white paper

Genowiz™

A truly robust Expression analyzer

Abstract

Gene expression profiles of 10,000 tumor samples, disease classification, novel gene finding, linkage analysis, clinical profiling of diseases, finding cis-and trans-elements of co-regulated genes. How much time would this immense task take? How many giga-bytes of memory would it require? And finally, how meaningful and easy will the results be to comprehend and trust?

DNA microarrays have revolutionized the way in which we study the complexity of biological effects. Analysis reams that exploit large data warehouses of gene expression experiments are necessary to realize the full potential of this technology. There are however, limited number of tools for data processing, storing and retrieving the information and analyzing the results in the context of existing knowledge. Large-scale, high-throughput experimental methods require material and information processing systems to match. In addition to being used to locate and track physical resources, they must also be able to manage very large quantities of data both before and after an experiment. They may also need to interface with laboratory instrumentation. Following hybridization of a microarray and the readout of gene expression levels, the data must be stored so that it is available for image processing, statistical and biological analysis.

Introduction

Technologies for gene expression studies have improved dramatically. Better analysis methods for large electronic data sets, reliable and effective tools for 'data mining' have become critical. The mountain of information that is the draft sequence of the human genome may be impressive, but without interpretation, all that remains is a mass of data. Gene function is one of the key elements researchers want to extract from the sequence, and the DNA microarray is one of the most important tool at their disposal.

Gene expression profiling using DNA microarrays generates reams of data. This is useless unless biologically meaningful information can be extracted and presented in some readily understandable fashion. The production of such meaningful information, involving many facets of image processing, statistical interpretation, analyses and data visualization is only possible with computer-based software. The overall process begins with cDNA or oligonucleotides spotted in two dimensional arrays onto glass slides, or synthesized on biochips using technology borrowed from the semiconductor industry. mRNAs isolated from the test and reference tissues are labeled after reverse transcription with either a red or green fluorescent dye, and hybridized to the microarray. After washing, the bound fluorescent dyes on the arrays are interrogated by a laser, producing two images, one of each color. The resulting ratio of the red and green spots on the two images provides information about the changes in expression levels of the genes across experimental conditions.

Problems and Needs

With the introduction of sophisticated laboratory instrumentation, robotics and large complex data sets, research is increasingly becoming a cross disciplinary endeavor requiring the collaboration of biologists, engineers, software, database designers, physicists and mathematicians. Techniques used in other fields can be extremely valuable to learn their proper applicability to biological problems. Microarray data analysis software developers and companies have created numerous software packages designed to analyze the images and hybridization intensity data obtained from the arrays.

Software

Microarray analysis software packages fall into three general categories. The first group consists of stand-alone packages designed for the general user. These products accept images from most microarray scanners, typically in the form of 16-bit TIFF files. They offer plenty of flexibility for analyzing data generated by different instruments and array types.

A second group consists of software packages configured to operate within specific array scanners. For example, some scanners offer an optional analysis software package that provides concurrent imaging and analysis of microarrays. This allows users to link the output images to tabular results and also get graphical ratio analysis results as each slide is scanned.

A third group consists of softwares crafted to analyze array-specific systems. The Microarray Suite Software, from industry pioneer Affymetrix Inc, services the company's GeneChip microarrays.

Image is Everything

Regardless of the category into which they fall, the essential task of array analysis software

packages is image analysis. Indeed, the primary goal is to measure the intensity of the arrayed spots and then convert those intensity values into quantified expression data. While this may appear straight forward, numerous problems can lead to questionable results. The initial issue is the assignment of a grid or template to define the spot locations from which the data will be captured. If not performed carefully, some arrays exhibit positional distortions, hybridization values can be assigned to the wrong spot which may result in error as only part of a spot has been examined. Spot irregularities stemming from the array fabrication process and bright or dark regions within specific spots caused by detritus also complicate the image analysis process.

Most software packages that carry out image analysis include an automatic alignment process that requires little, if any of user intervention. These alignment programs automatically apply a grid to the array and then extract hybridization values. Some packages, graphically flag spots as good, bad, or absent and then export these flags with all other numerical data. Still other algorithms successfully accommodate irregularly shaped spots as well as assign spot boundaries.

