Software for Quantitation and Visualization of Expression Array Data

BioTechniques 29:118-121 (July 2000)

David Strehlow

Boston University Medical Center,
Boston, MA, USA

ABSTRACT

Software is described that facilitates the analysis of phosphoimages from large array hybridizations. The Macintosh® PowerPC®-compatible application and its manual are available at no charge from <http://people.bu.edu/strehlow>. The software is compatible with both custom formats and array filters from three commercial manufacturers. It allows the rapid quantitation of every spot on images of hybridizations to large arrays. The user drags grids of squares over the spots on the image to define the coordinates of each spot, then aligns and edits the position of the grid. The software then corrects the positions as necessary and quantitates up to 27 000 spots per image. It stores the numerical values for each signal in a format called the fingerprint file. Fingerprint files can be directly averaged or compared, allowing the user to find mean values or differences in data from independent hybridization experiments. Data can be recalled from the fingerprint file and can be output in a variety of spreadsheet formats with several options for background correction. Finally, the software offers an output format that allows the convenient visualization of data points using animated, three-dimensional graphs.

INTRODUCTION

Large filter membranes containing thousands of expressed sequence tags or known cDNAs have recently been made available from several commercial vendors. Hybridization to these filters allows the determination of gene expression patterns. However, the full use of this information requires the ability to catalog numerical values for the expression of each clone on the filter. With proper computer software, it is possible to rapidly compare the expression levels of tens of thousands of genes in a given tissue. Also, it should be possible to compare expression levels of thousands of genes at many

stages during developmental programs, during oncogenesis and during progression of other disease states.

Commercial software is generally expensive, does not facilitate comparisons of whole expression patterns using pixel data and is limited in the filter types it will analyze. Many investigators perform only a small number of hybridizations and so find it difficult to justify the purchase of expensive software that might be used only once or a few times. Other free or less expensive software requires that the user compare images by superimposing one image on top of another and subtracting the pixel values. This superposition method requires that the images be identical trapezoids and that probe-specific activity be the same in all experiments. This is rarely the case, however.

A better method for comparing data from different hybridizations is not to compare images directly, but rather to quantitate each signal and compare the relative mathematical values for each clone on a pairwise basis. Since array filters typically contain thousands of data points, such software should accommodate large arrays of numerical values. DataMachine handles these large arrays and stores the clone pixel values with corresponding position data in a format called fingerprint files.

DataMachine Software

DataMachine allows the user to (a) open 16-bit phosphoimage data files; (b) quantitate total numerical pixel values representing any individual clone; (c) export total numerical pixel values of a user-determined number of clones with a variety of background correction methods; (d) find average and difference data from independent hybridizations; (e) export pixel values in a format that allows visualization of the data in 2-D spreadsheet formats or animated 3-D graphs; and (f) generate mapping functions to group data in 3-D space. Mapping functions can then be applied to independent hybridization images to distinguish and display differentially expressed clones.

For (b), the total numerical pixel value is the sum of all pixels that contribute to the spot image of a single clone in a phosphoimage. This value is limited to the maximum size of an imaginary rectangle that surrounds each spot but does not overlap neighboring spots. The actual size of the rectangle

depends on the filter manufacturer. If a unique clone is represented by two spots on the filter, the total numerical pixel value for each spot is found, and then the two values are averaged. See the Manual for further details.

The DataMachine version 1.0 software package will open and analyze industry-standard 127-dpi, 16-bit TIFF phosphoimages. [This includes Molecular Dynamics (.GEL) images (Amersham Pharmacia Biotech, Piscataway, NJ, USA) as well as Packard Cyclone images and others.] DataMachine will accommodate several types of filters. It has preset values that allow it to catalog clones from filters made by Genome Systems (Incyte Genomics, St. Louis, MO, USA), Clontech Laboratories (Palo Alto, CA, USA) and Research Genetics (now Invitrogen, San Diego, CA, USA). It can also be tailored to accommodate filters with a user-defined number of fields, rows, columns and spot sizes. Thus, it should work with nearly every phosphoimage and filter array configuration.

The DataMachine Method

DataMachine relies on a combination of user input and digital image recognition to identify the positions of each clone spot. The user first defines the type of filter being ana-

lyzed. If it is one of the three commercially available types recognized by DataMachine, then the software automatically identifies the number of fields, rows and columns of clones as well as the size of each spot on the filter. Alternatively, the user can customize these values to allow the analysis of filters of up to 30 fields, each containing 30 rows and 30 columns. Next, the user drags a grid of boxes (up to 30 × 30 boxes at one time) over the clones to define their positions (Figure 1A). The user can stretch the grid at each corner (Figure 1, B–E) or alter the positions of grid boxes individually (Figure 1F). Once the grid is defined, it can be fixed in position (Figure 2) and the process continued until the coordinates of all clones have been entered. Fixing a grid in position can, at the user's option, invoke computer coordinates correction. If this function is invoked, the user-designated clone positions may be altered by as much as ± 3 pixels horizontally or vertically to optimize the signal from any significant clones in the image. This function can be used to generate more consistent coordinates from several similar images.

