2 = 0.996).

CONCLUSION

ArrayExplorer is an innovative program for the analysis of gene expression arrays obtained with radioactively labeled cDNA probes. The combination of automatic and manual features in ArrayExplorer yields highly accurate and robust array data extraction. Gene expression intensities are automatically exported in a spreadsheet/text file format, which is convenient for further analysis and data mining.

ACKNOWLEDGMENTS

The authors would like to thank Ms. R. Stoyanova for numerous suggestions for improving and testing the program, as well as Dr. E. Ross for critical review of the manuscript. The development of this software was supported by funds from the 5th District AHEPA Cancer Res. Foundation, Inc. awarded to C. Patriotis, and through NIH SPORE P50-CA83638 (PI: R. Ozols). C. Patriotis is a Liz Tilberis Scholar of the OCRF, Inc. P. Patriotis was a Fellow of the Howard Hughes Medical Institute High School Student Program.

REFERENCES

1. **DeRisi, J., L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, and J.M. Trent.** 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. *14*:457-460.
2. **Heller, R.A., M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D.E. Woolley, and R.W. Davis.** 1997. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc. Natl. Acad. Sci. USA *94*:2150-2155.
3. **Ramsay, G.** 1998. DNA chips: state-of-the art. Nat. Biotechnol. *16*:40-44.
4. **Schena, M., R.A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier, and R.W. Davis.** 1998. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol. *16*:301-306.
5. **Schena, M., D. Shalon, R.W. Davis, and P.O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*:467-470.
6. **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.

Received 7 May 2001; accepted 13 August 2001.

Address correspondence to:

Dr. Christos Patriotis
Division of Medical Science
Fox Chase Cancer Center
Philadelphia, PA 19111, USA
e-mail: c_patriotis@fccc.edu