Background determination can also present problems, particularly when signal intensities are low, but software designers have developed various methods to determine background values. One of the most common methods involves the measurement of signals around each spot. This value is then subtracted before any ratio calculations are performed. Other approaches include subtracting a standard background value or determining background from specific areas on the array.

Spot Check

After an array of hybridization values has been gleaned from the image analysis software, hybridization ratios are calculated. Traditionally, a two-fold change in the ratio of a spot is accepted as the indication of up-or down regulation of a gene. Tabular representations of the results (often in tab delimited forms for export into spreadsheet and word processing software), ratio histograms, and scatter plots are common visualizations of the data.

As important as these results are, they hold little meaning if the underlying hybridization values are of poor quality. How can one determine if the expression changes are indeed statistically significant and thus worthy of additional study? Much of the product literature mentions statistical testing or rigid Quality Control testing of the images before data analysis. For example, softwares examine measurements such as diameter, circularity, pixel area, spot uniformity, and deviation from nominal position to determine confidence factors and a resulting pass/fail report for each spot. Similarly, some others have developed methods to place statistical confidence on the gene expression levels. Still others employ statistical methods such as analysis of variance (ANOVA), and scientists can eliminate false positives from over-expressed genes based on the statistical confidence measure selected.

Data Mining

With the hybridization data analyzed, the next goal is to search the data and arrange genes according to similarities or dissimilarities in their patterns of expression or identify functions for uncharacterized genes. This job is not trivial when thousands of genes are involved across several months of data collection.

The discovery of patterns in gene expression falls under the realm of data mining. For microarrays, data mining methods are derived from mathematical techniques known collectively as clustering. These multivariate statistical methods have become the essential tools for the elucidation of gene expression patterns in microarray data.

Data mining typically uses three types of clustering techniques. Hierarchical clustering is a common approach whereby data sets are split into classes and then subclasses, eventually forming a hierarchy displayed as a dendrogram. This technique has proved valuable in microarray data analysis. K-Means, a non hierarchical clustering method, repeatedly examines the data to create and refine clusters in order to maximize the significance of the intergroup distance. The third method, called Self Organizing Maps (SOMs), is a variation of the K-Means methodology. SOMs are a subfield of neural networks, a system of algorithms developed to explain how parts of the brain might self-organize into precise structures. Another useful data mining method, principal component analysis (PCA), is a "data reduction" technique used to identity uniquely expressing genes. PCA replaces a large number of variables with a smaller number,with little loss of information.

Visualization

Presenting all of this gene expression data in visual form, however, creates challenges to which software designers are finding colorful solutions. In addition to the typical dendrograms and red and green color block charts used by many data mining packages to depict cluster analysis results, some softwares use a unique 3-D plot with peaks and valleys to depict up and down regulated genes. They also include a 3-D scatter plot presentation of principal component analysis results, and mean, line, and bar charts for individual groups

of profiles. The graphics are very flexible and user friendly, and can be readily exported for report generation. With this same intuitive graphical environment, users can examine genotype and frequency distributions, and explore genotype phenotype associations.

Integration with Other Databases

Successful interpretation will rely on integrating experimental data with external information resources. Even more desirable would be a program that would be capable of suggesting possible explanations or hypotheses implied by the ensemble of information assembled by a process. Tools for exploring gene expression databases are still in their infancy. A major focus of infrastructure development to support large-scale gene expression studies will be in the area of electronic biological pathway databases and resources.

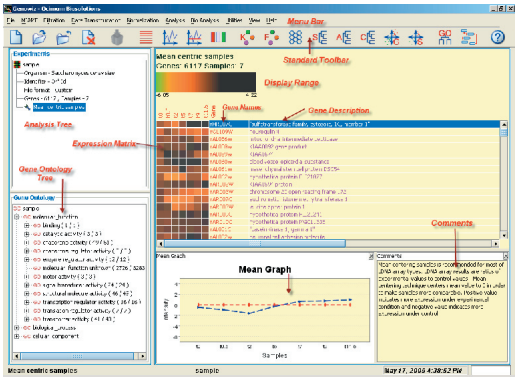
Improved Statistical Analysis

It is a fundamental assumption of many gene expression studies that knowledge of where and when a gene is expressed carries important information about what the gene does; therefore, an obvious first step is to organize genes on the basis of similarities in their expression profiles. Although cluster analysis has been the most widely used statistical technique applied to large-scale gene expression data, it is only one of several techniques that have been applied to data mining. Others methods include affinity grouping or market basket analysis, memory-based reasoning, link analysis, decision trees and rule induction, self-organizing maps and other types of neural networks and genetic algorithms. Undoubtedly these other techniques if used will prove useful in gene expression analysis.