Once all grid overlays have been positioned, the user can generate the fingerprint file. A new window will then appear that represents the clones in the image that have been entered into the fingerprint file as well as any that are not represented. The user can add clones or redefine their position by editing this file at any time. When fingerprint files have been generated for each phosphoimage, they can be combined to generate average or difference fingerprint files. In the former case, as many as nine images can be combined to find the average expression data at each clone position for all experiments. In the latter case, relative local values for each clone on each filter are automatically generated, and then differences between the relative values on any two fingerprint files can be found. The user can optionally ask that only clones that are significant on one of two filters be output. This option allows rapid identification of differentially expressed clones.

The data from either a mean comparison or a difference

Figure 1. Dragging and aligning a DataMachine grid over an array phosphoimage (see text for details).

Figure 2. Software feedback of the fixed positions of the array elements (see text).

comparison is saved in a new fingerprint file. Values from original, mean or difference fingerprint files can be output in spreadsheet format. The user can output any number of values, from a sample of the top values to all 27 000 values in a custom image. The output is automatically formatted so that each clone pixel value is presented next to its field, row and column designations. Output can be ranked according to either pixel value or position in the filter (field 1, row 1, column 1 being first and so on). Formatting by position allows the user to directly compare all clones on two filters by pasting the data to neighboring columns in a spreadsheet.

The user can also specify the type of background correction requested. Background correction options include using a blot average (or global average), a local average or no correction (raw data). Local average is useful for finding corrected values of clones on blots with nonuniform backgrounds. Raw data can be output and then corrected in spreadsheet format, if desired, using a selected clone or clones as standards on the filter. Commercial services generally find the average value of all pixels in the phosphoimage and use this value as the background. As long as the blot has a uniform background, the preference in our laboratory is to output and save the raw data, which we can then manipulate with background correction later. The user can also choose an arbitrary position on the blot as background. If the user samples the value of an arbitrary position on the blot, this value can then be subtracted from all raw values using a spreadsheet program.

Comparing Expression Fingerprints in 3-D

In addition to spreadsheet format, the user can request that the data be exported in a form suitable for display in a 3-D graph (Figure 3). The 3-D format is useful for representing data from a large data set. DataMachine will output the top

Figure 3. Representation of the high clones from an array hybridization in three dimensions (see text).

clone values in a file ready for importing into a free, limited version of the shareware Pro Fit. Viewing data in 3-D requires Pro Fit version 5.1, a trial version of which can be downloaded free from <http://www.quansoft.com/index.html>. The required 3-D module that plugs into Pro Fit to give it 3-D capabilities is available free at ftp://ftp.quansoft.com/pub/customer/quansoft/modules/Plot3D_210.hqx. Pro Fit will import the data and display it as an animated 3-D graph. DataMachine automatically sets the color of each clone in the 3-D graph according to the clone's pixel values. Red spheres in the graph are the most intense clones, blue are intermediate and green are the weakest. The position of each clone in the 3-D graph is determined by the position of the clone in the filter (e.g., filter position: field 1, row 1, column 1 translates to graph coordinates X, Y, Z = 0, 0, 0; filter position: field 1, row 1, column 2 translates to graph coordinates X, Y, Z = 1, 0, 0).

DataMachine additionally allows comparison of expression data sets in 3-D. At the user's option, DataMachine will group the highest value clones from one hybridization experiment and represent these near the origin in a 3-D graph. The

Figure 4. Comparison of high clones from two different array hybridizations after applying the same grouping function to each array and plotting in three dimensions (see text).

functions used to group the highest clones, the mapping functions, can be saved and applied to a second hybridization experiment. High clones that are the same in both experiments will be displayed near the origin for both experiments. High clones in the second experiment that are different from high clones in the first experiment will be displayed at isolated positions away from the origin (compare Figures 4A and 4B). Thus, differentially expressed clones are readily visualized in the animated 3-D space. DataMachine also contains a utility for converting from the XYZ coordinates in 3-D space to the field, row and column designations on the original filter. Therefore, the positions of clones visualized in 3-D space can be readily identified on the hybridization filter.

DISCUSSION

DataMachine has three valuable characteristics. First, it simplifies the task of quantifying array data. Second, it inserts numerical values for expression data in a file format that allows calculation of averages and differences in independent experiments. Third, it facilitates visualization of array data from either 2-D spreadsheet formats or animated 3-D formats.

ACKNOWLEDGMENT

The author would like to thank the Scleroderma Foundation for its generous ongoing support.

Received 3 August 1999; accepted 3 April 2000.

Address correspondence to:

Dr. David Strehlow
Department of Medicine
Boston University Medical Center
80 E. Concord St.
Boston, MA 02118, USA
e-mail: strehlow@bu.edu