Our Solution

Genowiz™ - A Comprehensive, Multi platform Package for Analyzing Gene Expression Data

- Genowiz™ includes a suite of advanced statistical analysis methods and offers choice for selecting analysis that is appropriate to user's dataset.
- Genowiz™ provides numerous visualization options to track down intricate correlation in microarray data. Genowiz™ allows researchers to organize experimental information, quickly and easily import data , normalize, perform cluster analysis, classify and visualize patterns, review gene information, and link analysis results to external tools.
- Genowiz™ is MIAME compliant and experimental details can be exported in MAGE- ML format. The gene list comparison and pattern simulation modules are excellent in tapping the biological significance of the data.
- Functional classification of genes, visualizing expression data in the context of metabolic, regulatory, signal transduction and disease pathways and retrieving scientific abstracts is an excellent resource for researchers to arrive at biological interpretations.
- Provision to edit and create new pathway maps, import own annotations, edit associated URLs, associate author information and generate publication quality graphs and reports make Genowiz™ a very flexible package.



Salient Features:

Easy Data Import and Handling of Multiple Experiments

Genowiz™ inherently recognizes about ten data formats pertaining to cDNA raw and processed data and Affymetrix processed data. Data of any other format including that from dye swap experiments can also be uploaded into the software. Formats which are not automatically recognized by Genowiz™ can be saved for the first time. A format once saved will be automatically recognized by the software when data of the same format is uploaded again.

Users can work with multiple experiments in one instance of the software and can perform cross experimental comparisons.

MIAME, MAGE-ML and Analysis Tracking

Experimental details can be provided using MIAME and the information can be exported in MAGE-ML format. Thus, Genowiz™ facilitates sharing of experimental details across the scientific community. All experimental analysis in Genowiz™ is tracked efficiently and saved in the database.

Excellent Pre-processing Options

Genowiz™ provides a range of data transformation, normalization and filtration options. These include:

Various adjusting parameters such as log

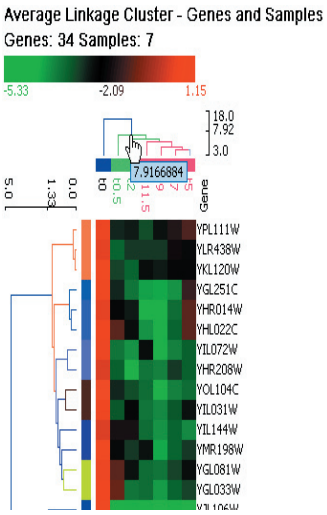
transformations, mean/median centering and Z-transformation etc. Normalization methods specific for raw (LOESS and Print tip normalization) and processed data have been provided catering to both single and double channel experiments. In addition normalization can be carried out globally or by selecting genes.

Data reduction options include replicate handling (biological and technical), parametric and non parametric tests for finding significant genes, multiple testing correction to reduce number of false positives etc. Options for dealing with missing values and filtration based on "CALLS" are also present.

Data Clustering

Clustering in Genowiz™ enables researchers to understand patterns in expression data easily by correlating the expression matrix with the graphs for the clusters. Clustering can be done for the genes or for the samples. In cases like hierarchical clustering, a two way clustering on both genes and samples can be done at the same time. Genowiz™ offers the flexibility to further partition the clusters so as to increase between cluster variance and decrease within cluster variance.

Options for partitional, hierarchical, SOM and PCA are present along with advanced clustering techniques like Gene Shaving. Visualization options for the clusters include mean, line, bar graphs, pie charts, 2-D PCA and scree plots.



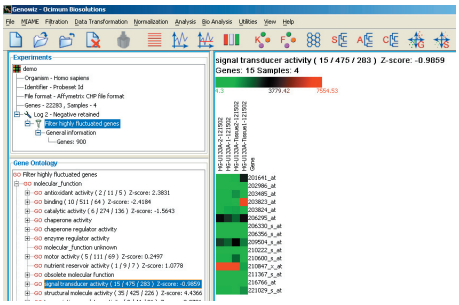
Data Classification

A user can classify data into pre-defined groups by using features like Discriminant PCA. The feature has been made very user friendly by enabling the saving of various training models used etc.

Annotations and Functional Classification

Information for genes in the data has been provided by connecting to several public domain databases. Desired annotations can be viewed and exported. Annotations are hyperlinked. Researchers can also connect to PubMed and retrieve scientific abstracts. User's also have the flexibility to upload own annotation information.

A functional classification can be performed to know the function or process of the genes and the component where they are acting. Several hypotheses can be drawn by looking at combinations of functional ontologies. Functional classification (gene ontology) in Genowiz™ is accompanied with a z-score report that attributes a statistical significance to an ontology term and directly shows the functions most effected in the data. Genes from a gene ontology cluster can be further analyzed by directly linking to a pathway.



Pathways

User can view, edit and create pathway maps. Hyperlinks can be set. Created and edited pathway maps can be saved and author details can be entered. Genes in a pathway can be saved for further analysis. A pathway map can be printed, saved as an image and can be exported in KGML format. Maps in KGML format can also be imported into the software for viewing.

Graphical Representation

Gene Expression data can be graphically explored in various ways in Genowiz™. Carefully designed visualization tools provide the following views for your data:

- Tabular representation
- Colored expression matrix
- Line Graph
- Bar Graph
- Mean Graph
- Gene Profile
- Pie charts
- 2D PCA Views
- M-A Plots
- Ch1/Ch2 Plots
- Scree Plots

Expression intensities are plotted as a colored matrix. Colors and size of the expression matrix, arrangement of the samples, background color etc. can be changed as desired. The expression matrix itself can be sorted (on right click) by intensity, mean intensities, standard deviation of intensities, gene ID, and so on. The graphs are also supplemented with context related information as tool tips. The graphs can be exported as publishing quality images.

Conclusion

Bioinformatics and data storage are the culmination of the microarray analysis process. Microarray data analysis offers an opportunity to generate functional data on a genome-wide scale and consequently provides much needed data for the biological interpretation of genes and their functions. The use of microarrays is booming in basic pharmaceutical research, because of the value gained by measuring the expression of numerous genes in parallel. The exploding amount of data generated by high throughput experiments, however, can become a minefield of options and opportunities for mistakes, particularly when it comes to statistical interpretation. Although attention is now being paid to experimental design and to methodologies for interpreting microarray data, further progress is needed. Experience has also shown that adopting standards for microarray experiment annotation, data representation and normalization can be crucial for converting mounds of data into useful and statistically significant conclusions. The possibilities are exciting, with dramatic new findings within reach. With all these tools, researchers can, with the click of a mouse, mine data to find genes of interest without too much sweating. It is like sitting in front of your TV and climbing Mount Everest!

United States: Ocimum Biosolutions Inc. Fortune Park VI, 8765, Guion Road, Suite G, Indianapolis, IN, 46268, USA
Phone: +1 317 228 0600, Fax: +1 317 228 0700, Email: us@ocimumbio.com

India: Ocimum Biosolutions Ltd, 6th Floor, Reliance Classic, Road No. 1, Banjara Hills, Hyderabad 500 034, A.P., India
Phone: +91 40 666 27200, Fax: +91 40 666 27205, Email: india@ocimumbio.com

The Netherlands: Isogen Biosolutions, (An Ocimum Biosolutions Company), Lagedijk Noord 18, Postbus 220, NL -3400 AE IJsselstein, The Netherlands
Phone: +31 (0) 30 68 78 788, Fax: +31 (0) 30 68 88 009, Email: europe@ocimumbio.com

Ocimum Biosolutions ...enabling R&D™

www.ocimumbio.com